|CURIEUX|

ACADEMIC REVIEW

November Issue

Issue 13 Part 1

Editing Staff

Chief-In-Editor

George Hasnah

Lead Editor

Amanuel Geremew

Ashley Du

Assisting Editors

Jing Yan Huang

Nidhi Potineni

Sophia Kong

Tony Yuan

Ridit Aggarwal

Manu Sharma

Rishi Gadikota

Eric Wang

Table Of Contents

Page 6: Smart PeriodCare: Conductivity-Based Saturation Detection and AI-Powered Pad Change Prediction By 1st Riana Agarwal and 2nd Vinay Vishwakarma

Page 21: The Framework for Reducing Recidivism in the United States: A Literature Review By Selina Zou

Page 35: A Review Analysis of AI in Ceramics Business Parth Kajaria

Page 46: Rigidity of the Midpoint Mean Value Property on Intervals of Fixed Length By Maksim Stepnov

Page 66: An Interactive 3D Simulation Framework for Evaluating Quantum Confinement and Emission Profiles of Quantum Dots Using a Particle-in-a-Box Model By Alya Aydinoglu

Page 79: Scheduling as an Access Barrier in Youth Sport: A Structured Evidence Review and Preliminary Planning Tool By Ranbir Goel

Page 93: Alzheimer's Disease, Air Pollution, and Biomarkers: A Mechanistic Analysis By Dhiya Prabhuram[1] and Vanitha Nithianandam, Ph.D.[2]

Page 104: The Final 50: Where Competitive Tiers Clash: Stroke Mechanics By Sophia Chen

Page 115: Frozen in Time: A Modern Look at Icing the Kicker By Rohan Thakare

Page 121: The New Wave Palestinian Music and Palestinian Identity in the Midst of Occupation and Displacement By Yusra Khalil

Page 134: Digital Navigational Aid for the Visually Impaired using Edge AI: Experimentally-Supported Advantages of Incorporating Safety-Oriented Deterministic Algorithms By Rudra Patel

Page 150: Krabbe Disease: The Current Understanding of Pathology and Therapeutic Approaches By Raymond Fadous

Page 172: How Does Socioeconomic and Geographic Background Affect Women's Financial Independence in India? By Kyra Aziz

Page 185: Value Capture and Strategic Dependence in the AI Stack: Evidence from Fintech, Healthtech, and Agtech in India and Brazil By Daria Mishina, Supervised by Professor Ganesh Mani

Page 207: Drone Swarms for Faster Wildfire Detection: From Simulation to Concurrent Flight By Ilayda Koker

Page 224: Reassessing the Neuropsychological Burden of Pediatric Hemophilia Across Treatment Eras By Rahim Noor

Page 243: Assessing Flood Mitigation Strategies for Urban Resilience: A Comparative Review of New York City and HoustonBy Eva Pirolli

Page 264: Protecting Children in Online Gaming: Legal Responses to Grooming and the Limits of Platform Accountability By Dongwon Shin

Page 282: An Exploration of the Link Between Music Preference, Self-Esteem, and Interpersonal Connection: Collective Self-Esteem By Tiffany Wang

Page 295: Evaluating Visual Perception Architectures for Autonomous Vehicles By Vihaan Siramdas

Page 308: The Role of Ultra-Processed Foods in Neurodegenerative Diseases through Gut Dysbiosis in Adolescents By Hannah Wang

Page 321: Comparing Heatmap Regression, Segmentation, and Multi-Frame Temporal Context for Surgical Tool-Tip Localization Under a Unified Evaluation Protocol By Tara Tony[1]

Page 333: A Cross-Program Benchmark of Electric Propulsion Design By Christian Sent

Page 346: Evaluating ASPM as a Genetic Modifier of Congenital Zika-Associated Microcephaly By Aanya Sood

Page 356: Astronomical Technologies and Cosmology in Ancient Greece, China, Egypt, and the Maya By Maya Fitzgerald

Page 365: Self-Reconfigurable Modular Robots for Ocean Exploration: Applications of Adaptive Robotic Systems By Xavier Song

Page 379: Competing on Values - A Multiple-Case Qualitative Analysis of Fashion Retail Platform Strategic Positioning By Ariana Riggi

Page 402: The Conceptual Frameworks of AI Productivity Paradox: A Theoretical Review By Chau Nguyen

Page 416: CSR Disclosure and Stock Returns: Evidence from Selected UAE-Listed Real Estate Firms By Riaan Shah

Page 428: Relativistic Effects from Special Relativity to Black-Hole Accretion Disks: Theory, Paradoxes, and Numerical Visualization By Beatrice Lee (student)[1] *

Page 450: Invisible in the Data: Middle Eastern and North African (MENA) Women and Pharmacogenomic Uncertainty in Clinical Drug Research By Arianna M. Brown

Page 463: Aerodynamic Drag Reduction and Thermal Management in Electric Vehicles: Trade-offs for Energy Efficiency and Driving Range By Ada Yilmaz

Page 486: To What Extent Do Taxes on High Income and High Wealth Individuals Affect Their Financial Behaviour and Location Decisions, and Why do These Responses Vary Across Contexts? By Pranav Krishna Narayanan

Page 494: Additive Manufacturing Technology in Civil Engineering and Its Application in Seismic Zones By Yerbol Janzakov

Page 518: A Study of Wing Aspect Ratio for Aerodynamic Efficiency Across Different Mission Profiles By Vatsal Singhal

Page 538: The BDNF Val66Met Polymorphism As A Potential Mechanism of Individual Differences in Cognitive Vulnerability to Sleep Deprivation: A Review By Hania Nabeel

Page 547: Ecological Impacts of Large-Scale Algae Cultivation on Aquatic Biodiversity: Evidence from Environmental Deoxyribonucleic Acid, Remote Sensing, and Water Quality Analyses By Sasha Kamath

Page 558: An Ethnographic Study on How Healthcare Classes Prepare Students to Work in the Healthcare Field By Ibrahim Ali

Page 570: Evaluating Structural Risk and Capital Access Barriers: Why Turkish Startups Experience Higher Early-Stage Attrition Than European Counterparts By Ayşenur Kurplikaya

Page 581: Relationship Between Stress and Buying Patterns By Kairrah Chaudhary

Page 590: Investigating the Role of TikTok in the Growth and Success of Gen Z Fashion Brands: A Case Study of Pink Palm Puff (PPP) By Raeka Ren

Page 603: The Effects of Achondroplasia on Quality of Life By Georgia Wei

Page 614: Does Interest-Rate Sensitivity Organise Volatility Transmission in Indian Equity Markets? By Aarav Shah

Page 631: Musical Symmetry and Parallel Modulation via the PLR Group By Jitong Lei

Page 643: Screening of Selected Clay Minerals for RNA Compatibility: Implications for Prebiotic Chemistry on Early Earth and Mars By Nidheesh Prathi

Page 655: Preserving Little Saigon: Economic Stability and Cultural Continuity in Westminster, California By Andy Cao

Smart PeriodCare: Conductivity-Based Saturation Detection and AI-Powered Pad Change Prediction By 1st Riana Agarwal and 2nd Vinay Vishwakarma

Abstract

Improper or infrequent pad replacement can lead to discomfort, skin irritation, leaks, and health risks, especially in the case of young girls, elderly patients, bedridden patients, and persons with impaired communication skills. Traditional monitoring is based on self-awareness and physical checking of the pads, being burdensome for caregivers and uncertain. Smart PeriodCare offers a system using both reusable and disposable functions, based on conductivity impact detection and application of AI in predicting the moment of switching the pads. Two copper tape electrodes placed in the absorbent layer can measure electrical conductivity changes according to the saturation level of the menstrual flow. A transmitter processes the information and makes it transmitted wirelessly via ESP-NOW to the receiver ESP32-C6 that gives colored LED and vibration signaling. The process of development was iterative, starting from 14 failures in integrating the pad and finishing with 35 tested pads with varying pad sizes, electrode length and materials. The final configuration included copper electrodes of 7cm in length and 120 ml of fake blood at a ratio of glycerol to water of 1 to 5. The physical detection system showed 85.71% accuracy (30 out of 35 cases) while the Random Forest system had 80.00% accuracy in prediction. The system integrates 6 predictive inputs with real-time sensing, demonstrating a practical dual-mode approach for timely and discreet hygiene monitoring.

Keywords

Smart PeriodCare, Conductivity Sensing, Saturation Detection, Predictive Pad Monitoring, Copper Electrodes, AI-Based Change Prediction, Reusable–Disposable System, Assistive Hygiene Technology

1. Introduction

Menstrual hygiene continues to be a worldwide challenge for health and dignity; this is especially a problem for adolescents and those who have difficulties keeping track of or highlighting their hygiene needs. UNICEF estimates that about 1.8 billion people menstruate worldwide, although issues like stigma, poverty, lack of facilities, and limited access to menstrual products continue to impact the management of menstrual health [16]. Even though there has been improvement regarding access to menstrual health products, it is worth mentioning that users might not be sure about the exact time of changing pads; this means that they could have difficulties with leakage and feeling uncomfortable.

Conventional disposable products, user education, regular manual checking and growing use of sensors dominate current methods of menstrual and hygiene management. Research on smart absorbents are focused on detection of fluid through conductivity,

1st Riana Agarwal and 2nd Vinay Vishwakarma, Adani International School, Ahemdabad, India rianajindal@gmail.com
2nd Vinay Vishwakarma Mentor, On My Own Technology Pvt. Ltd, Mumbai, India , vinay.vishwakarma@omotec.in

resistance, humidity, capacitance as well as biosensors. In 2023, a review by M.Ben et al [17] for non-invasive urinary leakage systems scrutinized 279 articles and found 12 relevant studies on sensor application, where conductivity sensors were by far the most widely used. Flexible humidity sensors and capacitance systems are the second most popular solution for automated monitoring of wetness and absorbed volume [18, 19]. Nevertheless, issues with positioning of the electrodes, false alarms, portability, comfort and interaction with sanitary products remain up-to-date.

To address these challenges, Smart PeriodCare is a reusable and disposable solution combining real-time conductivity sensing with predictive pad-change guidance. Two copper electrodes embedded within the absorbent region detect changes in electrical conductivity as fluid reaches the sensing zone. In parallel, a prediction dashboard uses six input parameters to generate three recommendations: No change needed, change soon and change recommended. Thus, the system combines detection with predictive decision support.

2. Related Work

In the article by Cho et al [1] a smart diaper system utilizing bluetooth technology to measure whether or not the system could help decrease the number of times caregivers were required to manually check diapers and increase efficiency in providing care. Overall, the results indicated that there were clinically useful estimations for volume as well as some level of promising agreement with measurements taken using standard methods, with room for improvement in both the rate of detection and usability of the system among older adults.

Sen et al [2] created a low-cost disposable wetness sensor using hydrogel and RFID technology. They make use of how absorbent a diaper is to produce a sensor that behaves differently based on whether or not there is moisture present, allowing for wireless detection of whether or not the diaper is wet without having to use a battery on the disposable portion of the sensor. This research is highly relevant to the development of scalable smart diaper systems and reducing electronic waste from disposable sensors.

This paper by Sennesael et al [3] presents a new method of applying printed carbon-ink sensors to create a contactless smart diaper system for caregivers by using a low-cost and easily produced disposable sensor to read from reusable electronics via a flexible strip that connects both devices. The authors studied the moisture response, dynamic range, and practical sensing of different types of disposable diapers across several trials of each.

This research by Tanweer et al [4] looks at the use of printed capacitive sensing coupled with front-end electronics as a means to monitor fluid volume in diapers. The proposed diaper design provides an accurate means of detecting voiding events while ensuring that the disposable capacitance sensor layer remains simple and low-cost. The results of this research provide valuable insight into how capacitive sensors can be used in conjunction with compact electronics to provide accurate measurements for wetness and volume.

This article by Tanweer at al [5] looks into the use of capacitive sensors in adult diapers to determine the amount of fluid collected. The way in which the type of sensor design,

placement and construction interact with the sensors' sensitivity to being wet, and provides guidelines for how to choose electrode spacing and depth of sensing within absorbent hygiene products. The findings in this study are useful, because they deal with the main challenge of transforming the saturated state of liquid to an electrical change that can be measured in a diaper.

This study by Dobson et al [6] describes an in-pad diagnostic platform to detect biomarkers for diseases found in menstrual flow. Rather than focusing on wetness, this study uses biochemical detection to create diagnostic sensing devices inside hygiene pads. This research provides important context for possible future developments of smart hygiene pads by outlining the progression in how sensor components will be used for both saturation detection and health screening applications.

This study by Tanweer et al [7] looks at printed electrodes for harvesting energy from urine to operate IoT nodes for smart diapers. The goal is to use urine as a power source, thus reducing reliance on disposable batteries for powering smart diapers with continuous wireless sensing capabilities. This paper will be useful for sustainable smart diaper design, especially when we want continuous wireless sensing.

This article by Ziai et al [8] presents an innovative Smart RFID tag used for detecting moisture levels in diapers. The research represents an early contribution to demonstrating the feasibility of passive wireless sensing without requiring active sensors attached to disposable diapers or any external power source. The results provide valuable insight into the design of future smart diapers that integrate disposable sensors with reusable readers.

In this early research by Siden et al [9] an initial smart diaper moisture-detection system was developed based upon using wireless sensing technology to detect diaper wetness in order to alert caregivers when their baby was wet. The historical value of this document lies in demonstrating some of the technical feasibility of smart diaper monitoring prior to the existence of modern technology for printing electronic circuits and connecting mobile devices.

The authors Tanweer et al [10] of this paper study printed coplanar capacitive sensors as a method to measure fluid (urine) volumes within adult diapers. They explore the effects of sensor layout and configuration on the reliability, sensitivity, and response time of the sensors to liquid accumulation. This work is constructive and engineering oriented in nature and will aid engineers in determining electrode spacing and design printed patterns.

This research article by Sennesael et al [11] develops a smart disposable diaper with a carbon-based, noncontact sensor for the purpose of providing an environmentally sound alternative to using electronic sensors on the diaper. The authors present their design for a disposable sensor that can be used with a reusable controller so as to minimize electronic waste. The focus of this paper is on environmentally sound practices, use of low-cost materials and manufacturing processes.

The article by Wang et al [12] presents an RFID sensor made from conductive yarn to allow for accurate and continuous monitoring of the urine produced by a child. The conductive yarn provides a non-rigid, textile-based sensing medium which will better integrate with the

diaper's design than traditional electronics. The use of RFID technology allows for remote reading of the urine from the diaper and a low-cost method of deploying these sensors.

The intention of this paper by Mukherjee et al [13] is to elucidate the technical challenges associated with the design space of textile-based smart sanitary napkins while also discussing potential future opportunities within that area. This paper is more useful as a reference on design constraints. As such this study will assist in guiding the development of the sanitary pad component of the overall system and provide insight on obstacles preventing real-world use.

The current research by Kim et al [14] will demonstrate the validity of the use of a smart diaper system in a population of nursing home residents. The primary focus of this research is on the automatic detection of voided volume and comparison of readings provided through the smart diaper with those taken using instrumented or clinical standards. This research is important because it involves the concept of using smart diapers in a laboratory setting to evaluate how that same technology performs with respect to real issues of concern for a vulnerable group of patients receiving long-term care.

The goal of this review by Bae et al [15] is to provide an overview of how to conduct safety evaluations and test methods for absorbent hygiene pads. This review will cover material safety, skin contact issues, absorbency behavior, and various ways to test. The paper is important because it gives a more complete picture of safety from a product perspective as it relates to the design of the sensor. It will also provide justification for making decisions regarding electrode placement, how close the electrode is to the skin, compatibility between materials, and long-term comfort and safety requirements when developing your smart sanitary pad prototypes.

3. Methodology

The suggested Smart PeriodCare system was created as a hybrid disposable-reusable platform that integrates detection of saturation using direct conductivity, creation of wireless alerts and predicting pad-change using artificial intelligence. The methodology consists of three steps:

1. Conductivity measurement and saturation detection
2. Wireless transmission and alert generation
3. Prediction of the suitable time for a pad change using AI

The system as a whole intends to offer the latest information about the current state of saturation and make a prediction on when the pad should be changed while the physical system consists of conductive electrodes placed into the absorbent part of the sanitary pad, and Seeed Studio XIAO ESP32-C6 controller. The second XIAO ESP32-C6 is a receiver that gives visual and tactile alerts.

3.1 Conductivity Measurement and Saturation Detection

(i) Principle of Operation: The functioning of the initial phase of the system relied on testing the level of electrical conductivity caused by the arrival of a conductive fluid at the electrodes separated from each other while being placed within the absorbent layer of the sanitary pad. In the dry state, the resistance of the circuit connecting the electrodes is extremely high. The arrival of the test liquid into the area of electrodes creates the path for electrical current, which leads to the decrease in resistance and increase in conductivity. This differentiation is essential because the changes in the amount of liquid absorbed and the reaction of the sensing system can vary depending on the type of pad material, its shape, the rate of fluid absorption, the place of electrodes, and the properties of the liquid.

(ii) Electrode Development: To check for the ability to achieve conductivity before proceeding with further tests, conductive electrodes were first assessed in isolation. Copper wire and nichrome wires of 0.6 mm each were studied as the possible electrodes. After that a 6mm copper tape electrode was tested with 4cm and 7cm lengths. The 7cm and 6mm copper tape electrode had proven to be successful since the longer length gave better conductivity range. The first trials that involved embedding the electrodes in the pads resulted in 14 pads being damaged or torn in the process of combining different pads.

(iii) Blood Replication Fluid: In order to achieve the aim of conducting the experiments under repeatable laboratory conditions, a blood-replication fluid was utilized instead of using human menstrual blood. The experimental fluid was made of glycerol and water in the ratio of 1:5. The solution was hence used since it provided a conductive fluid with the required flow behavior in order to run the sensing experiments. At last, 120 mL of blood-replication fluid was used in this experimental configuration. This value is interpreted as the final established experimental condition, which should not be treated as a standard definition of absolute saturation of a sanitary pad.

Multiple tests and configurations were performed in order to achieve the correct electrode and blood replication viscosity as mentioned in Table 1. All these trials were treated as engineering optimization work and were used to improve the positioning of the electrodes on the next rounds of testing.

Table 1: Experimental configurations evaluated with different sanitary pads

Pad Size	**Fluid / Test Condition**	**Electrode Configuration**	**Purpose of Test**	**Outcome / Decision**
XL pad	200 mL blood-replication solution	0.6mm Copper wire / Nichrome wire	Establish conductivity response in a larger pad	Possible chance of irritation and rashes from wire
S pad on curved structure	200 mL blood-replication solution	0.6mm Copper wire	Evaluate sensing under curved/realistic placement	Curved surface test successful. Electrode size too small

L pad	120 mL blood-replication solution	**Smaller electrodes**	Investigate sensing/leakage response with reduced electrode coverage	Compared with longer-electrode configuration
S pad	120 mL blood-replication solution	**7 cm vs 4 cm electrodes**	Determine suitable sensing length	**7 cm selected**
L pad on curved structure	120 mL blood-replication solution	**Nichrome vs copper electrodes**	Compare electrode materials	**Copper selected**
XXL Pad	120 mL blood-replication solution	**7 cm copper tape electrodes with 1:5 glycerol:water**	Final optimized configuration	**Selected for final prototype**

(iv) Electrical Measurement: The electrode's output was connected into the analog input of the Seeed Studio XIAO ESP32C6 based controller. The data received from the sensors would then be digitized by the microcontroller with the help of its ADC. The program for conducting experiments had a 12-bit ADC narrowed down to a digital resolution of 0-4095. The obtained ADC value could be converted into voltage. At first, the experiments utilized an empirical threshold of about 2500 ADC counts that helped in identifying the dry and conductive states. A voltage divider configuration was built using a known resistance value of 100k ohms to measure the conductivity. In the proposed approach, the higher ADC value was indicative of a dry status while the lower value meant higher conductivity in the area of the electrodes. The same can be seen in Figure 1 below.

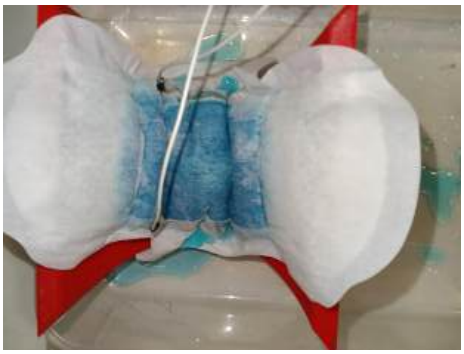


Fig 1: XL size pad with copper electrodes embedded for measuring ADC Values.
Hence the basic classification was validated as shown in Table 2.

Table 2: Conductivity classification parameters

Sensor condition	ADC response	Interpretation
Dry	> 2300	Low conductivity

Partially Saturated	2300 - 1800	Increased conductivity
Fully saturated	< 1800	Fully conductivity

(v) Two-Node Architecture for real-time detection: The system was developed using two XIAO ESP32-C6 microcontroller units so as to keep the disposable sensing element separated from the user alert mechanism. The first unit, referred to as the sender node, was connected to conductive electrodes built within the sanitary pad with the purpose of:

1. Collecting conductivity measurements.
2. Establishing the present saturation state of the device.
3. Translating the responses of the sensor into the saturation category.
4. Providing the status information by sending it wirelessly to the receiver unit.

The device 1 can be observed as shown in Figure 2.

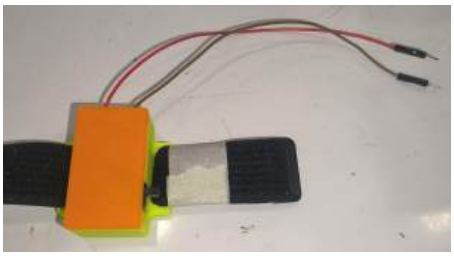


Fig 2: Device 1 with electrodes connection for Sanitary pad

The second unit, referred to as the receiver node, was made as an alert device. It had a task of receiving the information about the saturation state and delivering different kinds of visual and haptic feedback. The device 2 can be seen in Figure 3.

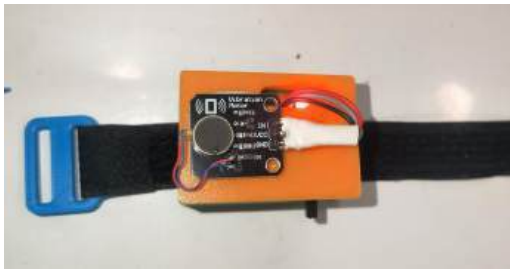


Fig 3: Device 2 with alert LEDs and Vibration motor module

While deploying, the ESP32 keeps sampling every input from the conductivity sensor and comparing these readings with the preset threshold parameters. Thus, this method enabled the sensing system to perform continuous supervision over the pad without intervention from the user or caregiver for taking manual readings.

(vi) Wireless Communication: Two ESP32-C6 devices utilize ESP-NOW technology to communicate wirelessly, resulting in low-latency communication. The sender sends the measured data in a continuous manner to the receiver. The data format includes information about the saturation level, sensor readings, and voltage data. The wireless receiver enables the possibility to reuse the device.

(vii) Alert Classification: The receiver interprets the transmitted saturation status and provides progressively stronger alerts. The demonstrated alert structure is as described in Table 3. The yellow LED serves as an early warning light that the pad is due for observation and replacement. A higher alert is indicated by the red LED and vibration motor once the measured conductivity reaches the level of higher alert. Both visual and haptic feedback were used so that the device could be utilized in scenarios where the user would not be able to actively watch the display. Haptic feedback ensures that the alert can stay relatively discreet. The complete system is as shown in figure below.

Table 3: Alert Generation structure

Detected condition	**Yellow LED**	**Red LED**	**Vibration**
Dry	OFF	OFF	OFF
Partial Saturation	ON	OFF	OFF
Full Saturation	OFF	ON	ON

The complete experimental setup with the integration of the sanitary pad with the sender device and the receiver device is as shown in Figure 4.

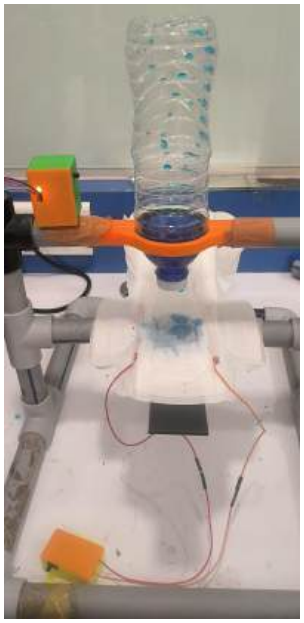


Figure 4: Complete experimental setup

3.2 AI-Based Pad-Change Prediction

The second section of Smart PeriodCare was designed for predictive estimation of sanitary pad change using the Random forest classifier. Although the physical system detects the present situation, the AI model makes a prediction of when there is a need for changing the pad based on the user inputs. The predictive AI model has been designed for software development only and is not meant to predict a medical diagnosis.

(i) Input Variables: The prediction model uses six user input variables -

1. Age
2. Pad type
3. Pad size
4. Flow type
5. Hours since the last pad change
6. Cycle day

Further the pad-type variable allows different absorbent constructions to be represented, including -

1. SAP
2. Fluff pulp
3. Airlaid paper
4. Cotton

Variables chosen included parameters that could affect use of the pads and the anticipated time frame for their replacement.

(ii) Dataset Preparation: The dataset was created with 6000 records for the purpose of developing and testing the artificial intelligence model . This data set consisted of the defined inputs as well as the target recommendation category. The three prediction classes were:

1. No Change Needed
2. Change Soon
3. Change Recommended

Data preparation was performed using Pandas, while categorical encoding and modeling were performed in the Python machine learning environment.

(iii) Random Forest Model: A Random Forest Classifier was selected as the machine-learning model for the prediction component. The Random Forest algorithm entails combining several decision trees to create a classification output based on the combined results of those decision trees. In this case, the model takes in the six usage/user variables and generates one of the three pad change recommendation classifications. The model was saved using Joblib and could then be used by the application backend when receiving a new prediction request.

Software Architecture: The AI prediction system was designed using a web-based framework that comprises. The model workflow is as shown in Figure 5. The input parameters

are entered by the user via the web application's interface. These values are collected by JavaScript and sent to the Flask backend using the prediction request. Following operations are performed by the backend:

1. Receives the user input.
2. Formats the input according to the trained model.
3. Applies the required preprocessing.
4. Loads the trained Random Forest model.
5. Generates the predicted pad-change category.
6. Returns the prediction to the browser.
7. Displays the recommendation and model confidence.

This kind of design allows for the integration of the prediction model in an intuitive interface, rather than directly interacting with the Python or machine learning code.

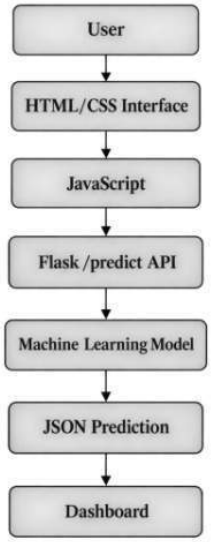


Figure 5: AI Model system architecture

(iv) Prediction Output: The algorithm outputs one of the following recommendation categories:

1. No change required - The present input situation suggests that no immediate change is expected.
2. Change soon - The model recommends that the user keep an eye on the pad and be ready for a change.
3. Change recommended - The model suggests changing the pad based on the input provided by the user.

Moreover, the user can see the confidence of the model in its prediction along with the predicted time-based output. It is important to note that the outputs provided by the system are

demo outputs only and should not be considered as validated clinically based recommendations. The prepared dashboard is as shown in Figure 6 below.

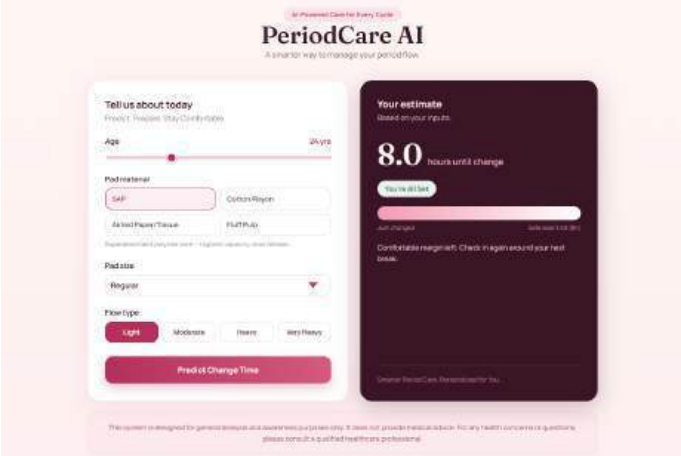


Figure 6: PeriodCare Dashboard

4. Results and Analysis

4.1 Physical Layer: Conductivity-Based Saturation Detection

The evaluation of the physical sensing layer was carried out with the aid of 35 pad tests which saw the testing of four different types of pads. During iterations of development, changes in the design and configuration of the electrodes were accomplished along the course of testing different sizes of pads, materials, and dimensions of the pads. As a result, copper electrodes of 7 centimeters have been used in the final setup. Further, the conductivity measurements indicated a significant difference between the readings of the ADC of the device as the amount of fluid used increased. Thus, the ADC value of 4095 was achieved at a dry state, and the ADC value for the partially saturated and fully saturated pads were obtained at respectively 2300 and 1800 ADC values. Across the 35 trials, the physical system correctly identified 30 cases, giving a detection accuracy as shown in Eqn (1) -

$$\frac{30}{35} \times 100 = 85.71\% \qquad (1)$$

Thus, 5 of 35 trials were incorrectly classified. Among the tested configurations, the XXL SAP pad produced the best observed performance. The results indicate that electrode placement and pad construction influence the conductivity response, supporting the need for configuration-specific calibration.

4.2 Predictive Layer: Random Forest Classification

The predictive layer was analyzed with reference to thirty-five instances, which were grouped into three classes, namely No change needed, Change soon and Change recommended. The prediction yielded twenty-eight successful predictions and seven unsuccessful ones, which made the overall accuracy forty percent. However, the class-wise results yield more information than just the overall accuracy. For No change needed, twelve were predicted but only fifteen

were actually classified as such: thus giving precision of sixty-six percent, recall of eighty-three percent and F1/87-score of seventy-four percent. In terms of the change soon classification, the precision is 100 percent, recall is sixty-eight percent and F1-score is eighty-percentage. Finally, for change recommended, the model yields a precision of seventy-seven percent and a recall of 100 percent. The overall weighted F1-score was 80.79%, indicating that the model's class-wise performance was broadly consistent with its 80% overall accuracy. The prediction model analysis is as shown in Table 4.

Table 4: AI model analysis

Metric	No Change Needed	Change Soon	Change Recommended
Actual cases	12	16	7
Predicted cases	15	11	9
Precision	66.67%	100.00%	77.78%
Recall	83.33%	68.75%	100.00%
F1-score	74.07%	81.48%	87.50%

4.3 Integrated System Analysis

The findings indicate that both layers function together efficiently. The physical layer defines the current state of sensing whereas the predictive layer uses six parameters including age, pad type, pad size, flow type, hours since change, and cycle day for indicating changes. The physical system scored an accuracy of 85.71% while the predictive model achieved an accuracy of 80.00% by obtaining a weighted F1-score of 80.79%. Hence, the overall framework provides more than just wetness detection by helping to obtain real-time information as well as offering predictive decision help. However, the findings are based on the prototype evaluation of the project. The predictive model utilizes a demo/synthetic dataset, thus the results obtained should not yet be seen as clinical validation. More data, independent validation, and practical testing are needed before implementation in healthcare or menstrual care.

5. Conclusion

The dual-layer method for accurately detecting saturation status and predicting when it is time for a new pad was demonstrated by the Smart PeriodCare system. The prototype was developed through trials on 35 pads across four types after 14 pads initially broken during attempts to integrate electrodes. The success in using the correct material and electrode dimensions (length, arrangement, and configuration) made it possible to use 7 cm electrodes. The research produced a system with 85.71% detection accuracy based on the successful classification of 30 out of 35 trials. The system generated three response types: no alert, yellow blinking LED with low vibrations, and red blinking LED with high vibrations. The second layer of Random Forest predictive model assessed 35 situations and made 28 correct predictions.

Finally, the overall prediction achieved an accuracy of 80% and a weighted F1 score of 80.79% with the class F1 score equal to 74.07% for No Change Needed, 81.48% for Change Soon, and 87.50% for Change Recommended. The model analyzed six parameters describing the user and the pad helping to generate three recommendations for changing the used pad.

For the future development of the system, development can be focused on combining the physical layer for sensing and the AI layer for prediction into an adaptive system. Information from the microcontroller regarding conductivity may be used as an additional input for the prediction model. Therefore, the two systems may be merged in such a way that one can obtain the required recommendation for the change through "prediction + real sensor situation," and this will allow a real time feedback system. The second development of the concept may imply replacement of copper wires/tapes with printed electrodes using conductive ink. Conductive ink can be used for screen printing and inkjet printing, and can be applied directly to a thin flexible substrate, which allows having a thinner, more even, and possibly scalable sensing layer. Printed electrodes may be made according to different geometry and integrated into the pad during manufacturing, while the receiver and sender module remains reusable and detachable.

Acknowledgement

I would like to express my sincere gratitude to my mentors Ms. Ira Doshi and Mr. Ariz Naiz of On My Own Technology Pvt. Ltd., for their comprehensive guidance and unwavering support throughout the entire project and research work.

Works Cited

[1] Cho, Jae Ho, et al. "A smart diaper system using bluetooth and smartphones to automatically detect urination and volume of voiding: prospective observational pilot study in an acute care hospital." Journal of medical Internet research 23.7 (2021): e29979.

[2] Sen, Pankhuri, et al. "Low-cost diaper wetness detection using hydrogel-based RFID tags." IEEE Sensors Journal 20.6 (2019): 3293-3302.

[3] Sennesael, Joryan, et al. "Carbon-ink sensing patterns for a contactless smart diaper system." 2024 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS). IEEE, 2024.

[4] Tanweer, Muhammad, et al. "Smart diaper with printed capacitive sensors and integrated front-end to monitor voided fluid volume." IEEE Sensors Journal 24.9 (2024): 14443-14453.

[5] Tanweer, Muhammad, et al. "Development of capacitive sensors to detect and quantify fluids in the adult diaper." International Conference on Bio-inspired Information and Communication Technologies. Cham: Springer Nature Switzerland, 2023.

[6] Dosnon, Lucas, et al. "A Wearable In‐Pad Diagnostic for the Detection of Disease Biomarkers in Menstruation Blood." Advanced Science 12.32 (2025): e05170.

[7] Tanweer, Muhammed, et al. "Sustainable printed electrodes for energy harvesting from urine to power IoT sensor nodes in smart diapers." International Joint Conference on Biomedical Engineering Systems and Technologies. SciTePress, 2024.

[8] Ziai, Mohamed A., and John C. Batchelor. "Smart radio‐frequency identification tag for diaper moisture detection." Healthcare technology letters 2.1 (2015): 18-21.

[9] Sidén, Johan, Andrei Koptioug, and Mikael Gulliksson. "The" smart" diaper moisture detection system." 2004 IEEE MTT-S International Microwave Symposium Digest (IEEE Cat. No. 04CH37535). Vol. 2. IEEE, 2004.

[10] Tanweer, Muhammad, et al. "Evaluation of printed coplanar capacitive sensors for reliable quantification of fluids in adult diaper." Mediterranean Conference on Medical and Biological Engineering and Computing. Cham: Springer Nature Switzerland, 2023.

[11] Sennesael, Joryan, et al. "An Environmentally Friendly Carbon-Printed Contactless Smart Diaper." IEEE Sensors Journal 25.2 (2024): 3309-3319.

[12] Wang, Lijun, et al. "Conductive fiber-based RFID sensor for real-time urine monitoring in smart diapers." Materials & Design (2026): 115777.

[13] Mukherjee, Manideepa. "Challenges and opportunities of textile based smart sanitary napkin design." Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers. 2019.

[14] Kim, Jae Heon, et al. "Accuracy of a smart diaper system for nursing home residents for automatically detecting voided volume: instrument validation study." JMIR Formative Research 8 (2024): e58583.

[15] Bae, Jihyun, Hoonjeong Kwon, and Jooyoun Kim. "Safety evaluation of absorbent hygiene pads: a review on assessment framework and test methods." Sustainability 10.11 (2018): 4146.
[16] “Menstrual Hygiene.” UNICEF, www.unicef.org/wash/menstrual-hygiene.
[17] Ben Arous, Myriam, et al. "Non-invasive wearable devices for urinary incontinence detection—a mini review." Frontiers in Sensors 4 (2023): 1279158.
[18] Tekcin, Meltem, et al. "Wearable and flexible humidity sensor integrated into disposable diapers for wetness monitoring and urinary incontinence." Electronics 11.7 (2022): 1025.
[19] Shichitani, Keisuke, and Kazuki Nakajima. "Urine non-contact sensor that evaluates multiple urination using capacitance change to inform when to replace the inner diaper." Frontiers in Digital Health 7 (2025): 1569144.

The Framework for Reducing Recidivism in the United States: A Literature Review
By Selina Zou

Introduction

Recidivism, the tendency for formerly incarcerated individuals to reoffend, remains one of the most pressing challenges facing the U.S. criminal justice system. Despite substantial investments in punitive policies and the maintenance of one of the largest prison systems in the world, America continues to experience stubbornly high rates of reoffending. According to the Bureau of Justice Statistics, over 70 percent of state prisoners are rearrested within five years of release, and over 50 percent are reincarcerated for a new crime or a technical violation. These numbers suggest that incarceration, the cornerstone of U.S. criminal justice policy, is failing to achieve one of its most critical goals: the prevention of future crime.

Historically, the United States has relied on a punitive correctional model predicated on deterrence, incapacitation, and retribution. This model assumes that isolating offenders through incarceration will either scare them into compliance or physically prevent them from committing additional crimes. Yet decades of research reveal that this approach has not produced long term public safety benefits. Instead, it has fueled mass incarceration and deepened social inequities without meaningfully reducing crime or recidivism.

At its core, the American justice system is caught in a fundamental conflict between two competing aims: satisfying the emotional demand for retribution against offenders and achieving the rational goal of preventing future crime. This paper argues that the system's deep seated emphasis on retribution has created a counterproductive and harmful cycle. By prioritizing policies that fulfill a societal desire for punishment, the system has adopted an approach centered on punitive incarceration that is not only empirically ineffective but often criminogenic, ultimately increasing long term harm and creating more victims. This punitive turn is not accidental; it is rooted in a political history intertwined with racial and social control, further detaching the system's practices from evidence based strategies for public safety.

A more effective and humane path forward requires a fundamental reorientation. The justice system must be organized principally around the prevention of future harm, not the fulfillment of retributive impulses. This requires a holistic, multi-level approach that acknowledges four essential truths: the inherent limitations of imprisonment; the necessity of addressing criminogenic needs through individualized, evidence based rehabilitation; the importance of supporting the social-ecological environments that shape an individual’s reentry process; and the urgent need to reform the structural biases embedded in the justice system itself. By synthesizing empirical research from criminology, sociology, and psychology, this paper will demonstrate why the retributive model has failed and propose a systemic framework designed to achieve genuine, lasting public safety.

Historical Context

To understand the current state of recidivism research, it is essential to contextualize the policy environment from which it emerged. The latter half of the 20th century saw a significant shift in American correctional philosophy. An earlier emphasis on rehabilitation and reintegration gave way to a more punitive “tough on crime” approach (McKean & Ransford, 2004). This period was marked by the introduction of mandatory minimum sentencing and a general political climate that favored punishment over treatment. The result was a dramatic and sustained increase in the use of incarceration. From 200,000 inmates in 1973, the U.S. federal and state inmate population grew at an absurd rate, surpassing one million in 1994 and reaching 1.4 million by the mid 2000s (Liedka et al., 2006).

This era was guided by the theoretical underpinnings of deterrence and incapacitation– the ideas that the threat of severe punishment would dissuade potential criminals and that physically removing offenders from society was the surest way to prevent them from committing more crime. Rehabilitative efforts were often defunded or delegitimized, reflected in a popular sentiment that "nothing works" to reform criminals (McKean & Ransford, 2004). For example, inmate participation in substance abuse treatment programs, a critical need for a large portion of the prison population, fell from 25 percent in 1991 to just 10 percent in 1997 (McKean & Ransford, 2004). It is within this context of mass incarceration and skepticism toward rehabilitation that a new generation of research began to emerge, rigorously questioning the effectiveness of a purely punitive system and seeking to identify what, if anything, truly works to reduce recidivism.

However, this shift cannot be fully understood without examining its racialized underpinnings. Vesla Weaver’s theory of frontlash illustrates how punitive crime policy emerged as a reaction to civil rights victories in the 1960s. After legal segregation was dismantled, political elites rechanneled anxieties about social change into the arena of crime control, presenting crime as a problem synonymous with racial disorder (Weaver, 2007). Policies once explicitly designed to oppress Black Americans were recast in the "race-neutral" language of law and order, yet their effects remained deeply racialized. Felon disenfranchisement, mandatory minimums, and the war on drugs disproportionately targeted Black communities, resulting in massive overrepresentation of Black men in the prison system.

This racially coded policy shift also altered the public’s concept of justice. As Dena Gromet (2012) notes, the justice system increasingly focused on punishment of the offender, often at the expense of addressing the actual harms experienced by victims or the long term health of communities. In this punitive environment, restorative options were sidelined, and the emotional needs of victims were instrumentalized to justify harsher penalties, particularly against marginalized groups.

Thematic Review of Literature

Theme 1: The Ineffectiveness and Criminogenic Nature of Incarceration

A significant body of research finds that incarceration is largely ineffective at reducing recidivism and may even be counterproductive. This stands in direct opposition to the core principle of specific deterrence, which theorizes that the experience of punishment should discourage an individual from committing future crimes. This phenomenon is often labeled a "criminogenic" effect.

Multiple meta-analyses and large scale studies have concluded that, compared to noncustodial penalties like probation, imprisonment appears to have a "null or mildly criminogenic effect on future offending" (Loeffler & Nagin, 2022). Studies, such as those exploiting the random assignment of cases to judges or discontinuities in sentencing grids, consistently find that the experience of post-conviction imprisonment has little to no impact on the probability of recidivism. Some research even suggests that longer sentences can increase recidivism (McKean & Ransford, 2004). A comprehensive review by Berger and Scheidegger (2021) concluded that while some studies find a minor deterrent effect for longer sentences, this finding is inconsistent, and "none of the studies suggested a strong aggregate level criminogenic effect," ultimately concluding the evidence is too mixed to draw universal conclusions.

This criminogenic pattern is particularly pronounced for specific forms of incarceration and certain populations. Pretrial detention, for example, consistently yields negative post-release outcomes (Loeffler & Nagin, 2022). Individuals held in jail before trial, often without conviction, experience disruptions to employment, family stability, and housing, which increase their vulnerability to future offending. Further illustrating these dynamics, Chiricos et al. (2007) provide a compelling example of labeling theory in action: in their study of Florida's "adjudication withheld" policy, individuals formally labeled as convicted felons were 17 percent more likely to recidivate within two years than those found guilty of the same offense but not officially labeled. This finding reveals that the social stigma and structural barriers associated with the felon label, such as loss of civil rights and diminished access to employment and housing, actively contribute to recidivism. Thus, incarceration is not merely a neutral sanction; it often deepens an individual’s entanglement within the criminal justice system.

Despite this evidence, the American criminal justice system remains deeply committed to incarceration, a commitment fueled in part by emotional and political factors beyond practical outcomes. Darley and Pittman (2003) argue that people are psychologically predisposed to seek retribution when they perceive intentional harm. This moral outrage often drives public demand for harsh punishment, overshadowing rational assessments of what actually reduces harm. Politicians capitalize on these emotions, using punitive crime policies as symbols of control and moral order, frequently invoking racialized fears to galvanize support (Weaver, 2007).

This dynamic raises a fundamental tension in the concept of justice: Is justice about punishment or restoration? Research by Gromet (2012) shows that when victims participate in restorative justice processes, such as receiving apologies or restitution, they are less likely to desire harsh punitive measures. However, the current system rarely offers victims this restorative option, defaulting instead to punitive responses that satisfy emotional desires for retribution but

do little to reduce future harm. Emotional satisfaction from punishment may provide temporary relief for victims or society at large, but if incarceration increases recidivism, it ultimately perpetuates cycles of harm.

To address this contradiction, the justice system must strive to balance emotional and rational considerations. One promising approach is to institutionalize restorative justice alongside traditional retributive systems, especially for nonviolent offenses and where appropriate for violent ones. Restorative justice conferences offer victims a meaningful forum to voice their experiences and allow offenders to make amends. Studies indicate that when victims feel heard and offenders are reintegrated into their communities rather than alienated, public safety improves and recidivism declines (Gromet, 2012).

Moreover, the punitive incarceration model disproportionately harms marginalized communities, exacerbating racial inequality. Black men are incarcerated at rates up to ten times higher than white men (Katsiyannis et al., 2018) and face greater scrutiny through arrests, surveillance, and sanctions for technical violations. Weaver (2007) contends that these disparities are not incidental but reflect a systemic pattern rooted in political backlash against civil rights advances. Mass incarceration thus operates as a mechanism of racialized social control, perpetuating cycles of disadvantage rather than effectively addressing crime.

Therefore, meaningful reform must combine evidence based rehabilitative practices with a structural reckoning of the racialized foundations of the current system. Only by addressing both practical outcomes and the underlying social and political dynamics can justice evolve beyond mere punishment toward restoration and true public safety.

Theme 2: Social and Environmental Determinants of Recidivism

A successful approach to recidivism must recognize that reoffending is not solely the product of individual pathology or poor decision making. It is profoundly shaped by the social and environmental contexts to which individuals return. An offender's chances of successful reintegration are heavily influenced by the characteristics of their community and the stability of their family unit. This ecological perspective shifts the focus from an exclusive concern with individual deficits to an understanding of how external factors can either support or undermine desistance from crime.

At the community level, the field of social ecology provides a robust framework for understanding these dynamics. Research on a large sample of released Florida prisoners by Mears et al. (2008) demonstrates that two key ecological factors, resource deprivation and racial segregation, independently and interactively influence recidivism. They found that former prisoners returning to resource deprived areas, characterized by high rates of poverty and unemployment, are significantly more likely to be reconvicted of a violent crime. Such areas are often marked by a breakdown of informal social control and a lack of legitimate economic opportunities, which can make a return to crime a more viable or attractive option (Mears et al., 2008). This focus on the environment is critical because the vast majority of returning prisoners

are released into "poor inner city communities with few services and little public sympathy for their plight". These communities are often the same ones that suffer from high rates of crime and failing schools, creating a feedback loop where incarceration destabilizes communities, and disadvantaged communities, in turn, foster high rates of recidivism (McKean & Ransford, 2004; Mears et al., 2008). Liedka et al. (2006) refer to the negative impact of high rates of incarceration on neighborhoods as a form of "collateral damage," which can degrade the very community resources and social networks needed for successful reentry.

At a more intimate ecological level, the family unit is arguably the most critical institution in the reentry process. For most former prisoners, relationships with family members are essential to successful reintegration. A meta-analysis by Katsiyannis et al. (2018) identified "family criminality" and "family rearing" as statistically significant predictors of adult recidivism, providing quantitative support for the importance of family context. Upon release, family members are the primary source of tangible and emotional support. A study by Naser and Visher (2006) of families of men returning to Chicago found that 76 percent of the men lived with a family member after release and 83 percent received financial support from them. This support serves as a crucial "buffering agent" against the challenges of reentry (Naser & Visher, 2006).

However, this reliance on family comes at a great cost. Incarceration places immense strain on the family unit, with 75 percent of relatives reporting that prisons were too far away and over half citing the high cost of phone calls as a barrier to maintaining contact (Naser & Visher, 2006). The median cost for a single visit was $55, and families spent a median of $50 on phone calls in the final month of a relative's sentence, substantial burdens for households already in precarious financial situations (Naser & Visher, 2006). Upon release, 83 percent of family members provided financial assistance, yet over half found it "pretty or very hard" to do so, and nearly a third experienced direct financial hardship (Naser & Visher, 2006). This support is often given under emotional strain, with nearly 10 percent of families reporting stress or anxiety tied to their loved one's return (Naser & Visher, 2006). Such persistent economic and emotional pressures on the very networks most critical to reentry reveal a systemic vulnerability: the justice system depends heavily on families to act as de facto reentry programs while offering them almost no support. Any effective recidivism strategy must therefore address not only the needs of the individual offender but also invest in the family and community ecosystems essential to a successful transition, breaking the cycle of disadvantage that incarceration perpetuates.

Theme 3: Systemic Bias and Inequity in the Justice System

While social and environmental factors create the external context for reentry, the criminal justice system itself is not a neutral actor. Its internal processes, from initial labeling to post-release supervision, are characterized by biases, inequities, and perverse incentives that actively contribute to the cycle of recidivism. A successful approach must therefore include a critique and reform of the system itself.

First, the system produces demonstrably different outcomes based on race and gender. Racial minorities, particularly Black men, are substantially overrepresented at every stage of the correctional system. A 2018 meta-analysis noted that Black males have had imprisonment rates 3.8 to 10.5 times higher than White men (Katsiyannis et al., 2018). This disparity is not merely a reflection of offending rates. Mears et al. (2008) found that young, nonwhite males have disproportionately higher rates of recidivism for violent and drug offenses, but they caution that this may reflect "differential law enforcement" practices. This suggests a systemic bias in how behavior is surveilled and sanctioned.

The impact of the system is also inequitable. The formal label of "convicted felon," a status imposed by the state, is a powerful criminogenic force. Chiricos et al. (2007) found that being formally adjudicated as a felon significantly increases the odds of recidivism by 17 percent compared to those found guilty of the same offense but for whom adjudication is withheld. The label creates lifelong structural barriers to employment, housing, and civil participation. Paradoxically Chiricos et al. (2007) found the labeling effect to be stronger for whites and for women. They theorize that because crime is stereotypically associated with minority men, the stigma of a felony conviction is more socially damaging for those who do not fit the stereotype. Similarly, a meta-analysis by Van Vugt et al. (2011) found that the link between an individual's moral development and their likelihood of recidivism was stronger for females than for males, suggesting that the system's response to individual characteristics is not gender neutral.

Perhaps the most compelling evidence of systemic failure is the mechanism of the "revolving door." Research by Harding et al. (2017) leveraging a natural experiment in Michigan found that being sentenced to prison, as opposed to probation, causes a dramatic increase in the probability of future imprisonment. Critically, this effect is not driven by an increase in new serious crimes. Instead, it is driven almost entirely by imprisonment for technical violations of parole. Parole supervision is typically far more stringent than probation, and this heightened surveillance increases the detection and punishment of low level rule breaking, such as moving without permission or failing a drug test. This behavior would not otherwise result in a prison sentence for someone who had not already been to prison (Harding et al., 2017). The system, therefore, creates its own recidivism crisis by subjecting one class of people (parolees) to a level of scrutiny that makes their return to prison almost inevitable, not because they are necessarily more dangerous, but because they are caught in a web of escalating surveillance and punishment. This demonstrates that a significant portion of recidivism is not a failure of the individual, but a direct and predictable outcome of the justice system's own design.

Theme 4: The Promise of Alternative Models-Rehabilitation and Reentry Programming

In stark contrast to the failures of punitive incarceration, a wealth of evidence demonstrates that well designed, evidence based rehabilitative programs can significantly reduce recidivism. Latessa & Lowenkamp (2006) has established a clear set of principles for effective intervention that, when properly implemented, yield consistent positive results. Their core

principles of effective intervention are *risk, need, treatment, and fidelity*. The Risk Principle states that intensive treatment should be reserved for high risk offenders, as placing low risk individuals in these programs can be counterproductive and even increase their likelihood of reoffending by disrupting prosocial networks and exposing them to antisocial peers. The Need Principle dictates that programs must target known "criminogenic needs,"dynamic risk factors highly correlated with criminal conduct, such as antisocial attitudes, substance abuse, and poor self control rather than noncriminogenic factors like general self esteem (Latessa & Lowenkamp, 2006; Hanson & Harris, 2000). The Treatment Principle emphasizes the use of behavioral models, such as Cognitive Behavioral Therapy (CBT), which are action oriented and teach offenders new prosocial skills. Finally, the Fidelity Principle stresses that programs must be delivered as designed by well trained staff to be effective.

Studies show that programs adhering to these principles work. In a large scale study of Ohio correctional programs, high risk offenders placed in effective residential programs saw significant reductions in recidivism, with some programs achieving a 30% or higher reduction, while low risk offenders placed in the same programs saw their recidivism rates increase (Latessa & Lowenkamp, 2006). Specific program types have also proven effective. Substance abuse treatment is critical, as an estimated 70 to 85 percent of inmates need it (McKean & Ransford, 2004). Additionally, mandatory drug treatment programs and drug courts– specialized court programs designed to handle cases involving individuals with substance use disorders– can reduce recidivism by about 31 to 32 percent (McKean & Ransford, 2004). Educational and vocational programs are similarly vital, given that 41 percent of state inmates lack a high school degree (McKean & Ransford, 2004). Participation in prison education has been shown to reduce recidivism by as much as 29 percent (McKean & Ransford, 2004). Employment programs that provide job preparedness and placement assistance, like New York's ComALERT program, report recidivism rates of 17 percent compared to 41 percent for nonparticipants (McKean & Ransford, 2004). The consistent thread is that interventions that address the root causes of criminal behavior and provide pathways to prosocial life are far more effective than punishment alone. The most successful settings for reducing recidivism are those in which rehabilitative programming is emphasized.

A modern, evidence based approach to recidivism can be significantly enhanced by leveraging new technologies to deliver and personalize rehabilitative interventions. As outlined by Sulasa and Kumar (2024), technology offers innovative tools for addressing the specific criminogenic needs of offenders, particularly juveniles, in ways that were previously not possible. This is not about replacing human intervention but augmenting it with powerful, data driven methods. Several key technologies show significant promise for rehabilitation:

- Virtual Reality (VR): VR can create immersive, simulated environments where offenders can safely practice critical skills. For instance, VR can be used to model real life scenarios that might trigger aggressive behavior or poor decision making,

allowing individuals to develop and rehearse coping skills and social skills in a controlled setting before facing these challenges in the community.

- Artificial Intelligence (AI): AI driven algorithms can analyze complex datasets to identify patterns, predict individual outcomes, and help tailor intervention strategies to an offender's unique risk factors and changing needs over time. This allows for a dynamic approach to risk management, moving beyond static assessments and adapting to an individual's progress.
- Wearable Devices and Biometric Monitoring: Wearables can track physiological indicators like heart rate and stress levels, providing real time data on an individual's emotional state. This information can enable timely interventions and personalized support, helping an individual manage distress or anger before it escalates into problematic behavior.

Ultimately, the combined message from Latessa & Lowenkamp's intervention framework and advancements in technology is clear: meaningful reductions in recidivism come not from where an offender serves their sentence, but from what is done during and after that sentence to address the factors driving criminal behavior.

Counterarguments and Rebuttals

Proponents of punitive criminal justice policies often raise two primary arguments: that harsh sanctions are a necessary deterrent to crime and that long term incapacitation is essential for public safety. While these arguments contain logic, they crumble when subjected to scrutiny and a long term view of public safety and sustainability.

The core idea is that longer and more certain prison sentences can reduce crime both by discouraging potential offenders and by physically preventing incarcerated individuals from committing new crimes in the community. Recent, methodologically rigorous research from the U.S. Sentencing Commission (USSC) provides the strongest contemporary evidence for this perspective. In a large-scale 2022 study examining over 32,000 federal offenders, the Commission found a statistically significant preventative effect for longer lengths of incarceration. Specifically, the report found that offenders sentenced to more than 120 months had odds of recidivism that were approximately 29 percent lower than a matched group of offenders who received shorter sentences, and those sentenced to more than 60 months up to 120 months had odds of recidivism that were approximately 18 percent lower than their counterparts. This preventative effect was found to be stable across different cohorts, as the results were "almost identical" to a previous study of offenders released in 2005, and offenders serving these longer sentences also took longer to recidivate upon release. This suggests that longer sentences may have a lasting impact on post-release behavior. From this perspective, longer prison sentences are a valuable and effective tool for enhancing public safety by reducing the likelihood of future crime among a significant portion of the offender population.

However, this argument is critically shortsighted when applied as a broad strategy. First, this incapacitation benefit applies only to a minority of the prison population; the same USSC study found no effect for the 59% of offenders who received sentences under 60 months. Second, the reality is that 97 percent of all state prisoners will eventually be released back into society (Lise McKean and Charles Ransford, 2004). Therefore, public safety cannot be measured solely by the crimes prevented while someone is locked away; it must be measured by the likelihood that they will offend again upon release. A strategy focused only on incapacitation is a temporary solution that fails to address the underlying drivers of criminal behavior. Moreover, the USSC study is limited to federal offenders, who represent a small and atypical subset of the overall U.S. incarcerated population. Federal offenders are often convicted of white collar crimes, immigration violations, or large scale drug offenses, which differ significantly from the state level offenses that make up the bulk of the prison system. This study also cannot fully eliminate unmeasured variables, such as differences in individual motivation, access to reentry programs, or community support upon release, which can heavily influence recidivism.

One cannot ignore the deeply emotional dimensions of justice, particularly for victims and their families. Crime is not merely a violation of law—it is a profound disruption of lives, relationships, and one's sense of safety in the world. For many victims, the legal outcome carries immense symbolic weight, representing not just a legal resolution but a societal acknowledgment of their suffering. A central argument in favor of punitive policy is that it provides a sense of closure, enabling victims to feel that justice has been served through clear and decisive offender accountability. Without such acknowledgment, victims may feel invisible or that their pain is being minimized. In the immediate aftermath of a crime, when emotions are raw and heightened, punishment can appear to be the most tangible and straightforward way to validate the harm done.

Public emotions toward rehabilitation often shift depending on the severity and violent nature of the crime. People are generally more open to rehabilitative approaches in cases involving nonviolent or accidental offenses but tend to grow less tolerant as crimes become more severe, intentional, or violent. The distinction between harm caused by negligence versus harm caused by malicious intent plays a significant role in shaping these attitudes. Rehabilitation may be embraced for mistakes or unintended consequences, but in cases where harm is deliberate, public sentiment often leans heavily toward retribution. This emotional divide fuels ongoing debates over who deserves second chances and under what circumstances. This raises difficult questions at the core of criminal justice policy: Should the ultimate goal be to satisfy victims' immediate emotional needs for punishment, or to focus primarily on preventing future crime to protect society as a whole? For offenders with violent tendencies, particularly those who show no remorse, must long term institutionalization be the default in order to prevent repeated harm? When, if ever, should such individuals be granted second chances? And how can the justice system fairly and accurately assess an offender's capacity for rehabilitation or their risk of

recidivism? These questions have no easy answers, but they remain central to discussions about balancing public safety with the principles of reform.

It is important to recognize that closure through punishment, while a valid emotional need, is not universally effective. Some victims discover that even the harshest sentence does not bring peace, particularly when emotional wounds run deep. In such cases, the pursuit of punitive closure can lead to drawn out legal battles, repeated appeals, and continued public attention, processes that can retraumatize rather than heal. Furthermore, an exclusive focus on punishment risks obscuring alternative forms of justice, such as truth telling, apologies, or reparations, that may offer victims a deeper and more personal sense of resolution.

Darley and Pittman (2003) reiterate this, cautioning that emotional satisfaction and justice do not always align. The public's desire for harsh punishment is often driven less by evidence based reasoning and more by moral outrage, a visceral reaction to perceived wrongdoing. As a result, justice systems risk prioritizing punitive responses that appease immediate anger but fail to address the root causes of crime. This dynamic can lead to counterproductive outcomes such as excessive incarceration, overcrowded prisons, and the neglect of rehabilitative or restorative strategies that might more effectively reduce recidivism and rebuild communities.

Dena Gromet's (2012) research challenges the assumption that victims universally demand the harshest possible penalties. Her studies found that justice, as experienced by many victims, is less about inflicting maximum punishment and more about achieving emotional closure, validation, and a restored moral balance. Gromet compared victims' responses to traditional retributive measures, such as lengthy prison sentences, with their reactions to restorative alternatives—apologies, restitution, or facilitated dialogue between victim and offender. Contrary to common belief, many victims reported greater emotional satisfaction, healing, and perceptions of fairness when engaged in restorative justice processes, even when these resulted in lighter sentences.

This suggests that recognition, accountability, and the opportunity for meaningful interaction often outweigh simple retribution in fulfilling victims' psychological needs. When victims have the opportunity to express their pain, ask questions, and receive acknowledgment of the harm done, they often feel more genuinely heard and respected. Such interactions can foster forgiveness, though not leniency, when offenders demonstrate sincere remorse and a willingness to make amends. Gromet found that this shift from punitive retribution to constructive accountability can lead victims to prefer solutions that focus on repairing harm rather than simply imposing suffering.

One of the major shortcomings of the traditional punitive system is that it removes the victim from the center of the process. The trial becomes a contest between the state and the offender, reducing the victim's role to that of a witness. Restorative justice seeks to reverse this dynamic by recentering the victim's voice. Victims are invited to articulate the personal impact of the crime, ask questions directly, and help shape reparative outcomes that are meaningful to

them. This sense of agency is critical to emotional healing, as it replaces feelings of helplessness with empowerment.

Restorative justice also addresses practical needs alongside emotional ones. Victims often gain insight into why the crime occurred, which can help resolve lingering uncertainty and reduce trauma. In some cases, they also receive tangible restitution that reinforces the perception that justice has been done. Importantly, Gromet's research shows that when victims perceive offenders as genuinely apologetic and willing to make amends, they are less likely to support excessively punitive sanctions and more likely to endorse solutions that foster offender reintegration into society.

The implications for policy are significant. If victims' emotional needs can be met through restorative practices, there is less justification for defaulting to harsh punitive measures that fuel mass incarceration without necessarily improving public safety. A hybrid system that incorporates restorative approaches alongside traditional measures could provide a more balanced form of justice—one that honors victims' experiences, holds offenders meaningfully accountable, and reduces the long term risks of recidivism.

However, these benefits do not erase the legitimate concerns about rehabilitation's limits, especially in cases involving individuals whose violent traits pose ongoing risks. Determining when someone poses too great a danger to be reintegrated, and how to make that determination fairly, remains an unsolved challenge. Public attitudes toward rehabilitation will likely continue to fluctuate depending on the crime's severity, the offender's perceived remorse, and the community's tolerance for risk. The justice system's task is to navigate these complexities without defaulting to policies driven solely by moral outrage or short term emotional satisfaction. Admittedly, this task of finding a fair balance between community safety and the potential for offender rehabilitation may be a task with no definitive solution.

Ultimately, justice must walk a difficult line: acknowledging and validating the emotional reality of victims, protecting public safety, and promoting policies that reduce future harm. Restorative justice offers promising tools for meeting these goals, but it is not a universal solution. In some cases, long term incapacitation may still be necessary to protect society. In others, the most effective path forward may involve reconciliation, reparation, and reintegration. The enduring challenge lies in discerning which approach serves both justice and safety in each unique case—and in doing so without losing sight of the human lives, both victim and offender, that hang in the balance.

Concluding Argument

The evidence reviewed in this paper strongly suggests that the prevailing punitive approach to criminal justice in the United States has largely failed to reduce recidivism. Decades of data consistently show that incarceration is an ineffective tool for promoting long term behavioral change and that its application can often be counterproductive. While proponents of punishment argue for deterrence and incapacitation, the research reveals a more complex reality:

prison can reinforce the very behaviors it seeks to eliminate, leaving individuals more alienated, stigmatized, and prone to reoffending.

Moreover, the American justice system's reliance on punishment, isolation, and surveillance is not simply the result of neutral policy choices. As Weaver (2007) demonstrates, this punitive model is deeply entangled with the nation's history of racial inequality. The rise of mass incarceration was not a data driven response to crime but a political reaction to civil rights gains. It was a deliberate strategy to reassert social control through ostensibly race-neutral mechanisms like sentencing reform and "law and order" policies. These structural realities have disproportionately harmed communities of color, deepening cycles of disadvantage and criminalization that extend far beyond individual offenders.

At the same time, justice in America has often been framed through an emotional lens. Research from Gromet (2012) and Darley and Pittman (2003) reveals that the justice system is designed not just to control behavior but to satisfy emotional demands for retribution. Moral outrage is a powerful psychological force, and punitive sanctions frequently serve as a societal outlet for this impulse. However, prioritizing emotional satisfaction over rational policy has led to greater long term harm. While punishment may offer immediate relief to victims or the public, it often exacerbates the very conditions that foster crime in the first place.

Over the past several decades, our understanding of recidivism has evolved from viewing it as a personal moral failure to recognizing it as a systemic outcome of structural inequality, legal stigma, and community disinvestment. Early policies that prioritized punishment and surveillance have failed to deliver sustained reductions in recidivism, leaving policymakers to confront the unintended consequences of a punishment driven system. This reality points toward a more promising and humane approach that is rooted in a new mindset that is problem solving, ecological, and systemically aware.

First, this mindset reframes recidivism not as a simple moral failure requiring retribution but as a complex behavioral problem to be understood and addressed. It focuses on identifying and analyzing the underlying drivers of reoffending—such as antisocial attitudes, substance abuse, and trauma—instead of focusing exclusively on the act of punishment. For certain offender populations, factors like deviant impulses, cognitive distortions, and antisocial orientation are far more predictive of recidivism than issues like denial or a lack of remorse. A problem solving mindset uses this evidence to diagnose the specific risks and needs of each individual, crafting interventions that target root causes rather than symptoms. This shift does not remove accountability; rather, it broadens the inquiry to ask why an offense occurred and how best to prevent future harm. While "tough on crime" policies often stem from an intuitive desire for retribution and a political need for simple, decisive messaging, an evidence based approach moves beyond this intuition to analyze problems at their roots.

Second, this new approach requires an ecological mindset that looks beyond the individual to their social environment. It recognizes that an individual's path after release is profoundly shaped by the context to which they return. Family plays a dual role, acting as a

crucial "buffering agent" by providing housing, emotional support, and financial assistance while simultaneously bearing immense strain and hardship throughout the reentry process. An ecological mindset also understands that community characteristics—such as resource deprivation, racial segregation, and a lack of social capital—create environments that can undermine even the most determined individual's efforts to desist from crime. High rates of incarceration destabilize neighborhoods, remove working age adults from the social fabric, and degrade the informal social controls that promote public safety. This viewpoint reframes individual outcomes as interconnected with the broader health and stability of families and communities, not as isolated personal choices.

Finally, a more sophisticated approach requires a systemically aware mindset. This involves turning a critical lens on the justice system itself, questioning how its own practices and structures actively contribute to the very outcomes it seeks to prevent. For example, the mechanics of the "revolving door" must be dissected: research shows that prison sentences often increase the likelihood of future incarceration, not because of new serious crimes, but because of technical violations of a stricter and more punitive parole system (Harding et al., 2017). Rather than viewing recidivism as a failure of personal willpower, this perspective acknowledges that systemic incentives and surveillance practices create cycles of punishment that are baked into the system's design. Moreover, racial disparities in arrests, charging, sentencing, and parole violations reveal a justice process that is anything but neutral.

A modern, humane justice system must therefore replace the outdated "punishment equals justice" mindset with one that prioritizes fairness, restoration, and equity. This requires targeted rehabilitation programs that address criminogenic needs, comprehensive family and community support systems, and reforms that confront the racial disparities embedded in the system's operation. Incorporating restorative justice practices offers a pathway to meet victims' emotional needs without defaulting to the criminogenic trap of mass incarceration. Restorative justice provides victims with an active role in the process while fostering offender accountability and community healing, a model that aligns both emotional and rational considerations.

Ultimately, creating a safer society does not require a choice between being "tough" or "soft" on crime; it requires being smart. It means replacing a failed, emotionally driven, and historically freighted system with one that is evidence based, humane, and economically sustainable. The path forward lies in dismantling the structural barriers to success and investing in the human potential for change. By integrating targeted rehabilitation, robust community support, and restorative principles, the United States can finally move beyond the revolving door of recidivism and build a justice system that fosters true safety and equity for all.

Works Cited

Berger, Elizabeth, and Kent S. Scheidegger. Sentence Length and Recidivism: A Review of the Research. Criminal Justice Legal Foundation, 2021.

Chiricos, Ted, et al. "The Labeling of Convicted Felons and Its Consequences for Recidivism." Criminology, vol. 45, no. 3, 2007.

Darley, John M., and Thane S. Pittman. "The Psychology of Compensatory and Retributive Justice." Personality and Social Psychology Review, vol. 7, no. 4, 2003.

Gromet, Dena M. "Restoring the Victim: Emotional Reactions, Justice Beliefs, and Support for Reparation and Punishment." Critical Criminology, vol. 20, no. 1, 2012.

Hanson, R. Karl, and Andrew J. R. Harris. "Where Should We Intervene? Dynamic Predictors of Sexual Offense Recidivism." Criminal Justice and Behavior, vol. 27, no. 6, 2000.

Harding, David J., et al. "Short- and Long-Term Effects of Imprisonment on Future Felony Convictions and Prison Admissions." Proceedings of the National Academy of Sciences, vol. 114, no. 41, 2017.

Katsiyannis, Antonis, et al. "Adult Recidivism in United States: A Meta-Analysis 1994-2015." J. Child Fam Stud, 2017.

Latessa, Edward J., and Christopher Lowenkamp. "What Works in Reducing Recidivism." University of St. Thomas Law Journal, vol. 3, no. 3, 2006.

Liedka, Raymond V., et al. "The Crime-Control Effect of Incarceration: Does Scale Matter?" vol. 5, no. 2, 2006.

Loeffler, Charles E., and Daniel S. Nagin. "The Impact of Incarceration on Recidivism." Annual Review of Criminology, vol. 5, 2021.

McKean, Lise, and Charles Ransford. Current Strategies for Reducing Recidivism. Center for Impact Research, 2004.

Naser, Rebecca L., and Christy A. Visher. "Family Members' Experiences with Incarceration and Reentry." Western Criminology Review, vol. 7, no. 2, 2006.

Sulasa, T. J., and Ramesh Kumar. "Role of Modern Technology for Rehabilitation in Mitigation of Juvenile Recidivism: A Critical Appraisal." Pakistan Journal of Life and Social Sciences, vol. 22, no. 2, 2024.

United States Sentencing Commission. Length of Incarceration and Recidivism. 2022.

Van Vugt, Eveline, et al. "Moral Development and Recidivism: A Meta-Analysis." International Journal of Offender Therapy and Comparative Criminology, vol. 55, no. 8, 2011.

Weaver, Vesla M. "Frontlash: Race and the Development of Punitive Crime Policy." Studies in American Political Development, vol. 21, 2007.

A Review Analysis of AI in Ceramics Business Parth Kajaria

Abstract

Technologies that provide increased clarity, speed, and precision are now supplementing several long-standing methods, such as hand-sketching designs, performing manual quality checks, or depending on intuition for color and glaze results. In order to examine how visualization software, AI-assisted design systems, virtual showrooms, and data-based manufacturing tools are transforming the industry, this review compiles recent research and industry examples. Additionally, designers are starting to use generative AI to experiment with surface effects, textures, and patterns that would be challenging or time-consuming to produce by hand. Thus, we need to safeguard the significance of human craftsmanship and creative judgment while also emphasizing that digital and AI technologies provide genuine prospects to increase creativity, efficiency, and customer interaction in ceramics. The entire value chain is being optimized by this technological change. Generative models like GANs and Stable Diffusion are speeding up the generation of patterns and forms in design. This enables designers to produce a variety of alternatives from a single sketch, and reinterpreting traditional themes for modern uses. AI can also play a role in sustainability. It can reduce material waste and leads to the optimization of kiln conditions to conserve energy. Predictive analytics helps business operations by improving trend predictions and offering intelligent, precise inventory and logistics management that reduces overproduction. The shift comes with its own set of difficulties. For example, the requirement for well-labeled, high-quality datasets and its practicality. Then the difficulty of integrating AI tools into conventional, frequently manual operations. This paper highlights another challenge which is the replication challenge. The most polished algorithms are still unable to properly capture the distinctive and unique traits inherent in many of the artisan works. Therefore, an emphasis on hybrid design systems where the artisan's skill and vision can effectively direct computing power is necessary for future success. These results of this paper demonstrate how AI and digital visualization are revolutionizing the ceramic value chain while also increasing operational effectiveness and opening the door to a new age of highly customized, data-driven design.

Keywords

Machine Learning; Computer Vision; Ceramic Design Innovation; Pattern Generation; Manufacturing & Sustainability

Introduction

The ceramics industry is humanity's one of the oldest and most enduring industries. Now there is a swift convergence toward digital technology, artificial intelligence (AI), and sophisticated visualization tools in its space. Works range from practical pots to architectural tiles. That has depended on highly proficient hand methods, instinctive assessment of glaze chemistry, and laborious trial-and-error. The entire value of the chain can be optimized and this

paper examines how contemporary technological interventions are altering the design, production, marketing, and customer engagement processes. Physical prototyping and hand-sketching in the early stages of ceramic design compelled designers to make important choices regarding form and color based on experience. This frequently resulted in expensive physical mock-ups. Now, a lot of this early-stage ambiguity has been eliminated because of technology like 3D modeling programs (like Blender and Fusion 360) and advanced visualization software. They enable the production of amazingly realistic digital prototypes that can be digitally altered and glazed. Generative AI can go beyond basic visualization and actively supports creativity. It is a more disruptive force than just AI. Designers can use models like GANs and diffusion models that help them design rules and receive unique and complex patterns. This speeds up the iterative design process and allows the exploration of ornamentation that would be very difficult or time-consuming to create by hand. Previously, quality control relied on error-prone manual inspection which led to significant material loss. Digital technologies are being implemented deeply into the production workflow to solve long-standing concerns of inconsistency, waste, and energy usage. Defect detection is now possible in real time because of the development of computer vision systems that use deep learning models, such as Convolutional Neural Networks. There are quantifiable environmental benefits in implementing technologies. Predictive analytics helps improve demand forecasting and also streamlines intelligent inventory and logistics management. The marketing, sales, and experience of ceramic products have also been completely transformed by the digital age. The traditional model is now supplemented by virtual showrooms and Augmented Reality (AR) solutions which enables customers to virtually place products in their homes. This eliminates visual uncertainty, boosts confidence, and shortens the sales cycle. Many algorithms in machine learning help examine consumer behavior and trends to predict changes in color schemes and formats. This way manufacturers can proactively match their product development with future market demand.

This review is predicated on the fundamental idea that digital transformation needs to be approached critically, despite these strong advantages in terms of speed, efficiency, and scale. Artificial intelligence (AI) is capable of analyzing preferences, forecasting kiln outcomes, and producing intricate patterns, but it is unable to completely mimic the inadvertent nuances, tactile experience, or innate creative judgment that characterize human craftsmanship and the distinctive quality of handcrafted ceramic products (Vancasso Tableware, 2019). Thus, this paper highlights the critical need for approaches that fully embrace the productivity advantages provided by digital and AI technology while protecting the importance of human artistry and creative control. This review helps provide a thorough overview of the current state of digital adoption in ceramics by gathering recent research and industry applications. It highlights both the real opportunities for innovation and the crucial issues pertaining to data, integration, and maintaining the irreplaceable value of the artisan.

Methodology

The narrative style review is used as the methodology in this study. It involves the methodical gathering, synthesis, and critical analysis of current scholarly literature, industry reports, and real-world case studies related to the use of artificial intelligence in the ceramics sector. This study aims to combine findings from several sources to identify prevalent themes, technical trends, advantages, difficulties, and research gaps. It focuses on interdisciplinary subjects such as AI-driven ceramic design and manufacturing, innovation, design, production, sustainability, and consumer engagement.

Themes

- **Customer Experience Enhancement**

Customer Experience Enhancement- investigates how immersive technology, AI, and visualization tools enhance consumer confidence, engagement, customisation, and decision-making throughout the buying process.

  - **Visualisation tools**

Focus on the use of photorealistic rendering technologies, augmented reality (AR), and 3D modeling software in digital prototype, product display, and pre-production validation.

  - **Design Recommendations**

Examine AI-assisted decision-making systems that make recommendations for forms, patterns, colors, and finishes based on customer preferences, market data, and past design trends

  - **Virtual Showrooms**

Examine AI-assisted decision-making systems that make recommendations for forms, patterns, colors, and finishes based on customer preferences, market data, and past design trends.

- **Design and Product Innovation**
  - **Generative AI for tile design and patterns**

It explores the creation of intricate designs inspired by culture and scalable ceramic patterns using GANs, diffusion models, and text-to-image systems.

  - **AI in sustainable design : minimising waste**

Reducing Waste Evaluations AI uses in material conservation, energy-efficient kiln control, process optimization, and defect identification.

  - **Predictive analysis for trend forecasting**

Predictive analysis for trend forecasting looks at how machine learning algorithms predict surface preferences, formats, color patterns, and consumer demand.

- **Manufacturing and Supply chain optimization**

- **AI driven quality control**

AI-enabled production planning, logistics optimization, and real-time operational efficiency are the main topics of manufacturing and supply chain optimization.

- **Smart Inventory and logistic management**

Predictive inventory solutions that lower overstocking, stockouts, and transportation risks are reviewed in Smart Inventory and Logistics Management.

- **AI driven quality control**

Analyzes computer vision systems used for real-time defect detection and automated inspection.

- **Marketing, Advertising and Sales**

Artificial intelligence uses data-driven decision-making and individualized client involvement.
The real-time optimization reduces the inefficiencies associated with conventional blanket marketing strategies while increasing return on advertising investment. AI-powered visual analytics also assist in determining which product layouts, graphics, and design narratives result in the greatest level of interaction.

Chatbots and virtual sales assistants are examples of conversational agents, which instantly respond to consumer questions. They deliver product advice, assist with configuration, pricing, and order tracking. Scalable customer service and improved accessibility is achieved.

- **Benefits and Challenges**

AI improves efficiency from an operational standpoint by lowering design trial-and-error, minimizing production errors, maximizing inventory levels, and enhancing supply chain coordination. By facilitating quick experimentation with intricate patterns, textures, and forms while permitting mass customization, generative AI creatively enhances design possibilities. But integrating AI also comes with a lot of difficulties. Effective model performance requires high-quality, well-labeled information, but in conventional ceramic manufacturing settings, these kinds of data are either hard to come by or inconsistent. Legacy system technical integration can be difficult and expensive, necessitating organizational transformation and staff retraining. Most importantly, the subtle workmanship, tactile sensitivity, and creative intuition that characterize handmade ceramic production are still difficult for AI systems to mimic. Hybrid approaches that strike a balance between computer efficiency and human creativity and cultural authenticity are essential for the successful future of AI in pottery.

Inclusion and Exclusion Criteria

Inclusion Criteria

- Academic studies
- Peer-reviewed journal publications
- Case studies and reports done by reputable ceramic producers and technology suppliers
- Publications in English language
- 2019 and 2025 time-frame for studies
- Research on ceramics design, production, sustainability, customer experience, or business optimization

Exclusion Criteria

- Research unrelated to ceramics, tiles, or the production of ceramics
- Articles with unclear methodology or unreliable sources
- Content that is only promotional or speculative and lacks depth of analysis
- Research that just looks at digital tools that aren't AI-driven or intelligent

Discussion

- **Customer Experience Enhancement**
  - **Visualisation tools**

By using a set of powerful visualization tools, any ceramics company may greatly enhance the customer experience in design and decision-making. By creating remarkably convincing digital prototypes of ceramic pieces using 3D modeling applications like Blender and Fusion 360, customers may examine shapes, textures, and glaze details well in advance of production. Customers may preview how ceramic items like planters or tiles would seem in their own environments using tools like SketchUp, RoomSketcher, and Homestyler, which increases confidence and reduces uncertainty. Thanks to augmented reality (AR) solutions like Shopify AR, Adobe Aero, or a custom Unity and Vuforia configuration, customers can use their phones to realistically place ceramic objects in their homes. In terms of marketing, Canva, Adobe Dimension, and Procreate facilitate the production of eye-catching images and mood boards with hand-painted glazing themes and photorealistic settings that appeal to consumers. Lastly, internal process planning is strengthened with technologies like Miro, Notion, and Trello, which guarantee efficient processes like kiln cycles and glazing schedules, increasing delivery speed and reliability. From the customer's initial engagement until the time a piece is installed in their house, these features collectively make the ceramic purchasing process more clear, pleasurable, and engaging.

Design Recommendations

AI has the potential to completely transform the ceramics industry. AI-powered design tools allow ceramic makers to explore and produce highly unique collections. Beyond typical hand drawing, they are able to produce distinctive forms, glaze patterns, and surface textures. Machine learning algorithms can analyze patterns in color, form, and customer preferences to recommend designs that are more likely to succeed in the marketplace. AI-driven visualization

technologies can enable realistic simulations of glazes, crackling effects, and hand-painted styles, providing customers with a preview of the final product before manufacture. By predicting kiln results, suggesting ideal firing temperatures, reducing errors, and improving uniformity, AI can aid in production setup optimization. The ceramics industry can provide more individualized goods, cut down on trial and error, shorten development cycles, and give clients a more engaging and assured purchasing experience by incorporating AI into both creative and operational operations. By producing distinctive shapes, patterns, and textures and minimizing human trial-and-error, AI improves ceramic design. In order to suggest designs that are more likely to sell, it examines consumer preferences and market trends. Customers may choose products with confidence before they are manufactured thanks to AI visualization tools that produce realistic previews of glazes and finishes. AI can improve uniformity and cut waste in manufacturing by forecasting kiln results and optimizing fire parameters. In general, AI improves the speed, creativity, personalization, and customer-friendliness of ceramic creation.

Virtual Showrooms

Digital and immersive technologies that enable clients to explore items more conveniently and interactively, several ceramics firms are revolutionizing the traditional showroom experience. Buyers can view high-quality 3D models, simulate products in real-life areas, and even experience augmented or virtual reality environments in virtual showrooms, which extend access beyond physical venues and make the decision process more interesting and educational.

For instance, buyers can thoroughly examine tiles, slabs, and sanitaryware in RAK Ceramics' 360-degree virtual showroom. With flagship shops in strategic areas like Jeddah and Riyadh in Saudi Arabia and a design hub in Milan, Italy, RAK Ceramics' headquarters are located in Ras Al Khaimah, United Arab Emirates, and its products are accessible to customers worldwide.Visitors can use virtual reality headsets to see the showroom, rotate and zoom in on objects, and take virtual measurements. Additionally, the business has grown into a metaverse setting, offering a completely digital area for collection browsing. This can enhance engagement, decision-making confidence, and brand perception as it would allow customers worldwide to interact with the brand in an immersive manner.

- **Design and Product Innovation**

  - **Generative AI for tile design and patterns**

Recent studies have shown how generative AI can revolutionize tile and ceramic design by facilitating large-scale customisation and innovation. For example, Jianfeng Lu, Xinyi Liu, Mengtao Shi, Chen Cui, and Mahmoud Emam (2023) created a sketch-based GAN model that could transform hand-drawn sketches into a variety of intricate representations of ceramic tiles. Similar to this, Xiaoyan Wang, Zhiling Zhong, Lanlan Wen, and Tao Liu (2024) replicated classical ornamentation patterns in contemporary pottery designs using generative AI techniques

like Stable Diffusion and LoRA. Miray Gür and associates (2024) investigated reinterpreting Iznik tile compositions for architectural purposes using AI technologies like DALL-E and MidJourney. These works show how AI bridges classical artistry with contemporary computational design by speeding up the creative process and enabling designers to experiment with complex patterns, colors, and textures that would be difficult or time-consuming to make manually. Works by Lu, Liu, Shi, Cui, and Emam (2023) shows how generative AI can change tile design by employing a GAN-based model to turn hand-drawn designs into a variety of realistic ceramic tile pictures. To guarantee that the generated tiles match both the input sketch and realistic textures, their approach employs a sketch extraction network and a sketch-based generation network with cross-domain adversarial loss. The concept allows designers to quickly and easily experiment with new patterns and textures while retaining creative control by generating several design variations from a single sketch by perturbing the latent space.

- **AI in sustainable design : minimising waste**

By reducing production waste and enhancing process efficiency, AI-driven methods are rapidly helping ceramics become more sustainable. Recent research has employed deep learning models, such as convolutional neural networks (CNNs), for real-time defect detection on ceramic surfaces, enabling producers to discover and eliminate defective parts early and minimize material loss. In order to improve quality control and reduce scrap, generative models like DCGAN have been used to create realistic ceramic textures while also employing Faster-R-CNN to detect faults with high accuracy. By automatically adjusting process variables, reinforcement learning techniques have also been utilized to optimize manufacturing parameters in ceramic sintering processes, lowering waste and energy consumption. The study by Cumbajin et al. (2024) shows how AI can improve production efficiency and reduce material waste to improve sustainability in ceramic manufacturing. The researchers created a deep learning-based computer vision system to instantly identify surface flaws in ceramic objects. The technique minimizes scrap and rework, saving raw materials and energy, by detecting defective pieces early in the production process before glazing, fire, or packaging. The AI-driven inspection tool reduces the need for manual labor while supporting environmental goals by automating quality control and achieving high accuracy (~98%) and F1-score (~97.3%). This strategy, which is integrated into the manufacturing chain of a partner Portuguese company, is an example of how AI may help produce ceramics more effectively and sustainably by reducing resource consumption, emissions, and overall material waste.

Predictive analysis for trend forecasting

AI and machine learning-powered predictive analytics is used to predict consumer preferences and design trends for tile and surface products. Predictive models spot current trends in areas such as color schemes, formats, and textures before they become popular. For example, computers can detect shifts toward warm, neutral tones or an increase in specific tile sizes, such elongated forms, by tracking image engagement on Pinterest and Instagram. Thanks to this

data-driven insight, manufacturers and designers can direct their collections proactively, connecting product development with future demand rather than reacting to it. When paired with machine-learning insights into how these preferences evolve over time, predictive analytics is a strategic tool for reducing risk, getting rid of overproduction, and increasing market relevance in the ceramics business.

Manufacturing and Supply chain optimization

Smart Inventory and logistic management

AI-driven inventory and logistics management is revolutionizing the ceramics manufacturing sector with more precise demand forecasts, ideal stock levels, and efficient distribution. Machine learning algorithms can analyze historical sales data, market trends, supplier lead times, and external factors (such seasonality) to predict optimal inventory requirements and reduce overstock and stockouts. Because raw materials and finished goods are heavy and brittle, AI-powered optimization reduces holding costs and logistical risk in the ceramics industry by helping identify how much inventory to maintain, where to store it locally, and when to repurchase. For delicate ceramic products, AI may concurrently suggest efficient loading and routing techniques, boosting container utilization and reducing damage during transit.

Marketing, Advertising and Sales

AI enhances marketing, advertising, and sales strategies that enables companies to more effectively reach target audiences and personalize customer experiences. Machine learning algorithms analyze consumer behavior, purchase history, social media engagement, and search patterns to predict preferences for tile kinds, colors, and finishes. Then marketers tailor their campaigns precisely. By showcasing popular collections and suggesting similar products, AI-powered recommendation engines integrated into e-commerce platforms improve the likelihood of cross-selling and upselling. By refining campaign targeting and bidding strategies in real time, AI in digital advertising guarantees a higher return on investment for promotional expenditures. Additionally, chatbots and virtual assistants increase customer engagement and conversion rates by offering individualized help and assisting customers with the decision and purchase process.

Benefits and Opportunities

Through generative design and predictive modeling, AI can help create highly customized patterns and forms while minimizing trial-and-error, which lowers material waste and accelerates time-to-market. Computer vision technologies improve quality control and drastically reduce reject rates by enabling automated problem detection. AI-driven process optimization, especially in kiln control, can increase output, lower energy use, and encourage greener manufacturing. Business-wise, advanced analytics provide more effective demand

forecasting, logistics optimization, and inventory management, which lowers overproduction and boosts productivity. Meanwhile, by suggesting designs that suit the client's preferences and style, chatbots and AI-powered marketing tools improve target audiences, increase sales, and make customers happy.

Challenges and Limitations

Large, well-labeled datasets of tile designs, flaws, and production parameters are necessary for AI models. They often fall short in conventional ceramic processes; another limitation is the availability and quality of data. Incorporating AI into existing manufacturing workflows can be difficult, requiring changes to people skill sets, production processes, and quality control methods. There are also worries about an excessive dependence on AI, potential bias in predictive models, and intellectual property rights for patterns generated by AI. Lastly, while AI can enhance design and procedures, it cannot fully replace human creativity, expertise, or artisan knowledge, all of which are still essential for upscale, handmade ceramic products. The challenges of using artificial intelligence to replicate artisan ceramics are highlighted in the Vancasso Tableware essay. Algorithms to effectively replicate these subtleties, which give handcrafted ceramics their depth and uniqueness is challenging.

Conclusion

Artificial intelligence and digital visualization technologies are radically changing the ceramics industry throughout the whole value chain, from design and production to marketing, sales, and customer interaction, as this review has shown. Instead of being experimental add-ons, tools like computer vision systems, generative AI, sophisticated visualization software, and predictive analytics are becoming more and more incorporated into modern ceramic methods. Client confidence is also enhanced using intelligent recommendation engines and targeted marketing, and virtual showrooms. They also boost client confidence, reduce sales cycles, and increase worldwide market reach. Human craftsmanship has cultural nuance and creative intuition and sensitivity, which cannot be entirely replicated by AI. Widespread adoption is still hampered by issues with data accessibility, system integration, ethical issues, and intellectual property rights. Above all, relying too much on automation runs the risk of eroding the artistic value that distinguishes superior ceramic work. Hybrid human-AI systems where digital intelligence enhances rather than replaces artisanal expertise and creative judgment are where AI in the ceramics sector is headed. Global markets are increasingly competitive based on AI usage and implementation in multiple industries. Ceramic companies can achieve sustained innovation, operational resilience, and improved customer experiences if they integrate AI carefully. The goal is also to maintain human authorship and cultural authenticity.

Works Cited

Blender Foundation. “Home of the Blender Project - Free and Open 3D Creation Software.” Blender.org, 2019, www.blender.org.

“Autodesk Fusion | 3D CAD, CAM, CAE, & PCB Cloud-Based Software | Autodesk.” Autodesk.com, 7 June 2024, www.autodesk.com/ca-en/products/fusion-360/overview

“SketchUp: 3D Design Software - Bring Your Vision to Life.” Trimble.com, 2025, sketchup.trimble.com/en.

RoomSketcher. “RoomSketcher.” RoomSketcher, RoomSketcher - Visualizing Homes, 2017, www.roomsketcher.com/.

Homestyler. “Homestyler - Free 3D Home Design Software & Floor Planner Online.” Homestyler, www.homestyler.com/.

“Augmented Reality for Ecommerce with Shopify AR.” Shopify, 2023, www.shopify.com/ca/augmented-reality.

Lehnen, Cormac. “Industry Meets Craft.” Ceramic Review, 8 Apr. 2024, www.ceramicreview.com/articles/clay-meets-code/.

Sahu, Yash. “Artificial Intelligence in Handmade Ceramics: Enhancing Designs and Optimization for Wholesale Success.” Ruralhandmade.com, 2024, ruralhandmade.com/blog/artificial-intelligence-in-handmade-ceramics-enhancing

RAK Ceramics, 2020, www.rakceramics.com/canada/en/media/news/rak-ceramics-launches-showroom-virtual-tour/

Lu, Jianfeng, et al. “A Sketch-Based Generation Model for Diverse Ceramic Tile Images Using Generative Adversarial Network.” Intelligent Automation & Soft Computing, vol. 37, no. 3, 2023, pp. 2865–2882, https://doi.org/10.32604/iasc.2023.039742

Wang, Xiaoyan, et al. “Research on the Generation of Ceramic Decorative Patterns from the Perspective of AIGC.” Mathematical Modeling and Algorithm Application, vol. 3, no. 3, 8 Jan. 2025, pp. 21–26, https://doi.org/10.54097/5w8n1m87.

Miray Gür, et al. “Communicating AI for Architectural and Interior Design: Reinterpreting Traditional Iznik Tile Compositions through AI Software for Contemporary Spaces.” Buildings, vol. 14, no. 9, 15 Sept. 2024, pp. 2916–2916, www.mdpi.com/2075-5309/14/9/2916, https://doi.org/10.3390/buildings14092916.

Cumbajin, Esteban, et al. “A Real-Time Automated Defect Detection System for Ceramic Pieces Manufacturing Process Based on Computer Vision with Deep Learning.” Sensors, vol. 24, no. 1, 31 Dec. 2023, pp. 232–232, https://doi.org/10.3390/s24010232.

“How AI Is Predicting 2025-2026 Subway Tile Trends.” Nova Tile and Stone, 7 Nov. 2025, www.novatileandstone.com/blog/subway-tile-2/how-ai-is-predicting-2025-2026-subway-tile-trends-107

---. “REDEFINING CERAMIC TILE DESIGN with ARTIFICIAL INTELLIGENCE - ACPV ARCHITECTS.” ACPV ARCHITECTS, 4 July 2023, acpvarchitects.com/news/redefining-ceramic-tile-design-with-artificial-intelligence/?utm.

Manetti, Alessandro, et al. “Design Trend Forecasting by Combining Conceptual Analysis and Semantic Projections: New Tools for Open Innovation.” Journal of Open Innovation: Technology, Market, and Complexity, vol. 7, no. 1, 10 Mar. 2021, p. 92, https://doi.org/10.3390/joitmc7010092

Sheth, Arch. “How AI Applications Are Reshaping the Future of Ceramic Manufacturing.” Syndell Technologies, Syndell, 18 Apr. 2025, syndelltech.com/ai-in-ceramic-industry/?

“Artificial Intelligence for Predictive Inventory Management in Manufacturing - Innovatech.” Innovatech, 2025, www.innovatech.info/use-case/artificial-intelligence-for-predictive-inventory-management-in-manufacturing/?

Gerosa, Sel. “Advanced Industrial Inventory Management Analytics.” Mosaic Data Science, 18 Feb. 2022, mosaicdatascience.com/2022/02/18/advanced-industrial-inventory-management-analytics

Lee, Sarah. “Tech-Driven Marketing in Ceramics.” Numberanalytics.com, 2025, www.numberanalytics.com/blog/tech-driven-marketing-ceramics

“The Role of AI in Ceramic Tile Design: Merging Tradition with Innovation.” CBIS 2025, 13 Mar. 2025, www.cbisexpo.com/post/the-role-of-ai-in-ceramic-tile-design-merging-tradition-with-innovation

Patra Admin. “How AI Is Transforming Ceramic Tableware Design | Patra Porcelain.” PATRA, the Leading Brand of Porcelain & Bone China Tableware, 2 May 2025, www.patraporcelain.com/articles/the-impact-of-ai-on-ceramic-tableware-design-and-innovation/

Rose, Elena. “AI & Handcrafted Ceramics: The Replication Challenge.” Vancasso, 2019, www.vancassotableware.com/blogs/dining-decoded/ai-replicating-handcrafted-ceramics-challenges?srsltid=AfmBOooydEBoMmj79vWNhI0ek2Z3330fb5MpyPcFpqsTbpqjQw2GH2Bp.

Rigidity of the Midpoint Mean Value Property on Intervals of Fixed Length
By Maksim Stepnov

Abstract

The mean value theorem gives an interior point at which the derivative is equal to the average rate of change but it does not state where that point is. A standard converse says that if the midpoint works for every interval then the function is a polynomial of degree maximum two. This paper asks what happens when only one fixed interval length is required. The result is that parabolas are not the only solutions, there is an infinite amount of counterexamples, in the form $e^{at}cos(bt)$, and the smallest growth rate that occurs is $\alpha_1 = 5.5373565660\ldots$ This paper then proves that a solution created from a finite amount of similar models and growing more slowly than $e^{\alpha_1|t|}$ must be a polynomial with degree equal or less than two, and that α_1 cannot be replaced by a larger number. The paper also gives an algebraic description of the problem using differentiation and translation operators. For every growth bound, the corresponding solution is a cyclic module over the ring created by differentiation. It can be shown with a quotient of this ring by an explicit polynomial. The rigidity result is equivalent to showing that this polynomial is z^3. Finally, it is shown that this is specific to the midpoint; if the point is moved further away from the center, the growth threshold disappears immediately.

Introduction

The mean value theorem guarantees the existence of a point where the derivative equals the average rate of change over an interval, but it does not state where that point is. If f is continuous on *[a,b]* and differentiable on *(a,b)*, then there exists some $\xi \in$ *(a,b)* such that $f'(b) = (f(b) - f(a))/(b - a)$, A good question is, for which functions can this point always be chosen to be the midpoint $\xi = (a + b)/2$?

Every quadratic polynomial has this property, if the midpoint is a mean value point for every interval, then *f* must be a polynomial of degree maximum two[1,3,5,8].

Theorem 1. *$f: R \to R$ is differentiable, and satisfies $f(a + h) - f(a) = h\frac{df}{dt}\left(a + \frac{h}{2}\right)$ for every $a \in R$ and $h > 0$. Then f is a polynomial of degree equal or less than 2.*

Proof. Put $a = t - \frac{h}{2}$,

$$f\left(t + \tfrac{h}{2}\right) - f\left(t - \tfrac{h}{2}\right) = h\tfrac{df}{dt}(t)$$

for every t and $h > 0$; both sides change sign when h does so (1) is correct for all real h. Theorem 3 shows that f has derivatives of every order. Differentiating (1) twice with respect to h and cancelling gives

$$\frac{1}{4}\left[\frac{d^2f}{dt^2}\left(t+\frac{h}{2}\right)-\frac{d^2f}{dt^2}\left(t-\frac{h}{2}\right)\right]=0 \qquad for\ all\ t\ and\ h,$$

so $\frac{d^2f}{dt^2}$ has the same value everywhere and f is a polynomial *equal or less than* 2. However, $\left(t+\frac{h}{2}\right)^2-\left(t-\frac{h}{2}\right)^2=2th=h\cdot\frac{d}{dt}\left(t^2\right)$, so every similar polynomial works.

Question 2. *Let $f: R \to R$ be differentiable and assume*

$$f(a+1)-f(a)=\frac{df}{dt}\left(a+\frac{1}{2}\right) for\ every\ a\in R,$$

so that the midpoint is a mean value point of every interval with length one. Is f a polynomial of degree equal or less than 2?

By the fundamental theorem of calculus, $f(a+1)-f(a)=\int_a^{a+1}\frac{df}{dt}(s)ds$, so writing $g:=\frac{df}{dt}$, condition (2) says exactly that

$$\int_a^{a+1} g(s)ds=g\left(a+\frac{1}{2}\right) \qquad for\ every\ a\in R.$$

that the *midpoint rule* of numerical integration with one subinterval with width 1, is exact for g on every interval with length one. Since $f(t):=\int_0^t g(s)ds$ recovers a solution (2) from a continuous solution (3). Theorem 1 is the fact that the midpoint rule is true for polynomials of degree *equal or less than 1* and when the step size may change, the content is that one constant step size allows a lot more.

The standard theorem suggests that the answer is "parabolas," it is not. There are an infinite amount of counterexamples.

$$u_1(t)=e^{\alpha_1 t}cos\left(\beta_1 t\right), \qquad \alpha_1=5.5373565660\ \ldots, \quad \beta_1=14.9953525556\ \ldots,$$

a fast growing oscillation completing one cycle approximately every 0.419 units. It satisfies (2) and it isn't a parabola.

Every one of them grows at a minimum as fast as $e^{\alpha_1|t|}$, so a solution that is not a polynomial cannot be small. Finding this is the main result of the paper, and the constant α_1 is actually sharp, it cannot be replaced by anything larger than it.

α_1 is not arbitrary. Trying $f(t)=e^{\lambda t}$ in (2) collapses the problem to a single equation, and $\lambda=2(x+iy)$ and separating real from imaginary turns that equation into

$$sinhxcosy=x, \qquad coshxsiny=y.$$

Section 3 shows the possible growth rates form an increasing sequence $\alpha_1 < \alpha_2 < \cdots$, and it shows α_1 is the smallest.

Methods

Note that (2) assumes $\frac{df}{dt}$ exists at every point since a ranges over all of R. Replacing a by $t - \frac{1}{2}$ changes (2) in the symmetric form used,

$$\frac{df}{dt}(t) = f\left(t + \tfrac{1}{2}\right) - f\left(t - \tfrac{1}{2}\right).$$

Theorem 3 *Every solution of (6) has derivatives of all orders.*

Proof. A differentiable function is continuous so the right side of (6) is continuous, so $\frac{df}{dt}$ is continuous and f has a continuous first derivative. If f has n continuous derivatives then the right side also does, therefore does $\frac{df}{dt}$, giving f its $(n + 1)$st.

Another note is that the interval length does not matter. If f has the midpoint property at h, meaning $f(a + h) - f(a) = h\frac{df}{dt}\left(a + \frac{h}{2}\right)$ for all a, then $g(t) := f(ht)$ has it at length 1 by chain rule, and bound $|f(t)| \le Me^{c|t|}$ becomes $|g(t)| \le Me^{ch|t|}$. So every statement continues to length h with each growth constant c replaced by c/h, the threshold α_1/hshorter intervals suggest the condition more frequently and are more rigid.

The equation as a quadrature error

Reformulation (3) is the statement that a quadrature error vanishes, and that error has a standard integral form, used twice below. Write

$$L[f](t) := f(t + 1) - f(t) - \frac{df}{dt}\left(t + \tfrac{1}{2}\right),$$

so that (2) reads $L[f] \equiv 0$.

Theorem 4 If g *has two continuous derivatives then for every* $m \in R$,

$$\int_{m-1/2}^{m+1/2} g(s)ds - g(m) = \int_{-1/2}^{1/2} k(v)\frac{d^2g}{dt^2}(m + v)dv, \qquad k(v) := \frac{1}{2}\left(\frac{1}{2} - |v|\right)^2,$$

and $\int_{-1/2}^{1/2} k(v)dv = \frac{1}{24}$. *So, for every* f *with three continuous derivatives,*

$$L[f](t) = \int_{-1/2}^{1/2} k(v)\frac{d^3f}{dt^3}\left(t + \tfrac{1}{2} + v\right)dv.$$

Proof. Substituting $s = m + u$ and writing $g(m + u) - g(m) = \int_0^u \frac{dg}{dt}(m + w)dw$ turns the left side into the double integral

$$\int_{-1/2}^{1/2} \int_0^u \frac{dg}{dt}(m + w)dwdu,$$

over the triangle $0 \le w \le u \le \frac{1}{2}$ together with the triangle $-\frac{1}{2} \le u \le w \le 0$, the second one counted negatively. Reversing the order of integration, the variable u runs over $\left(w, \frac{1}{2}\right)$ in the first and over $\left(-\frac{1}{2}, w\right)$ in the second, so the double integral equals

$$\int_{-1/2}^{1/2} sgn(w)\left(\frac{1}{2} - |w|\right)\frac{dg}{dt}(m + w)dw = -\int_{-1/2}^{1/2} \frac{dk}{dv}(w)\frac{dg}{dt}(m + w)dw,$$

since $\frac{dk}{dv}(w) = -sgn(w)\left(\frac{1}{2} - |w|\right)$. One integration by parts, with $k\left(\pm\frac{1}{2}\right) = 0$, gives the stated identity, and $\int_{-1/2}^{1/2} k = 2\int_0^{1/2} \frac{1}{2}\left(\frac{1}{2} - v\right)^2 dv = \frac{1}{24}$. For (8) take $g = \frac{df}{dt}$ and $m = t + \frac{1}{2}$; the fundamental theorem of calculus turns the integral of g into $f(t + 1) - f(t)$, and $g(m) = \frac{df}{dt}\left(t + \frac{1}{2}\right)$.

Since $k > 0$ on $\left(-\frac{1}{2}, \frac{1}{2}\right)$, formula (8) says a solution of (2) is precisely a function whose third derivative averages to zero against k at every point. Two things follow immediately. The first constrains the shape of any counterexample.

Theorem 5 *If f solves (2) and $\frac{d^3f}{dt^3}$ never changes sign, then f is a polynomial of degree max.* 2.

Proof. By Theorem 3, the integrand in (8) is continuous and of one sign, so the integral's vanishing together with $k > 0$ on the open interval forces $\frac{d^3f}{dt^3}\left(t + \frac{1}{2} + v\right) = 0$ for every t and every $v \in \left(-\frac{1}{2}, \frac{1}{2}\right)$.

So every counterexample has a third derivative that changes sign, a first indication that the counterexamples must oscillate. The second consequence is the polynomial case. Taking $f = t^j$ in (8),

$$L\left[t^j\right] = j(j - 1)(j - 2)\int_{-1/2}^{1/2} k(v)\left(t + \frac{1}{2} + v\right)^{j-3} dv = \frac{j(j-1)(j-2)}{24}t^{j-3} + \cdots \qquad (j \ge 3),$$

a polynomial of degree exactly $j - 3$; in particular $L[1] = L[t] = L\left[t^2\right] = 0$ and $L\left[t^3\right] = \frac{1}{4}$. So for a polynomial p of degree n we have $L[p] = 0$ if and only if $n \le 2$, because for $n \ge 3$ the

leading coefficient of $L[p]$ is a nonzero multiple of that of p. The polynomial solutions are precisely the parabolas, exactly as in the classical case.

Counterexamples

Substituting $f(t) = e^{\lambda t}$ for a complex number λ divide by $e^{\lambda t}$ turns (2)

$$e^{\lambda} - 1 = \lambda e^{\lambda/2}, \qquad that\ is, \qquad e^{\lambda/2} - e^{-\lambda/2} = \lambda.$$

Writing $\lambda = 2z$ and using $sinhz = \frac{1}{2}\left(e^{z} - e^{-z}\right)$,

$$sinhz = z.$$

Put $z = x + iy$ with x, y real. The formula together with Euler's formula gives $sinh(x + iy) = sinhxcosy + icoshxsiny$, so (10) is a real pair (5). Each pair (x, y) gives a real solution (2), taking real and imaginary parts of $e^{2(x+iy)t}$ shows that

$$e^{2xt}cos(2yt) \qquad and \qquad e^{2xt}sin(2yt)$$

both satisfy (2). These are the *modes* of the equation, $a = 2x$ is the *growth rate* of the mode. Two cases are immediate. If $y = 0$ the pair becomes $sinhx = x$, which for real x only at $x = 0$, since $sinhx > x$ for $x > 0$ and $sinh$ is odd. If $x = 0$ it becomes to $siny = y$, only at $y = 0$. So $(0, 0)$ is the only solution on all axis, it corresponds to 0 growth rate and the solutions of growth rate 0 are the parabolas of (9). So the question is whether (5) has any solutions with $x \neq 0$, it has infinite solutions.

Operators

For real number s let T_s be *translation* operator, $\left(T_s f\right)(t) := f(t + s)$ and $D := \frac{d}{dt}$ and $I := T_0$ for the identity. Then (7),

$$L = T_1 - I - DT_{1/2},$$

and (2) is the equation $L[f] = 0$. Let A mean the collection of all operators from D and the T_s by taking the real linear combinations and composites. Since $T_s T_u = T_{s+u}$ and D is commuted with every T_s, the collection A forms a commutative ring under addition and composition. It is commutative because differentiation commutes with every translation operator.

A **ring** is a collection of objects that can be added and multiplied according to the usual rules. In A, addition is the usual addition of operators, and multiplication is composition. This multiplication is commutative because differentiation commutes with translation.

An **ideal** is a subset of a ring that is closed under addition and remains inside the subset after multiplication by any element of the ring. For a fixed function, the operators in A that remove it form an ideal.

A **module** is similar to a vector space, except that its scalars come from a ring rather than a field. The solution spaces of (2) will be modules over the polynomial ring $R[D]$.

Theorem 6 *There is a map* σ *assigning to each* $P \in A$ *a function* $\sigma(P)$ *of a complex variable, determined by* $\sigma(D)(z) = z, \qquad \sigma\left(T_s\right)(z) = e^{sz}$,

such that $\sigma(P + Q) = \sigma(P) + \sigma(Q)$ *and* $\sigma(P \circ Q) = \sigma(P)\sigma(Q)$. *It satisfies*

$$P\left[e^{\lambda t}\right] = \sigma(P)(\lambda)e^{\lambda t} \qquad for\ every\ P \in A,\ every\ \lambda \in C.$$

For operator (12), $\sigma(L)(z) = e^z - 1 - ze^{z/2} =: \Delta(z)$,

the function that reappears in Theorem 11.

Proof. Both D and T_s act on $e^{\lambda t}$ by multiplication by a number: $D\left[e^{\lambda t}\right] = \lambda e^{\lambda t}$ and $T_s\left[e^{\lambda t}\right] = e^{\lambda s}e^{\lambda t}$. Sums and composites of operators that multiply $e^{\lambda t}$ by numbers again multiply it by the sum, respectively the product, of those numbers, so (13) holds with σ defined as stated, and σ converts sums into sums and composites into products. Reading (12) through σ gives the stated symbol.

Three things follow at once. First, (13) says every exponential is an eigenvector of *every* operator in A at the same time, with $\sigma(P)(\lambda)$ as the eigenvalue; so asking which exponentials solve $L[f] = 0$ is the same as asking where $\sigma(L)$ vanishes, which is how (10) arose. Second, for a fixed function f the set

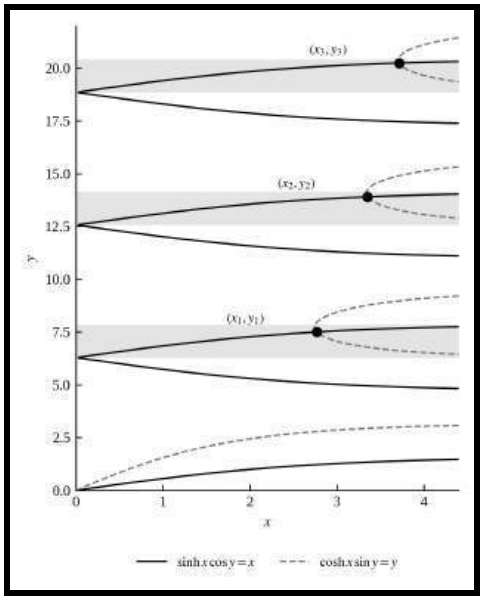
$$I_f := \{P \in A : P[f] = 0\}$$

is an ideal of A, it is closed under addition, and if $P[f] = 0$ then $(Q \circ P)[f] = Q[0] = 0$ for every Q. Question 2 asks which f have $L \in I_f$. Third, each $P \in A$ computes $P[f](t)$ from the values of f in a bounded window around t, and functions removed by such an operator are what Schwartz called *mean-periodic*; Theorem 15 below is his description of the solution space, for our particular L.

Results

Solving the two equations

Throughout this section (x, y) is a solution of (5). In the first equation both sides are odd in x and even in y; in the second both sides are even in x and odd in y. So changing the sign of x, or of y, or of both, again gives a solution, and solutions come in groups of four, $(\pm x, \pm y)$. It is therefore enough to find those with $x > 0$ and $y > 0$.



The pair (5) as two families of curves in the plane. A point solves the system exactly when a solid curve meets a dashed one, and the first three crossings are marked. The shaded strips are the bands J_k of Theorem 7; the crossings occur only there.

Each equation of (5) defines a family of curves in the (x, y) plane, and a solution of the system is a point where a curve of one family meets a curve of the other; Figure 1 shows both families. The rest of this section proves that the crossings are exactly one per horizontal strip, and locates them.

Theorem 7 (Solutions) *Let* (x, y) *solve (5) with* $x > 0$ *and* $y > 0$. *Then*

$$cosy > 0, \qquad siny > 0, \qquad \frac{tany}{y} = \frac{tanhx}{x},$$

and consequently y *lies in one of the intervals* $J_k := \left(2\pi k,\ 2\pi k + \frac{\pi}{2}\right)$ *with* $k \geq 1$ *an integer.*

Proof. Since $x > 0$ we have $sinhx > x > 0$, so the first equation gives $cosy = x/sinhx$, a number between 0 and 1. The second gives $siny = y/coshx > 0$. A real number whose sine and cosine are both positive is in $\left(2\pi k, 2\pi k + \frac{\pi}{2}\right)$ for integer $k \geq 0$. Dividing the second equation by the first gives $cothxtany = y/x$, which rearranges to $tany/y = tanhx/x$. Now $tanhx < x$ for every $x > 0$, so the right side is less than 1; and $tany > y$ on $\left(0, \frac{\pi}{2}\right)$, so if k was 0 the left side would be larger than 1. Therefore, $k \geq 1$.
So no solution has y below 2π, to locate the solutions down eliminate y.

Theorem 8 *Define* $Y(x) := cothx\sqrt{sinh^2 x - x^2} \qquad (x > 0).$

Then Y is strictly increasing, $Y(x) \to 0$ *as* $x \to 0^+$, *and* $Y(x) \to \infty$ *as* $x \to \infty$; *so Y takes every positive value once.*

Proof. Solve the two equations in (5) for cos y and sin y, then square and add, to get $x^2/sinh^2 x + y^2/cosh^2 x = cos^2 y + sin^2 y = 1$. Solving for y and using $y > 0$ gives $y = coshx\sqrt{1 - x^2/sinh^2 x}$, which is the labeled $Y(x)$.

For monotonicity, write $s(x) := x/sinhx$ and factor the same expression as

$$Y(x) = coshx \cdot \sqrt{1 - s(x)^2}.$$

Now $\frac{sinhx}{x} = \sum_{n\geq 0} \frac{x^{2n}}{(2n+1)!}$ is strictly increasing on $(0, \infty)$, since every term is non decreasing and the term $\frac{x^2}{6}$ is strictly increasing. So s is strictly decreasing with values in $(0, 1)$, therefore $1 - s^2$ is positive and strictly increasing, and so is its square root. Therefore, Y is a product of two positive strictly increasing functions and is therefore strictly increasing.

For the limits, $s(x) \to 1$ and $coshx \to 1$ as $x \to 0^+$, so $Y(x) \to 0$; and $s(x) \to 0$ and $coshx \to \infty$ as $x \to \infty$, so $Y(x) \to \infty$. A strictly increasing continuous function with these limits takes every positive value exactly once.

Theorem 9. *For each integer $k \geq 1$ there is exactly one solution (x_k, y_k) of (5) with $x_k > 0$ and $y_k \in J_k$, and these, together with $(0, 0)$ and the sign changes $(\pm x_k, \pm y_k)$, are all the solutions. $x_1 < x_2 < x_3 < \cdots$ and $x_k \to \infty$, so the growth rates $\alpha_k := 2x_k$ create a strictly increasing sequence and α_1 is the smallest one that is not zero.*

Proof. Fix $k \geq 1$ and let y range over J_k. By Theorem 8 there is exactly one $x > 0$ with $Y(x) = y$; call it $X(y)$, and note X is increasing because Y is. We claim that $(X(y), y)$ solves (5) exactly when

$$A(y) := \frac{tany}{y} \quad equals \quad B(y) := \frac{tanh(X(y))}{X(y)}.$$

To see this, write $p = cosy$, $q = siny$, $P = x/sinhx$ and $Q = y/coshx$ with $x = X(y)$. The relation $y = Y(x)$ says $P^2 + Q^2 = 1$, and $p^2 + q^2 = 1$; and (14) says precisely $qP = pQ$. Two unit vectors in the plane with $qP - pQ = 0$ are equal or opposite. Since $y \in J_k$ gives $p > 0$ and $x > 0$ gives $P > 0$, they are equal, $p = P$ and $q = Q$. The converse is Theorem 7 together with Theorem 8.

Compare the two sides of (14). On J_k,

$$\frac{dA}{dy}(y) = \frac{ysec^2 y - tany}{y^2} = \frac{sec^2 y}{y^2}\left(y - \frac{1}{2}sin2y\right) > 0,$$

because $y > 2\pi > \frac{1}{2} \geq \frac{1}{2} sin2y$ there; and A runs from 0 at the left end of J_k to $+\infty$ at the right end. Meanwhile $t \mapsto tanht/t$ is positive and strictly decreasing on $(0, \infty)$, and X is increasing, so B is positive and strictly decreasing on J_k, with finite values at both ends. So $A - B$ is strictly increasing, negative near the left end and $+\infty$ at the right end, and therefore vanishes at exactly one point of J_k. That gives exactly one solution per band, and Theorem 7 says there are no others with $x, y > 0$.

For the ordering, $y_k = Y(x_k)$ lies in J_k, so x_k lies in the interval $I_k := \left(X(2\pi k), X\left(2\pi k + \frac{\pi}{2}\right)\right)$. Since $2\pi(k+1) > 2\pi k + \frac{\pi}{2}$, these intervals are disjoint and move steadily to the right, so x_k increases with k. Finally the second equation of (5) gives $y_k \leq coshx_k$, and $y_k > 2\pi k \to \infty$, so $x_k \to \infty$.

The second equation of (5) can be solved for x outright, and doing so pins down how the growth rates behave for large k.

Theorem 10 (Exact formula and growth of the rates). *Every solution of (5) with $x, y > 0$ satisfies $coshx = \frac{y}{siny}$, that is, $x = arccosh\left(\frac{y}{siny}\right)$.*

Moreover $cosy_k \to 0$ and $siny_k \to 1$ as $k \to \infty$, and

$$x_k = log(4\pi k) + O\left(\frac{1}{k}\right), \quad so \quad \alpha_k = 2log(4\pi k) + O\left(\frac{1}{k}\right).$$

Proof. Equation (15) is the second equation of (5) divided by $siny$, which is positive by Theorem 7; and $cosh$ is increasing on $x > 0$, so it may be inverted.

By Theorem 9 we have $x_k \to \infty$, and the first equation of (5) gives $cosy_k = x_k / sinhx_k$, which tends to 0 because $sinhx$ outgrows x. Since $siny_k > 0$ and $sin^2 + cos^2 = 1$, this forces $siny_k \to 1$.

For the asymptotics, we use two elementary estimates. First, for $w \geq 2$,

$$arccoshw = log\left(w + \sqrt{w^2 - 1}\right) = log(2w) + log\frac{1+\sqrt{1-w^{-2}}}{2},$$

and since $\sqrt{1-u} \geq 1 - u$ for $0 \leq u \leq 1$, the last term lies between $-w^{-2}$ and 0; so

$|arccoshw - log(2w)| \leq w^{-2}$. Second, writing $c := cosy_k$, we have $siny_k = \sqrt{1-c^2}$ and hence $\left|logsiny_k\right| = \frac{1}{2}\left|log\left(1-c^2\right)\right| \leq c^2$ once $c^2 \leq \frac{1}{2}$. Applying both to (15),

$$x_k = log\left(2y_k\right) - logsiny_k + O\left(y_k^{-2}\right) = log\left(2y_k\right) + O\left(c^2\right) + O\left(k^{-2}\right).$$

Now $2\pi k < y_k < 2\pi k + \frac{\pi}{2}$ gives $4\pi k < 2y_k < 4\pi k + \pi$, so $0 < log\left(2y_k\right) - log(4\pi k) < log\left(1 + \frac{1}{4k}\right) < \frac{1}{4k}$. This already shows $x_k = log(4\pi k) + o(1)$, and in particular $x_k = O(logk)$. Feeding that back,

$coshx_k = y_k/siny_k$ is of size $2\pi k$, so e^{x_k} is of size $4\pi k$ and $c = x_k/sinhx_k = O(logk/k)$, whence $c^2 = O\left((logk)^2/k^2\right)$. Collecting the three error terms gives

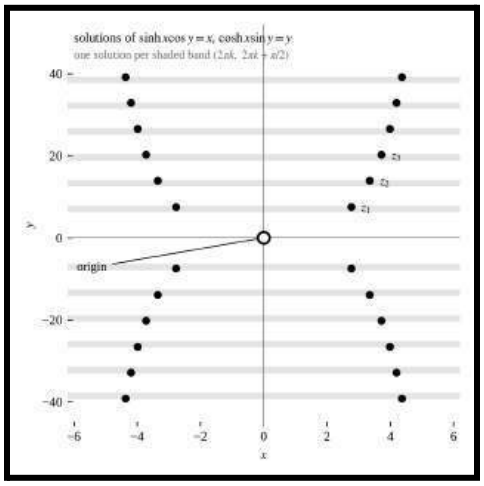
$\left|x_k - log(4\pi k)\right| \leq \frac{1}{4k} + O\left((logk)^2/k^2\right)$, which is the claim. Doubling gives the statement for α_k.

Numerically $k\left(x_k - log(4\pi k)\right)$ is 0.2377 at $k = 1$ and 0.2509 at $k = 1000$, so the error really is of size $\frac{1}{4k}$, exactly the bound the proof produces.

Since $y_k > 2\pi$, every mode has $\beta_k = 2y_k > 4\pi$, so no counterexample oscillates slowly. Table 1 lists the first six solutions and Figure 2 plots them.

The first six solutions of $sinhxcosy = x$, $coshxsiny = y$ with $x, y > 0$, and the growth rates of the corresponding modes. All digits shown are correct: the values were computed in mpmath at 50 decimal digits, and a script in the supplement regenerates every number in this paper.

k	x_k	y_k	growth rate $\alpha_k = 2x_k$
1	2.7686782830	7.4976762778	5.5373565660
2	3.3522098849	13.8999597140	6.7044197697
3	3.7167676798	20.2385177078	7.4335353595
4	3.9831416403	26.5545472655	7.9662832807
5	4.1932514704	32.8597410051	8.3865029409
6	4.3667951177	39.1588165201	8.7335902353



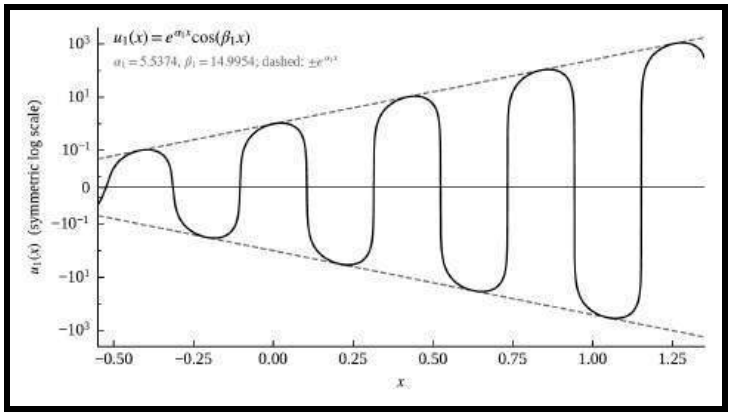
Solutions (x, y) *of the pair (5), plotted as points* $x + iy$*. The origin is a solution of a special kind and accounts for* $1,\ t,\ t^2$*. Every other solution lies in one of the shaded bands* $y \in \left(2\pi k, 2\pi k + \frac{\pi}{2}\right)$*, exactly one per band (Theorem 9).*

Counterexample 1

Taking $k = 1$ and forming the mode (11) gives

$u_1(t) = e^{\alpha_1 t} cos\left(\beta_1 t\right), \qquad \alpha_1 = 2x_1 = 5.5373565660\ \ldots, \quad \beta_1 = 2y_1 = 14.9953525556\ \ldots,$

which is the function announced in (4). Its period is $2\pi/\beta_1 = 0.4190088418\ \ldots$, so it completes more than two full oscillations per unit of length while its envelope multiplies by $e^{\alpha_1} \approx 254$. Figure 3 shows it. It answers Question 2 in the negative, and by Theorem 9 no counterexample of this form grows more slowly.



The first counterexample $u_1(t) = e^{\alpha_1 t} cos\left(\beta_1 t\right)$*, on a symmetric logarithmic scale so that the oscillation and the growth are visible together. The dashed lines are the envelope* $\pm\, e^{\alpha_1 t}$*.*

Threshold

We can now state the rigidity result. Call a function an *exponential sum* if it has the form

$$f(t) = p(t) + \sum_{k=1}^{N} e^{a_k t}\left(A_k cos\left(b_k t\right) + B_k sin\left(b_k t\right)\right),$$

with p a polynomial, N finite, the pairs $\left(a_k, b_k\right)$ distinct, $b_k \geq 0$, and $\left(A_k, B_k\right) \neq (0, 0)$. Every counterexample produced above is of this form, and so is every finite combination of them.

Theorem 11 (Threshold for exponential sums). *Let f be an exponential sum satisfying (2). Then p has degree at most 2, and each pair $\left(a_k/2, b_k/2\right)$ solves (5); in particular $\left|a_k\right| \geq \alpha_1$ for every k. Consequently, if in addition*

$|f(t)| \leq Me^{c|t|}$ *for all* t, *with* $0 \leq c < \alpha_1 = 5.5373565660\ldots$,

then no exponential terms can occur and f is a polynomial of degree at most 2.

Proof. The operator L of (7) is linear, and it sends each mode to a multiple of itself: for complex λ,

$$L\left[e^{\lambda t}\right] = \left(e^{\lambda} - 1 - \lambda e^{\lambda/2}\right)e^{\lambda t} =: \Delta(\lambda)e^{\lambda t},$$

so taking real and imaginary parts, L carries the two-dimensional space spanned by $e^{at}cos(bt)$ and $e^{at}sin(bt)$ into itself, and kills it exactly when $\Delta(a + ib) = 0$. Also $L[p]$ is a polynomial for any polynomial p.

The functions $1, t, \ldots, t^{degp}$ together with the modes $e^{a_k t}cos\left(b_k t\right)$, $e^{a_k t}sin\left(b_k t\right)$ for distinct pairs $\left(a_k, b_k\right)$ are linearly independent: they are solutions of a linear constant coefficient differential equation whose characteristic roots 0 and $a_k \pm ib_k$ are distinct, and such solutions are independent. Applying L to (16) and using $L[f] = 0$, the polynomial part and each two-dimensional mode block must therefore vanish separately. So $L[p] = 0$, which by (9) forces $degp \leq 2$; and $\Delta\left(a_k + ib_k\right) = 0$ for each k, which is exactly the statement that $\left(a_k/2, b_k/2\right)$ solves (5). By Theorem 9 a nonzero solution of (5) has $|x| \geq x_1$, so $\left|a_k\right| = 2|x| \geq 2x_1 = \alpha_1$.

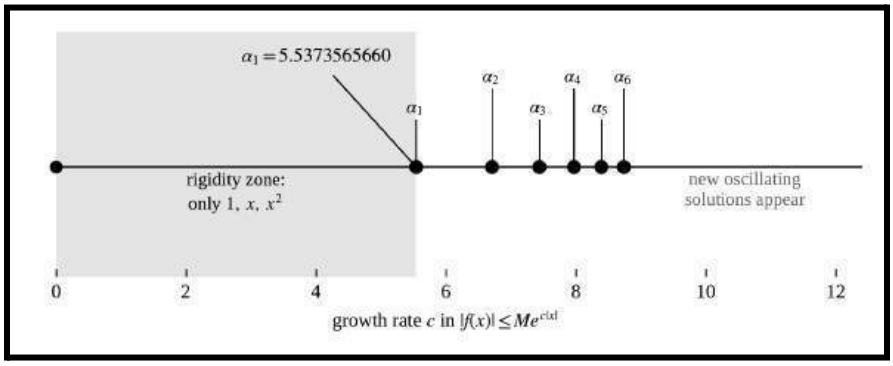
For the last claim, suppose some exponential term is present and let $a := max_k a_k$, assuming first $a > 0$. Multiplying f by e^{-at} and letting $t \to +\infty$, every other term tends to 0 and the polynomial part does too, so $e^{-at}f(t) - (Acos(bt) + Bsin(bt)) \to 0$ for the surviving block. The function $Acos(bt) + Bsin(bt)$ has amplitude $\sqrt{A^2 + B^2} > 0$, so there are arbitrarily large t with $|f(t)| \geq \frac{1}{2}\sqrt{A^2 + B^2}e^{at}$, which is incompatible with $|f(t)| \leq Me^{ct}$ once $c < a$. Hence $a \leq c < \alpha_1$, contradicting $\left|a_k\right| \geq \alpha_1$. If instead every a_k is negative, the same argument

applied as $t \to -\infty$ gives the same contradiction. So no exponential term occurs.

Remark 12. *The restriction $b_k \geq 0$ and the absence of terms like $te^{at}\cos(bt)$ cost nothing. A term $te^{\lambda t}$ satisfies $L\left[te^{\lambda t}\right] = \Delta(\lambda)te^{\lambda t} + \frac{d\Delta}{d\lambda}(\lambda)e^{\lambda t}$, so it can belong to a solution only if both $\Delta(\lambda)$ and $\frac{d\Delta}{d\lambda}(\lambda)$ vanish. One checks directly that $\frac{d\Delta}{d\lambda} \neq 0$ at every nonzero root, so such terms never appear.*
The constant in Theorem 11 cannot be improved, and the reason is simply that u_1 exists.

Theorem 13 (Sharpness). *The function u_1 of (4) satisfies (2), is not a polynomial, and obeys $\left|u_1(t)\right| \leq e^{\alpha_1|t|}$ for every t. Hence α_1 in Theorem 11 cannot be replaced by any larger constant.*
Proof. That u_1 satisfies (2) is the construction of (11). For $t \geq 0$ we have $\left|u_1(t)\right| \leq e^{\alpha_1 t} = e^{\alpha_1|t|}$, and for $t \leq 0$ we have $\left|u_1(t)\right| \leq e^{\alpha_1 t} \leq 1$. Since u_1 is unbounded and changes sign infinitely often it is not a polynomial. If Theorem 11 held with some $c' > \alpha_1$ in place of α_1, then u_1, which meets the growth bound with $c = \alpha_1 < c'$, would be forced to be a polynomial.

Figure 4 summarises the picture on a single axis.



The classification on one axis. Below $\alpha_1 = 5.5373565660$ the only solutions are 1, t and t^2. Each later growth rate α_k admits four further modes, $e^{\pm\alpha_k t}\cos(\beta_k t)$ and $e^{\pm\alpha_k t}\sin(\beta_k t)$.

Remark 14 (The bound has to be two-sided). *The absolute value in $e^{c|t|}$ matters. Asking only that $f(t) = O(e^{ct})$ as $t \to +\infty$ is not enough for any c, however small, because solutions of (5) come in $\pm$ pairs: alongside u_1 there is $v_1(t) := e^{-\alpha_1 t}\cos(\beta_1 t)$, which is bounded for $t \geq 0$ and is not a polynomial. The symmetry is forced, since $t \mapsto f(-t)$ turns one solution of (2) into*

another.

The general case

Theorem 11 assumes the solution is built from finitely many modes. That assumption can be dropped.

Theorem 15 (General threshold; proof in[9]). *Let* $f: R \to R$ *be differentiable and satisfy (2), and suppose* $|f(t)| \leq Me^{c|t|}$ *for all* t *with* $0 \leq c < \alpha_1$. *Then* f *is a polynomial of degree at most* 2. *More generally, for any* $c \geq 0$, *every solution obeying that bound is a finite combination of the modes whose growth rate is at most* c *in absolute value, together with a polynomial of degree at most* 2.

We do not prove this here, because the argument is of a different kind: one takes Laplace transforms of f on the two half-lines separately, observes that the two transforms are given by a single formula whose poles sit exactly at the numbers $2\left(x_k + iy_k\right)$ catalogued above, and moves a contour of integration across the strip $|Res| \leq c$, picking up one residue per admissible mode. That belongs to complex analysis rather than calculus, and it is carried out in the companion preprint[9]. General theories of this type for equations that refer to both past and future values go back to Rustichini[6] and Mallet-Paret[4]; see also Bellman and Cooke[2] for the classical theory of equations of this shape.

What matters for the present paper is that the constant is the same in both statements. The elementary analysis of (5) is what produces α_1, and no transform can change it.

Remark 16. *Some growth restriction really is needed in Theorem 15. Infinite superpositions* $\sum_k c_k e^{\alpha_k t} cos\left(\beta_k t\right)$ *with rapidly shrinking coefficients converge and differentiate term by term, and they satisfy (2) while growing faster than any single exponential. So the finiteness in (16) is a genuine hypothesis, not a technicality.*

The solution space as a module

The operator language of Theorem 6 lets the whole classification be packed into a single algebraic statement. Fix $c \geq 0$ and let $m = m(c)$ be the number of indices k with $\alpha_k \leq c$, so $m = 0$ exactly when $c < \alpha_1$. Let

$$M_c := \left\{f: R \to R : \ f\ is\ differentiable,\ L[f] = 0,\ and\ |f(t)| \leq Me^{c|t|}\ for\ some\ M\right\}.$$

By Theorem 15 this is the span of $1, t, t^2$ together with the $4m$ functions $e^{\pm\alpha_k t} cos\left(\beta_k t\right)$ and $e^{\pm\alpha_k t} sin\left(\beta_k t\right)$ for $k \leq m$, a real vector space of dimension $3 + 4m$. Differentiating any of these

stays inside the list, so M_c is carried into itself by $D = \frac{d}{dt}$, hence by every real polynomial in D. This makes M_c a module over the polynomial ring $R[D]$, with $p \cdot f := p(D)[f]$.

Theorem 17 (One function generates everything). *Set*

$$v_c(t) := t^2 + \sum_{k=1}^{m}\left(e^{\alpha_k t} + e^{-\alpha_k t}\right)cos\left(\beta_k t\right), q_c(z) := z^3 \prod_{k=1}^{m}\left(z^2 - 2\alpha_k z + \alpha_k^2 + \beta_k^2\right)\left(z^2 + 2\alpha_k z + \alpha_k^2 + \beta_k^2\right).$$

Then every $f \in M_c$ equals $p(D)\left[v_c\right]$ for some real polynomial p, and $R[D]/\left(q_c(D)\right) \cong M_c$ as $R[D]$-modules. So M_c is cyclic, generated by the single function v_c.

Proof. Write $f_0 := t^2$ and, for $1 \le k \le m$, $f_k^{\pm} := e^{\pm\alpha_k t}cos\left(\beta_k t\right)$, so that $v_c = f_0 + \sum_k\left(f_k^{+} + f_k^{-}\right)$. Put $q_0(z) := z^3$ and $q_k^{\pm}(z) := z^2 \mp 2\alpha_k z + \alpha_k^2 + \beta_k^2$.

Each summand generates one block. Since $\frac{d}{dt}t^2 = 2t$ and $\frac{d^2}{dt^2}t^2 = 2$, the smallest subspace containing f_0 and closed under D is $span\{1, t, t^2\}$, and $q_0(D)\left[f_0\right] = 0$. For real a and $b \neq 0$,

$$\frac{d}{dt}\left(e^{at}cos(bt)\right) = ae^{at}cos(bt) - be^{at}sin(bt), \qquad \frac{d}{dt}\left(e^{at}sin(bt)\right) = ae^{at}sin(bt) + be^{at}cos(bt),$$

so the smallest D closed subspace containing $e^{at}cos(bt)$ is the two-dimensional span of $e^{at}cos(bt)$ and $e^{at}sin(bt)$; on it $D^2 - 2aD + \left(a^2 + b^2\right)$ vanishes, because $a \pm ib$ are the roots of $z^2 - 2az + a^2 + b^2$. Taking $a = \pm\alpha_k$ and $b = \beta_k$ gives $q_k^{\pm}(D)\left[f_k^{\pm}\right] = 0$, and the blocks so obtained are exactly the ones listed above.

The polynomials $q_0, q_1^{+}, q_1^{-}, ..., q_m^{+}, q_m^{-}$ are pairwise coprime: their complex roots are 0 and the $4m$ numbers $\pm\alpha_k \pm i\beta_k$, all distinct because the α_k are distinct and positive and the β_k are positive. Their product is q_c, and $q_c(D)$ annihilates every summand of v_c, so $q_c(D)\left[v_c\right] = 0$.

Let $N := \{p(D)\left[v_c\right] : p \in R[D]\}$, and let g be any one of the summands, with factor q_g. The polynomial $r := q_c/q_g$ is divisible by the factor belonging to each of the other summands, so $r(D)\left[v_c\right] = r(D)[g]$. As r and q_g are coprime there are real polynomials a, b with $ar + bq_g = 1$, whence

$$a(D)\left[r(D)\left[v_c\right]\right] = a(D)r(D)[g] = \left(1 - b(D)q_g(D)\right)[g] = g.$$

So every summand lies in N, and N is closed under D, so N contains every block and therefore $N = M_c$.

Finally $p \mapsto p(D)\left[v_c\right]$ is a surjective map of $R[D]$-modules $R[D] \to M_c$ whose kernel contains the ideal $\left(q_c\right)$, so it induces a surjection $R[D]/\left(q_c\right) \to M_c$. As real vector spaces both sides have dimension $deg q_c = 3 + 4m$, so that surjection is an isomorphism.

Remark 18. *This restates the main theorem in one line. The property of Question 2 is rigid below growth rate c precisely when $q_c(z) = z^3$, and Theorem 11 says that this happens precisely for $c < \alpha_1$. Each time c passes an α_k, one more quartic factor is multiplied in and the module grows by four dimensions.*

Remark 19. *The module is over $R[D]$ and not over anything that undoes D: antidifferentiation does not preserve M_c, since t^2 solves (2) while $\int_0^t s^2 ds = \frac{t^3}{3}$ does not, $L\left[t^3\right]$ being $\frac{1}{4}$ by (9). The failure is measured exactly. If F is any antiderivative of a solution f, then L commutes with D, so $\frac{d}{dt}L[F] = L[f] = 0$ and $L[F]$ is some constant κ; since $L\left[t^3\right] = \frac{1}{4}$, the function $F - 4\kappa t^3$ is again a solution.*

Remark 20 (The first two cases). *For $c < \alpha_1$ we have $m = 0$, so $v_c = t^2$ and $q_c(z) = z^3$, giving $M_c \cong R[D]/\left(D^3\right)$: the entire solution space consists of t^2 together with its first two derivatives $2t$ and 2. For $\alpha_1 \le c < \alpha_2$ we have $m = 1$, $v_c(t) = t^2 + \left(e^{\alpha_1 t} + e^{-\alpha_1 t}\right)cos\left(\beta_1 t\right)$,*

$$q_c(z) = z^3\left(z^2 - 2\alpha_1 z + \alpha_1^2 + \beta_1^2\right)\left(z^2 + 2\alpha_1 z + \alpha_1^2 + \beta_1^2\right),$$

of degree 7. The seven dimensions are 1, t, t^2 and the four modes of rate $\pm\alpha_1$, and every one of them is a real linear combination of v_c and its first six derivatives. Adding the next band multiplies q_c by another quartic and raises the dimension to 11.

Moving the point off centre

Everything above is special to the midpoint, and it is easy to see how fragile that is. Fix $\theta \in (0, 1)$ and ask instead for $f(a + 1) - f(a) = \frac{df}{dt}(a + \theta)$. The same substitution $f(t) = e^{\lambda t}$ now gives

$$G_\theta(\lambda) := e^\lambda - 1 - \lambda e^{\theta\lambda} = \sum_{n \ge 2}\left(1 - n\theta^{n-1}\right)\frac{\lambda^n}{n!} = 0,$$

and $\theta = \frac{1}{2}$ returns the equation studied above. Write $c_*(\theta)$ for the corresponding threshold, the smallest $|a|$ over the nonzero growth rates. Then $c_*\left(\frac{1}{2}\right) = \alpha_1$. One might expect c_* to vary smoothly, and to guess that the midpoint, being the symmetric choice, is where it is smallest.

Both guesses are wrong.

Theorem 21 (Off-centre points). *For every* $\theta \neq \frac{1}{2}$ *in* $(0,1)$*, equation (17) has a nonzero real solution, negative if* $\theta < \frac{1}{2}$ *and positive if* $\theta > \frac{1}{2}$*. Moreover, writing* $\varepsilon := \theta - \frac{1}{2}$*, there are* $\eta > 0$ *and a smooth odd function* $\lambda(\varepsilon)$ *on* $(-\eta, \eta)$ *whose values are nonzero real solutions of (17) for* $\varepsilon \neq 0$*, with* $\lambda(\varepsilon) = 24\varepsilon + O(\varepsilon^3)$.

Consequently $c_*(\theta) \to 0$ *as* $\theta \to \frac{1}{2}$*, while* $c_*\left(\frac{1}{2}\right) = 5.5373565660\ldots$*; so* c_* *jumps at* $\theta = \frac{1}{2}$*, and the midpoint is a local maximum, not a minimum.*

Proof. From (17) the lowest term is $\left(\frac{1}{2} - \theta\right)\lambda^2$, so for $\theta < \frac{1}{2}$ the function G_θ is strictly positive at all small nonzero λ. On the other hand, as $\lambda \to -\infty$ we have $e^\lambda \to 0$ and $\lambda e^{\theta\lambda} \to 0$, since an exponential beats a linear factor, so $G_\theta(\lambda) \to -1$. By the intermediate value theorem G_θ has a root at some $\lambda_\theta < 0$. For $\theta > \frac{1}{2}$ the lowest term is negative, so $G_\theta < 0$ at small positive λ, while $G_\theta(\lambda) \to +\infty$ as $\lambda \to +\infty$ because e^λ outgrows $\lambda e^{\theta\lambda}$ for $\theta < 1$; the same argument gives a root $\lambda_\theta > 0$.

For the behaviour near the midpoint, set $\varepsilon := \theta - \frac{1}{2}$ and

$$F(\lambda, \varepsilon) := \frac{G_{1/2+\varepsilon}(\lambda)}{\lambda^2} = \sum_{n \geq 2} \left(1 - n\left(\frac{1}{2} + \varepsilon\right)^{n-1}\right) \frac{\lambda^{n-2}}{n!},$$

the series converging for all (λ, ε), so F is a smooth function of two variables. Reading off the first two terms, $F(0, \varepsilon) = \frac{1}{2} - \theta = -\varepsilon$ and $\partial_\lambda F(0,0) = \frac{1}{6} - \frac{1}{8} = \frac{1}{24}$, so

$$F(0,0) = 0, \qquad \partial_\lambda F(0,0) = \frac{1}{24} \neq 0, \qquad \partial_\varepsilon F(0,0) = -1.$$

The implicit function theorem therefore supplies $\eta > 0$ and a smooth $\lambda(\varepsilon)$ on $(-\eta, \eta)$, unique among small solutions, with $F(\lambda(\varepsilon), \varepsilon) = 0$, $\lambda(0) = 0$ and

$$\frac{d\lambda}{d\varepsilon}(0) = -\frac{\partial F/\partial\varepsilon(0,0)}{\partial F/\partial\lambda(0,0)} = 24.$$

It remains to see that λ is odd. A direct computation gives $G_{1-\theta}(-\lambda) = -e^{-\lambda} G_\theta(\lambda)$, so the root sets of $G_{1/2+\varepsilon}$ and $G_{1/2-\varepsilon}$ are negatives of one another, so $-\lambda(-\varepsilon)$ is also a small root for the parameter ε, and uniqueness gives $\lambda(\varepsilon) = -\lambda(-\varepsilon)$. An odd smooth function has no even-order terms in its Taylor expansion, so the ε^2 term is absent and $\lambda(\varepsilon) = 24\varepsilon + O(\varepsilon^3)$.

Finally $e^{\lambda(\varepsilon)t}$ is a solution of the off-centre equation with growth rate $|\lambda(\varepsilon)|$ and is not a polynomial, so $c_*(\theta) \leq |\lambda(\varepsilon)| \to 0$ as $\varepsilon \to 0$.

The slope 24 is not a new constant: it is $1/\int k$, the reciprocal of the weight in Theorem 4,

the same number that appears in (9) and in $L\left[t^3\right] = \frac{1}{4}$. The numerics agree with the theorem well outside the range it covers: at $\theta = 0.45$ the real root is -1.21459 against $24\varepsilon = -1.2$, and at $\theta = 0.49$ it is -0.24012 against -0.24. The mechanism behind the jump is a lost symmetry. At $\theta = \frac{1}{2}$ the pair (5) has no solution with $y = 0$, because that would need $sinhx = x$ with $x \neq 0$; this is what keeps every counterexample oscillating and expensive. Break the symmetry and that obstruction disappears: a real root emerges from the origin at first order in $\theta - \frac{1}{2}$, the cheapest counterexample becomes a plain exponential with no oscillation at all, and it is cheap precisely because it is a small perturbation of a constant.

Discussion

Written as (6), the equation ties $\frac{df}{dt}$ to values of f both ahead of t and behind it. Equations that look only backward, such as $\frac{df}{dt} = f(t-1)$, are easy to handle, write f on an interval of length one and integrate, and that data describes every solution. Here neither direction is available, and there is no initial value problem at all.

For a linear differential equation of order n the solutions are fixed by n numbers at a single point, whereas here the solution space is infinite dimensional and yet no finite list of numbers determines a solution. Theorems 3 and 5 are the first hints of this: the equation creates derivatives of all orders out of nothing, and forces the third of them to change sign, so the values of f on any interval are already heavily constrained by the values elsewhere.

The classification in Theorem 11 says that the constraint takes a very clean form. Growth rate acts like a budget. With a budget below α_1 nothing but a parabola can be afforded; raise it past α_1, then past α_2, and four new modes become available each time.

Theorem 17 says what the budget buys. Each time the budget is raised past an α_k the solution space gains four dimensions, and the enlarged space is still generated by one function together with the operation of differentiating. So the growth threshold is not only a size condition on solutions; it is a condition on which factors the polynomial q_c is allowed to have, and the rigidity statement is the single equality $q_c(z) = z^3$. This is also the reason the classification has the shape it does: quotients of $R[D]$ by a polynomial are the simplest modules there are, and the whole family of solution spaces here is of that kind.

Requiring the midpoint mean value property at a single interval length, rather than at all of them, changes the answer without destroying it. Parabolas lose their monopoly to an infinite family of growing oscillations, one for each solution of $sinhxcosy = x$, $coshxsiny = y$ (Theorem 9). What survives is a sharp threshold: among exponential sums, any solution growing more slowly than $e^{\alpha_1|t|}$ with $\alpha_1 = 5.5373565660\ldots$ is a polynomial of degree at most two (Theorem 11), and that constant cannot be raised (Theorem 13). The same threshold governs all

solutions of at most exponential growth (Theorem 15). The effect belongs to the midpoint alone: moving the evaluation point off centre by ε creates a real growth rate of size about 24ε, and the threshold collapses (Theorem 21).

Three questions remain open. First, what does $c_*(\theta)$ do away from $\theta = \frac{1}{2}$? This is harder to explore numerically than it looks, because the solutions for $\theta \neq \frac{1}{2}$ do not fall into neat bands, and a search started from the asymptotics misses some; at $\theta = 0.35$ such a search reports -4.0399904 as the smallest, while a careful count finds $3.8722154 \pm 11.6143549i$, which are smaller. Second, suppose f has the property at two interval lengths $h_1 \neq h_2$ at once, so that $L_{h_1}[f] = L_{h_2}[f] = 0$ for the operators $L_h := T_h - I - hDT_{h/2}$. Then the whole ideal of A generated by L_{h_1} and L_{h_2} annihilates f, and by (13) the only exponentials available to f are the common zeros of the two symbols. Since $\sigma\left(L_h\right)(z) = \Delta(hz)$, a nonzero common zero would mean that the scaling $z \mapsto \left(h_2/h_1\right)z$ carries one nonzero root of Δ to another, that is, that two solutions of (5) lie on one ray through the origin. Numericallyop=[[the angle $arctan\left(y_k/x_k\right)$ increases strictly for $1 \leq k \leq 2000$, which would rule that out and leave f a polynomial of degree at most two, but I have no proof. Third, α_1 is at present only described as $2x_1$ for a solution of two transcendental equations; a closed form would be of interest.

Acknowledgements

I thank Boris Belik for proofreading the manuscript and for suggesting the substitution used in Theorem 8. All numerical values in this paper were computed in mpmath at 50 decimal digits.

Works Cited

1. J. Aczél. Lectures on functional equations and their applications. Academic Press, 1966.
2. R. Bellman, K. L. Cooke. Differential-difference equations. Academic Press, 1963.
3. H. Haruki. A property of quadratic polynomials. The American Mathematical Monthly. Vol. 86, pg. 577-579, 1979.
4. J. Mallet-Paret. The Fredholm alternative for functional differential equations of mixed type. Journal of Dynamics and Differential Equations. Vol. 11, pg. 1-47, 1999, 10.1023/A:1021889401235.
5. B. Richmond, T. Richmond. How to recognize a parabola. The American Mathematical Monthly. Vol. 116, pg. 910-922, 2009, 10.4169/000298909X477023.
6. A. Rustichini. Functional differential equations of mixed type: the linear autonomous case. Journal of Dynamics and Differential Equations. Vol. 1, pg. 121-143, 1989, 10.1007/BF01047828.
7. L. Schwartz. Théorie générale des fonctions moyenne-périodiques. Annals of Mathematics. Vol. 48, pg. 857-929, 1947.
8. M. Stepnov. Localization theorem for the weighted mean value theorem. Preprint, 2026.
9. M. Stepnov. A transform proof of the midpoint rigidity theorem. Companion preprint, 2026.

An Interactive 3D Simulation Framework for Evaluating Quantum Confinement and Emission Profiles of Quantum Dots Using a Particle-in-a-Box Model By Alya Aydinoglu

Abstract

Spatial confinement fundamentally dictates particle energy levels in quantum mechanics, giving nanomaterials like quantum dots (QDs) unique, size-dependent optical properties. While QDs are critical components in modern lighting, solar energy, and QLED displays, demonstrating their underlying physics often requires complex frameworks. Drawing on Aleksey Ekimov's pioneering principles, this study investigates how closely a simplified 3D infinite potential well ("particle-in-a-box") model can qualitatively simulate QD emission spectra. We implemented this model in Python to solve for eigenstates and energies in a cubic potential well, deriving emission wavelengths from electron-hole recombination energies. The resulting interactive, open-source educational tool allows users to input spatial quantum numbers along the three Cartesian axes alongside the dot size. In real time, the code calculates and displays the electron state, hole state, bandgap, total transition energy, emission wavelength, and corresponding color of the dot. This project demonstrates how simplified models can generate clear insights into complex material behaviors, bridging abstract theory with observable real-world technology. Ultimately, this open-source simulation serves as an accessible STEM resource for future researchers and students seeking to visualize size-dependent quantum effects in semiconductor nanocrystals.

Introduction

Quantum Dots (QDs) are nanoscale semiconductor particles that exhibit size-dependent optical properties driven by quantum confinement effects. When the physical dimensions of these particles approach or drop below the exciton Bohr radius, the electrons and holes become spatially confined, resulting in quantized, discrete energy levels [1-5]. This discrete energy spectrum enables QDs to absorb and emit light at specific, size-dependent wavelengths. QDs are in high demand for applications such as biomedical imaging, solar cells, photodetectors, and quantum dot light-emitting diode (QLED) technologies [6,7]. Furthermore, compared to natural atoms, QDs could provide a distinct advantage in optical lattices due to their unique thermal properties when modeling the Fermi-Hubbard model in high-temperature superconductors [8].

QLED displays are particularly notable for leveraging the exceptional optical properties of QDs, including high color saturation, tunable emission, and high operational stability [6,7,10]. Traditional liquid-crystal LED displays suffer from lower power efficiency and limited functional flexibility because they rely heavily on color filters and backlights. In contrast, QLEDs are self-emissive, offering precise individual subpixel light control. Additionally, precise control over QD size and composition during chemical synthesis allows the emission spectra to be custom-tailored to optimize display performance. Because of their narrow emission spectra and robust temperature profiles as semiconductor materials, QDs remain prime candidates for the

future of optoelectronic devices, cementing QLEDs as the next generation of display technology [6,7,10]. However, a deeper understanding of these underlying quantum mechanical principles remains crucial for driving future material designs and technological improvements.

This study aims to develop and analyze a theoretical model of quantum dots to better understand their quantum confinement effects and resulting optical properties, particularly within the context of QLED applications. To accomplish this, we have developed a Python-based simulation that functions as an intuitive, open-source educational software tool to visually demonstrate these quantum models. The remainder of this paper is structured as follows: we first present a review of relevant literature, followed by the theoretical framework and computational modeling approach. We then present the simulation results alongside a detailed discussion, concluding with an outline of our final findings and potential directions for future research.

Literature Review

This section reviews fundamental quantum confinement phenomena, emission mechanisms in QLED devices, advances in quantum dot arrays for quantum simulation, computational modeling techniques, and the latest material choices, extracted from external research. There will be a significant emphasis that focuses on Quantum Dot optical properties, device applications, and theoretical modeling, giving a basis for the simulation model and data gathering [1-3,6,8].

Understanding the quantum mechanical behavior of electrons in confined systems is essential for modeling quantum dots. This study is grounded in the early experimental work of Alexey Ekimov on quantum size effects in semiconductor crystals and in the later theoretical and experimental work on semiconductor clusters and CdSe quantum dots [1,3-5]. These studies laid the foundation for understanding how the size and surface of quantum dots affect properties such as light absorption and emission spectra. Heating glass containing semiconductor components can lead to the formation of semiconductor nano crystallites, and quantum confinement of electrons in these crystallites produces the quantum size effect [1,3,4]. Ekimov's hypothesis suggests that as the size of quantum dots decreases, the energy levels become more discrete, which directly affects their absorption and emission spectra. The size-dependent behavior, supported by quantum confinement, aligns with the principles of the particle-in-a-box model. Applying the logic from Ekimov's hypothesis, alongside acknowledging how the quantum dot electron exhibits wave-like properties confined within the particle-in-a-box boundaries, then as the box becomes smaller, the electron's lowest energy level increases, leading to a blue shift in the emitted light. This explains how quantum dots can be engineered to emit specific wavelengths based solely on size, and for semiconductor nano crystallites, the fundamental optimization is the threshold energy for optical absorption.

From a theoretical standpoint, the semiconductor physics of electron wavelengths and quantized energy levels are central to understanding quantum dots. In confined systems like quantum dots, electron waves propagate with reduced inertia, resulting in the "effective mass" of the electron being less than its actual mass. This altered mass enhances the visibility of

confinement effects. Quantum confinement becomes significant in structures with dimensions on the order of roughly 100 angstroms, or 10 nanometers. Since the wavelength of an electron depends on both its energy and mass, size becomes a critical factor in tuning optoelectronic properties. Within this confinement, electrons behave like particles in a box, constrained to quantum wave functions that fit evenly within the dot's boundaries. The lowest energy level corresponds to a standing wave with a single antinode, the next to a wave with two, and so on — creating a harmonic series of energy levels. These levels rise rapidly, as electron energy is inversely proportional to the square of the wavelength. This harmonic nature — arising from the quantized standing wave patterns — is one of the most unique and defining features of quantum dots [3,4,11].

Furthermore, dots of different sizes produce different results in the harmonic spectra, a feature supported by Ekimov's theoretical framework. The discrete and predictable emission spectra that arise from these quantized energy levels are a direct result of quantum confinement. Each energy level corresponds to a higher harmonic, and their inverse-square dependence on wave-length gives rise to the unique spectral signatures of each quantum dot [3].

Quantum dots (QDs) serve as highly versatile platforms for simulating atomic and molecular systems, bridging foundational condensed matter physics with modern quantum computing architectures. The established paradigm of quantum dots as "artificial atoms" underpins their characteristic discrete electronic energy states and highly tunable optical profiles engineered within a solid-state environment. In the literature, 2-D QD arrays fabricated via electrostatic gating patterns or advanced epitaxial material growth have demonstrated the capacity to simulate complex quantum phenomena, including the Fermi-Hubbard model. This highlights the capacity of coupled quantum systems to explore collective many-body electronic interactions. When compared to trapped-atom systems, solid-state quantum dot platforms offer advantages, such as more accessible cryogenic temperature thresholds and intrinsic Coulombic interactions. While much of the existing research centers on these complex, multi-dot quantum computing simulations, evaluating the localized, single-dot electronic properties remains critical. This fundamental atomic analogy provides the theoretical foundation for implementing simplified particle-in-a-box approximations to model single-carrier wave functions and energy profiles [8,12].

The optoelectronic behavior of quantum dots is most visible in device applications such as QLEDs, where their size-tunable emission plays a critical role. Optoelectronics are the study of electronic devices with optical signals. Through an intersection of electrical charge flow, optoelectronic information is carried, processed, or detected using optoelectronic devices, through photons and electrons. It is important to reflect how electron configurations happen in both the ground state and the excited state, with the electrons having both an upward and downward spin in the ground spin. This is important to know because light emission is a result of the energy transfer from the excitation state to the ground state. This information could be presented in color, intensity, and polarization of light. Optoelectronic devices include LEDs, photodiodes, solar cells, laser diodes, and QLEDs. These devices use optoelectronic information

for means such as data transmission, display technologies, sensors, medical screening, and more [6, 7].

Particularly in QLED displays, the color is encoded in the emission spectra of quantum dots as previously mentioned, with the emitter being a semiconductor nanoparticle. In QLEDs, electrons and holes are injected from respective transport layers and recombined in the quantum dot layer, emitting photons with energies corresponding to the bandgap. This process is governed by the Planck–Einstein relation, which ties the emitted photon's wavelength to the energy gap — a foundational concept in quantum optics. Furthermore, the thin layer structure of the QLED device makes it reliable in flexible displays [6, 10].

While theoretical models have been established, recent rese-arch also explores practical fabrication and manipulation of Quantum Dots across institutions. Institutions such as CNET in France, NTT in Japan, the University of Cambridge, the State University of New York at Stony Brook and Princeton University have used methods of fabrication with the harmonic spectra. Pierre Guéret et al. at IBM Zürich has already developed a quantum dot tuned through an electronic gate, where the applied potential changes the size and thus the energy levels of the dot [9]. Delft and MIT further reveal that quantum dot energy levels are influenced not only by size but also by charge quantization and the dot's capacitance, deepening the understanding of quantum dot behavior. Furthermore, innovation in materials science is also proving itself, such as ZnO nanoparticles and PEDOT: PSS being used to manipulate the energy levels and charge confinement within QLED structures [7,10]. These engineered confinement improvements using these nanoparticles improve overall QLED performance, further proving that quantum dots are the technology of the future.

Despite these significant advances, challenges remain, primarily in integrating quantum dots into viable circuits that outperform traditional silicon-based technologies. While the theoretical potential of quantum dots is immense, the ultimate viability of quantum circuits depends on how well they compete against advancements in conventional silicon technology. Moreover, while emission, absorption, and lasing spectra can be finely tuned, scaling this technology for complex, programmable architectures continues to present a substantial hurdle with interconnections and internal architectures being adapted dynamically to solve these new arisen issues.

In conclusion, the current literature provides strong support for the modeling of quantum dots as confined systems governed by discrete energy states. However, there remains a gap in bridging simplified theoretical models with the nuanced, material-specific behaviors observed in real devices. This study seeks to address that gap by applying an educational yet insightful particle-in-a-box model to simulate quantum dot emission, thereby exploring how size-dependent quantum effects translate into macroscopic optical properties. This study applies a simplified but informative particle-in-a-box model to simulate emission wavelengths of quantum dots, with a focus on how these results connect to practical QLED behavior.

Theoretical Background

We have developed a simulation tool that is based on the three dimensional particle-in-a-box model, implemented in Python using scientific and visualization libraries. By adjusting parameters such as quantum dot size and quantum numbers, the simulation allows real-time observation of how energy levels and emission wavelengths shift due to quantum confinement. Theoretical foundations, simulation implementation, and limitations are discussed below.

Theoretically, the three-dimensional particle-in-a-box model provides a fundamental framework for understanding the quantum confinement in quantum dots. In this model, an electron is confined within a cubic potential well with infinitely high walls, and its wave function must vanish at the boundaries. Starting from the Schrödinger equation [11],

$$-\frac{\bar{h}^2}{2m}\nabla^2\Psi(x,y,z) = E\Psi(x,y,z) \qquad \text{(Eq1)}$$

where $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ is the three-dimensional Laplacian operator, $\bar{h}$ is the reduced Planck constant, and m is the mass of the particle. The solution for a cubic box of side length L yields normalized wave functions of the form:

$$\Psi n_x, n_y, n_z(x,y,z) = \sqrt{\frac{8}{L^3}}\, sin\left(\frac{n_x \pi x}{L}\right) sin\left(\frac{n_y \pi y}{L}\right) sin\left(\frac{n_z \pi z}{L}\right) \qquad \text{(Eq2)}$$

where $n_x, n_y, n_z \in N^+$ are the quantum numbers that define the discrete energy levels along each spatial dimension. This solution leads to a quantized energy expression that accounts for all three spatial components:

$$E = \frac{\bar{h}^2 \pi^2}{2m\left(L10^{-9}\right)^2}\left(n_x^2 + n_y^2 + n_z^2\right) \qquad \text{(Eq3)}$$

Although this idealized model does not incorporate real-world factors such as effective mass variation, surface effects, or finite potential barriers, it still reveals how reducing the size L of the quantum dot increases energy level separation — a hallmark of quantum confinement [1,3]. In semiconductor quantum dots, both electrons and holes are confined. Their individual energies are computed using the above equation, substituting their respective effective masses m_e and m_h. The total transition energy between a confined electron and hole state, combined with the intrinsic bandgap energy E_g^{bulk} of the material (e.g.,CdSe), gives the exciton recombination energy [2,3]:

$$E_{total} = E_g^{bulk} + E_e + E_h \qquad \text{(Eq4)}$$

where E_e and E_h are defined respectively as:

$$E_e = E_{nx} + E_{ny} + E_{nz}\,, \qquad E_h = E_{nx,h} + E_{ny,h} + E_{nz,h}. \qquad \text{(Eq5)}$$

Because photon energy is related to wavelength by $E = \frac{hc}{\lambda}$, the emitted photon's wave length is:

$$\lambda = \frac{hc}{E_{total}}, \quad \lambda_{nm} = \frac{hc}{E_{total}} 10^{9} \qquad \text{(Eq6)}$$

Modeled by the wavelength (λ), where h stands for Planck's constant and c is the speed of light. Finally, to visualize the probability distribution of the particle's position within the quantum dot, the probability density function is used:

$$P(x, y, z) = |\psi(x, y, z)|^2 \qquad \text{(Eq7)}$$

While this model is not sufficient for precision design of quantum dot devices, it captures the essential physics of how size influences emission energy, making it valuable for educational simulations and first-principles understanding of nanostructured materials [1, 3].

In principle, quantum simulators are physical systems engineered to emulate the behavior of other quantum systems [13, 14]. While digital quantum simulation (DQS) uses gate-based quantum computers, analog quantum simulation leverages specific quantum hardware such as cold atoms or trapped ions. However, due to technical constraints and decoherence, practical quantum simulators are limited in precision. As a result, classical simulations such as the one implemented in this project remain valuable for modeling simple systems like quantum dots under idealized conditions. A real life quantum simulator requires a controllable quantum mechanical system emulating evolved quantum systems. This will require a Digital Quantum Simulation (DQS) quantum computer, and while DQS uses gate-based quantum computers, analog quantum simulation leverages specific quantum hardware such as cold atoms or trapped ions. Even with Analog Quantum Simulation being currently used to solve these physics problems, the controls vary from system to system. These vary through laser pulses, radio frequency pulses, or electro-magnetic fields.

Alas, simulated systems are not coherent, causing errors in the quantum simulation which end up being overlooked. Although the required level of precision and control is relatively low, errors are still not minimized. Progress has been achieved within the past decade, with Quantum simulators almost guaranteeing tremendous impact in many fields, providing a basis to explore new physical phenomena like the quantum dot. So, while Richard P. Feynman can look over us and be proud, with the risk for error considered, this is why simple simulations through Python prove to be advantageous.

Computational modeling provides several advantages when studying quantum systems like quantum dots. Simulations allow for the isolation of parameters, facilitate visual trend analysis, and support early hypothesis testing before experimental validation. However, idealized models like the infinite potential well ignore many real-world complexities such as Coulomb interactions, finite barrier heights, temperature effects, or surface phenomena. As such, while

educational and insightful, these simulations should be interpreted within their limits.

Experimental Results

The simulation framework was developed in Python using Visual Studio Code, utilizing NumPy for core numerical computations and a combination of Plotly and Dash for real-time, interactive 3D visualizations. The application features an intuitive user interface equipped with adjustable sliders to control quantum dot diameter (1-10nm) and the spatial quantum numbers that define the first excited energy states for single charge carriers (electrons and holes) along the three Cartesian axes. Modifying these parameters dynamically updates the wave function distributions, emission wavelengths, and corresponding transition energies. Large language models were utilized during development strictly to verify mathematical equations and guide the software implementation. The simulation platform is publicly accessible at the GitHub repository (https://github.com/AlyaAydinoglu/Quantum_Dot_Simulations), and the structural layout of the parameter input screen is illustrated below in Figure 1.

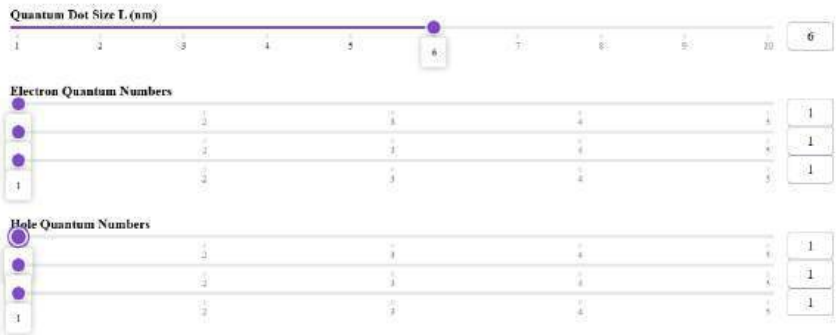


Figure 1. Parameter input interface and control layout for the interactive quantum dot simulation.

As illustrated in the user interface (Figures 1 & 2), initializing the simulation with an electron state of (1,1,1), a hole state of (1,1,1), and a quantum dot diameter of 6.0nm yields a calculated emission wavelength of 637.7nm, a predicted emission color of orange, and a transition energy of 1.944eV. This is the absolute ground-state exciton configuration with the lowest-energy electronic excitation possible within the nanostructure. Our results were compared with experimental data for cadmium selenide (CdSe) quantum dots, which exhibit a true transition energy of 2.051eV and a corresponding peak wavelength of 605nm. The simulation's predictions deviate from the empirical measurements by less than 5%, and the predicted orange color strongly correlates with the physically observed emission spectrum. The resulting spatial distribution of the quantum dot system is visualized in Figure 3.

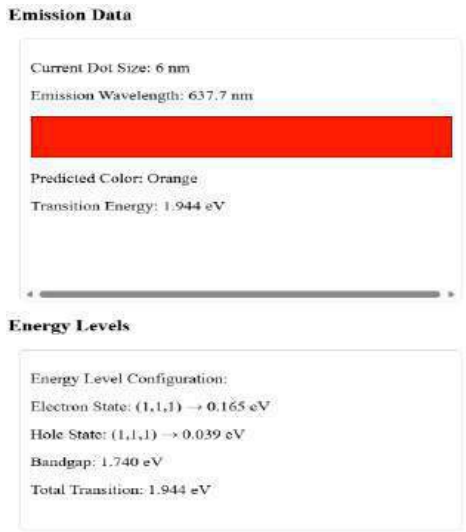


Figure 2. Calculated electronic transition properties and predicted optical emission parameters from the particle-in-a-box model simulator.

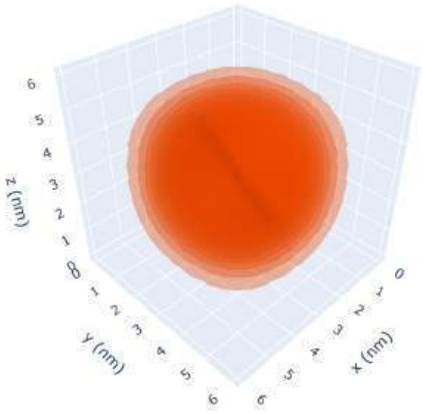


Figure 3. Simulated spatial mapping of the absolute ground-state exciton configuration for a 6.0nm particle, demonstrating carrier localization along the Cartesian axes.

To further evaluate the model under higher-order conditions, a secondary configuration was simulated utilizing an electron/hole state of (2,1,1)/(2,1,1) within a quantum dot diameter L of 8.0nm. Because the spatial quantum numbers of the co-confined carriers are identical, the overlap integral remains maximized, resulting in a highly allowed optical transition characterized by an intense orange emission. Due to this symmetric (2,1,1) spatial configuration, the respective probability density mappings of the electron and hole are structurally identical. As illustrated in Figure 4, the spatial distribution features two distinct probability lobes aligned along the x-axis, separated by a vertical nodal plane down the center since the quantum number is 2. Conversely, the wave function profiles along the y- and z-axes maintain their smooth, continuous ground-state geometries as quantum numbers remained at baseline quantum numbers.

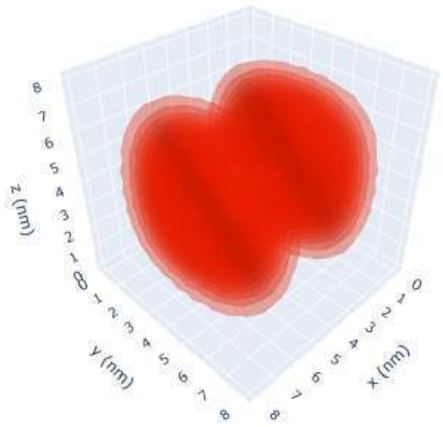


Figure 4. 3D spatial mapping and nodal structures of the identical electron and hole wave functions for the (2,1,1)/(2,1,1) configuration.

Subsequently, the simulation was executed across multiple electron/hole configurations by sweeping the quantum dot diameter L from 2.0 to 10.0nm in discrete increments. This generated a real-time characteristic curve plotting emission wavelength as a function of quantum dot size, which was color-coded to reflect the predicted optical output. For the initial trial, the ground-state configuration (1,1,1)/(1,1,1) was evaluated. The resulting profile (Figure 5) demonstrated a sharp, non-linear increase in emission wavelength at smaller diameters, which asymptotically approached a plateau as L increased, spanning a range from 346.6nm to 683.7nm. The simulation strongly correlates with the empirically reported 280–670nm emission band for CdSe quantum dots. Specifically, the onset of blue emission occurs near L = 2.5nm, yellow near L = 4.0nm, and red near L = 7.0nm, all of which strictly align with experimental literature values. The comprehensive dataset compiling these sizing metrics is provided in Table 1.

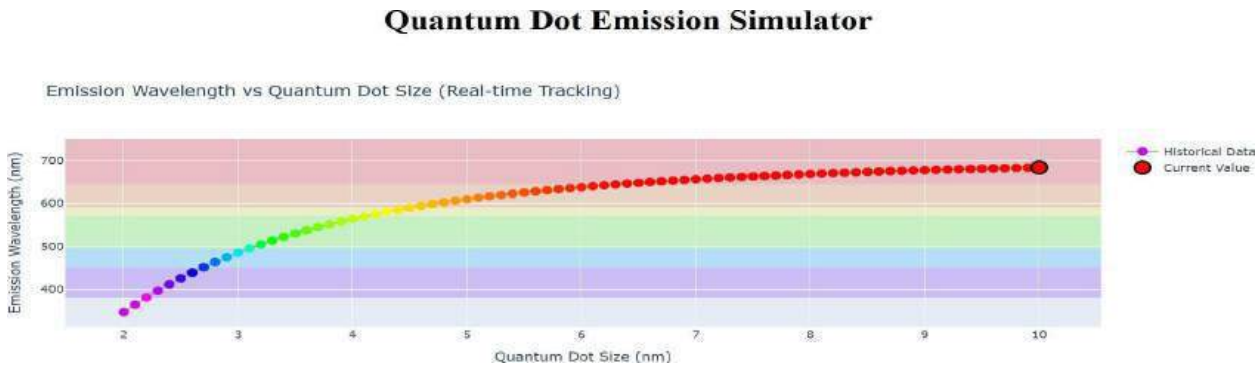


Figure 5. Simulated emission wavelength as a function of quantum dot diameter (L) from 2.0 to 10.0nm for the ground-state configuration, color-coded to represent the corresponding optical spectrum shifting from ultraviolet to red.

Table 1. Computed emission-wave length dataset for electron/hole quantum number configuration (1,1,1)/(1,1,1)

L(nm)	λ (nm)	L(nm)	λ (nm)	L(nm)	λ (nm)
2.0	346.6288	4.7	598.2022	7.4	661.5401
2.1	364.0080	4.8	602.1869	7.5	662.7970
2.2	380.5456	4.9	605.9788	7.6	664.0091
2.3	396.2513	5.0	609.5893	7.7	665.1785
2.4	411.1433	5.1	613.0291	7.8	666.3071
2.5	425.2458	5.2	616.3080	7.9	667.3968
2.6	438.5874	5.3	619.4355	8.0	668.4492
2.7	451.1998	5.4	622.4202	8.1	669.4660
2.8	463.1162	5.5	625.2701	8.2	670.4488
2.9	474.3711	5.6	627.9929	8.3	671.3990
3.0	484.9989	5.7	630.5956	8.4	672.3180
3.1	495.0337	5.8	633.0849	8.5	673.2072
3.2	504.5089	5.9	635.4669	8.6	674.0677
3.3	513.4569	6.0	637.7475	8.7	674.9009
3.4	521.9689	6.1	639.9321	8.8	675.7078
3.5	529.8945	6.2	642.0259	8.9	676.4894
3.6	537.4421	6.3	644.0335	9.0	677.2469
3.7	544.5784	6.4	645.9594	9.1	677.9811
3.8	551.3288	6.5	647.8079	9.2	678.6931
3.9	557.7172	6.6	649.5830	9.3	679.3836
4.0	563.7658	6.7	651.2882	9.4	680.0536
4.1	569.4959	6.8	652.9272	9.5	680.7038
4.2	574.9269	6.9	654.5032	9.6	681.3349
4.3	580.0774	7.0	656.0192	9.7	681.9478
4.4	584.9647	7.1	657.4783	9.8	682.5430
4.5	589.6047	7.2	658.8831	9.9	683.1213
4.6	594.0125	7.3	660.2362	10.0	683.6832

To investigate the influence of asymmetric carrier states, the emission characteristics were evaluated by varying the electron spatial configuration to an excited (2,1,1) state while keeping the hole fixed in the (1,1,1) ground state (Figure 6). In this configuration, the emission spectrum exhibits a pronounced blue shift toward shorter wavelengths across the entire sizing domain. This wavelength suppression remains uniform; for instance, the onset of the red emission band is delayed, appearing only when the quantum dot diameter reaches approximately 8.0nm. At a minimal diameter of L = 2.0nm, the simulated wavelength is 245.0nm, which is significantly shorter than the 346.6nm recorded in the symmetric ground-state trial. Conversely, at the upper boundary of L = 10.0nm, the wavelength

asymptotically approaches 662.0nm, which is comparable to the 683.7nm observed in the first configuration. This systemic shift toward higher transition energies is consistent with established quantum confinement theories for asymmetric excited-state carrier configurations.

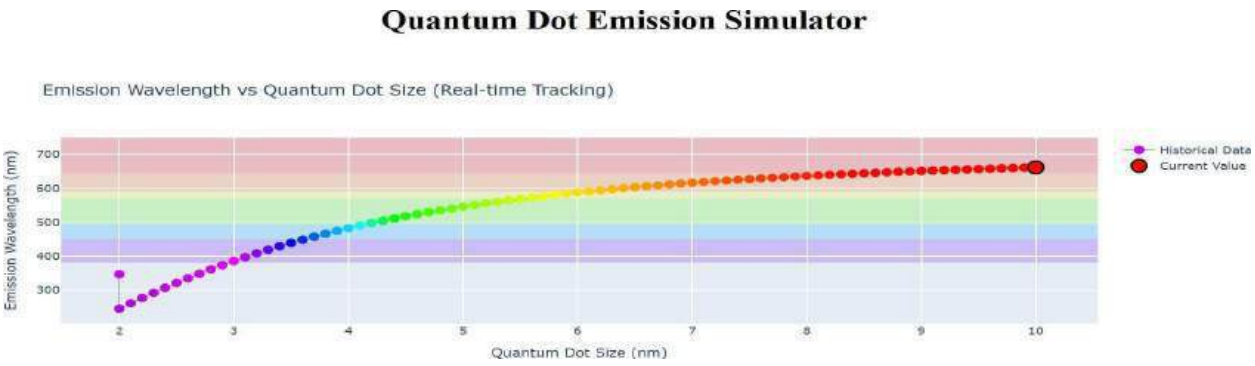


Figure 6. Emission wavelength as a function of quantum dot size for electron/hole quantum-number configuration (2, 1, 1)/(1, 1, 1).

These simulations demonstrate the systematic influence of quantum confinement dynamics and localized quantum number configurations on the primary emission characteristics of CdSe quantum dots. For both trials, the non-linear increase in emission wavelength with expanding quantum dot diameter matches the expected trend dictated by quantum confinement theory [1, 3]. Restricting the physical boundaries of the nanostructure intensifies carrier confinement, expanding the effective energy bandgap and prompting a characteristic blue shift toward shorter wavelengths. Conversely, increasing the quantum dot diameter relaxes this spatial confinement, systematically narrowing the bandgap and resulting in a red shift toward longer wavelengths.

The observed evolution of the emission profile can be explained by two competing mechanisms within the nanostructure. First, increasing the quantum dot diameter L relaxes spatial confinement and reduces the effective energy bandgap, which consequently shifts the emission wavelength toward longer, lower-energy spectral regions. Second, elevating the carrier quantum numbers populates higher-order eigenenergies, increasing the total transition energy and inducing a distinct blue shift toward shorter wavelengths. These wavelength shifts are coupled to the electronic bandgap through the Planck–Einstein relation, which forms the fundamental physics framework governing quantum dot emissions. Ultimately, this particle-in-a-box simulation platform provides an efficient, intrinsic method to model the interplay between spatial geometries and optical properties.

Smaller quantum dots exhibit a strong quantum confinement effect and emit higher-energy blue light, whereas larger nanocrystals emit lower-energy red light, making them highly valuable for the red subpixels in commercial QLED displays [6,7]. Furthermore, tuning the carrier quantum numbers provides an additional degree of freedom to precisely modulate the emission wavelength, offering a predictive pathway for engineering color-pure QLED

architectures. While this model effectively evaluates CdSe quantum dots, quantifying the Stokes shift would be a critical next step to optimize subsequent material layering, thereby enhancing both operational stability and quantum efficiency. On a macro level, this framework connects idealized physics to practical optoelectronic device engineering by translating electronic configurations into predicted emission spectra. However, because this platform utilizes a simplified particle-in-a-box approximation, it does not account for multi-particle Coulombic interactions (excitonic binding energies) or phonon-mediated lattice coupling. Future iterations will address these ideal-behavior assumptions to improve modeling fidelity in heavily engineered systems.

Conclusion

This study evaluates the simulated emission characteristics of CdSe quantum dots by systematically varying nanostructure dimensions and carrier quantum configurations. Utilizing an interactive three-dimensional infinite potential well model, a simulation platform was engineered to provide real-time tracking of optical spectral shifts under dynamic parameter adjustments. As dictated by quantum confinement theory, expanding the quantum dot diameter relaxes carrier confinement to yield longer emission wavelengths, whereas elevating the electron spatial quantum numbers populates higher eigenenergies, producing a definitive hypsochromic shift. Degenerate electronic configurations yielded structurally identical optical outputs, validating the mathematical and physical symmetry inherent to isotropic three-dimensional systems.

These findings underscore the critical role of precise geometric tuning and energy-level engineering when optimizing optical profiles for advanced optoelectronic applications, such as quantum-dot light-emitting diodes (QLEDs). Predictable control over emission spectra offers a distinct design advantage for next-generation display architectures. However, because this study relies on an idealized particle-in-a-box model, it intentionally omits multi-particle Coulombic interactions, phonon-mediated coupling, and material-specific lattice dynamics. Consequently, while the framework serves as an efficient conceptual baseline, expanding its predictive fidelity for complex experimental settings remains a key developmental objective. Future iterations will extend the model to incorporate excitonic binding energies, diverse semiconductor materials, and direct calibration loops against empirical spectral data. Ultimately, this platform serves as a high-utility pedagogical tool enabling rapid, visual evaluation of quantum mechanics principles, while establishing a foundational framework for future industrial and research-driven device modeling.

Works Cited

1.Ekimov, A. I. ve Onushchenko, A. A. (1981). Quantum size effect in three-dimensional microscopic semiconductor crystals. JETP Letters, 34(6):345–349.

2.Brus, L. E. (1984). Electron-electron and electron-hole interactions in small semiconductor crystallites: The size dependence of the lowest excited electronic state. The Journal of Chemical Physics, 80(9):4403–4409. doi:10.1063/1.447218.

3.Brus, L. E. (1986). Electronic wave functions in semiconductor clusters: experiment and theory. The Journal of Physical Chemistry, 90(12):2555–2560. doi:10.1021/j100403a003.

4.Norris, D. J. ve Bawendi, M. G. (1996). Measurement and assignment of the size-dependent optical spectrum in CdSe quantum dots. Physical Review B, 53(24):16338–16346. doi:10.1103/PhysRevB.53.16338.

5.Nobel Prize Outreach. The Nobel Prize in Chemistry 2023. NobelPrize.org.

6.Shirasaki, Y., Supran, G. J., Bawendi, M. G., ve Bulović, V. (2013). Emergence of colloidal quantum-dot light-emitting technologies. Nature Photonics, 7(1):13–23. doi:10.1038/nphoton.2012.328.

7.Talapin, D. V. ve Steckel, J. (2013). Quantum dot light-emitting devices. MRS Bulletin, 38(9):685–691. doi:10.1557/mrs.2013.204.

8.Hensgens, T., Fujita, T., Janssen, L., Li, X., Van Diepen, C. J., Reichl, C., Wegsche-ider, W., Das Sarma, S., ve Vandersypen, L. M. K. (2017). Quantum simulation of a Fermi–Hubbard model using a semiconductor quantum dot array. Nature, 548(7665):70–73. doi:10.1038/nature23022.

9.Guéret, P., Blanc, N., Germann, R., ve Rothuizen, H. (1992). Confinement and single-electron tunneling in Schottky-gated laterally squeezed double-barrier quantum-well heterostructures. Physical Review Letters, 68(12):1896–1899. doi:10.1103/PhysRevLett.68.1896.

10.Choi, M. K., Yang, J., Hyeon, T., ve Kim, D.-H. (2018). Flexible quantum dot light-emitting diodes for next-generation displays. npj Flexible Electronics, 2:10. doi:10.1038/s41528-018-0023-3.

11.Griffiths, D. J. ve Schroeter, D. F. (2018). Introduction to Quantum Mechanics. 3rd ed. Cambridge University Press.

12.Hanson, R., Kouwenhoven, L. P., Petta, J. R., Tarucha, S., ve Vandersypen, L. M. K. (2007). Spins in few-electron quantum dots. Reviews of Modern Physics, 79(4):1217–1265. doi:10.1103/RevModPhys.79.1217.

13.Feynman, R. P. (1982). Simulating physics with computers. International Journal of The-oretical Physics, 21(6–7):467–488. doi:10.1007/BF02650179.

14. Georgescu, I. M., Ashhab, S., ve Nori, F. (2014). Quantum simulation. Reviews of Modern Physics, 86(1):153–185. doi:10.1103/RevModPhys.86.153.

Scheduling as an Access Barrier in Youth Sport: A Structured Evidence Review and Preliminary Planning Tool By Ranbir Goel

Abstract

Children and youth benefit, through organized sport, physically, psychologically and socially. But there is a disparity in access to organised sport, especially among young people of low socio-economic status. Previous studies have defined a set of factors—financial cost, transport, parental time, school demands, household demands, information, and social support—that are interrelated influences on participation and retention. This is a systematic evidence review to explore the role of scheduling in this broader system of barriers. The review concludes that scheduling should be considered to be a component of institutional access conditions, not the only factor responsible for dropouts. The relationship between session times and employment, school commitments, childcare, transport availability, safety issues and other children in the household. The review also highlights a practical challenge: whilst written research documents these barriers it is often not easy to make information on family availability available for use in programme scheduling. The article presents a first step evidence-informed planning tool to address this gap that consists of five stages: mapping availability, identifying competing demands, assessing transport and equity, piloting alternative schedules, and monitoring participation. It is not offered as a validated scale or causal model. It is a practical approach for low resource programmes, and a foundation for future empirical testing.

Keywords

Social and behavioural sciences, Sport Sociology, Youth Sport, Scheduling, Participation Equity

Introduction

People typically think of the relationship between organised sport and physical activity, social development, psychological well-being and community participation. To assess the psychological and social benefits of sport participation among youth, a systematic review of research with children and adolescents was conducted in which sport participation was found to be associated with psychological and social benefits, such as improved self-esteem and social interaction. The review also highlighted that a lot of the evidence was cross-sectional, meaning there would not be good causal inferences to make between participation and later outcomes (Eime et al.). Access is therefore a precondition for the benefits of organised sport. A programme may not be beneficial for a child if the family won't be able to afford, reach, understand, or attend it regularly. Organised sport participation is unequally distributed by households. Studies with youth have found that youth who come from low-income families are less likely to participate and encounter challenges related to costs, transportation, sense of belonging, and access to appropriate opportunities (Tandon et al.).

The issue is often cited as a monetary issue. Families with limited resources may be unable to register due to registration fees, equipment, uniforms or travel expenses. However, even with financial support there are other obstacles. Even when parents receive financial assistance, they may still have to deal with transportation, work schedules, time restrictions and other logistical issues (Clark et al.). Scheduling is a key component to this access challenge. The day and time of a session affects attendance, ability to provide transport, safety of the route and the ability to attend while being able to combine childcare, school and household duties and work with attendance. A programme can be formally available to all children yet actually only available to families that have flexible time and reliable transport.

Scheduling also constitutes an organisation choice. The programme coordinator can choose the time of the sessions based on the availability of the venue, the preferences of the staff, the way it has been done in the past and the expectations of when children have free time. The decisions may not be taken with an intention to exclude, but may have different effects. Children whose parents or guardians work late may be excluded from a weekday afternoon. If it is a late session there may be issues for families who don't have private transport. For communities with seasonal labour, migration, examinations, and/or childcare changes, a fixed schedule may not be appropriate.Low-income parents cited these barriers as being interconnected in research. The findings from this survey indicate that sport participation is more than just a matter of wanting to participate; there are issues that arise from parents' feelings about sport. It also depends on the availability of the resources and conditions of the household and how well they meet the needs of the organisation.

This is a particular problem for community programmes that serve low income groups. Limited funding, staff, facilities and administrative capacity are common to such programmes. Possibly can't provide many venues, or more than one session per day. For this reason, they need to use tools that can assist them in making better decisions with a limited amount of resources. This article provides a systematic review of the research on youth-sport participation, dropout, parental facilitation, transport, socioeconomic inequality and access to programmes. It asks: Who are the individuals, household, organization and environmental barriers that impact participation and retention in organized youth sport for low-income families?

What are the interaction effects between scheduling and transport, work, school, child-care, financial and informational situations?Which evidence-informed planning domains might be useful for community sport programmes to recognise and minimise access barriers related to the planning of their programmes?The article has two contributions. First, it brings together the evidence that scheduling exists in a larger socio-ecological context of constraints. Second, it is suggesting a preliminary planning tool for community sport programmes, based on evidence, to assist in determining what the scheduling risks are, to seek input from the families, to experiment with other scheduling options and to track participation.The tool is not offered as a tested instrument to assess the effectiveness of a programme. It won't say that changing sessions will not result in dropouts. Rather, it offers a practical approach which can be tested and improved via future work.

Review Method

Review design

A structured evidence review and synthesis of narratives were used in this study. It aimed to uncover mechanisms that were likely to recur in the relationship between scheduling and youth-sport access and to develop a first step planning tool based on these mechanisms.

The article does not state that it is a systematic review or meta-analysis (Snyder; Page et al.). Although the description of the present review is the more accurate "structured evidence review," the term systematic review is used throughout the text for ease of reference.

The results of the review are thematic; since pooled effect estimates were not calculated (Campbell et al.). If the review employs other methods of synthesis, it is critical that the reader is able to differentiate the evidence from the authors' interpretation of the evidence.

Search strategy

The evidence search focused on academic research concerning:

- Youth and organised sport participation.
- Sport dropout and retention.
- Low-income families and sport.
- Parental facilitation of children's participation.
- Scheduling and time constraints.
- Transportation and neighbourhood access.
- Socioeconomic inequality in sport.
- Community sport programmes.

Search terms included combinations of "youth sport participation," "organised sport dropout," "low-income families," "parental support," "sport transport barriers," "sport scheduling," "time barriers to sport," "community sport access," "socioeconomic inequality in sport," and "sport participation retention."

The search was iterative. Broad topic searches were used to identify relevant studies, followed by targeted searches for specific articles and research themes. The final evidence base consisted of peer-reviewed research and review articles addressing youth-sport participation, access, socioeconomic inequality, parental facilitation, transport, and retention.

Inclusion criteria

The evidence was deemed relevant if it could satisfy most of the following criteria:

- It involved children, adolescents or youths.
- It looked at organised, community, school or club sport.
- Covered participation, drop-out, retention, access to or barriers to.
- It analyzed families with low income or socioeconomically diverse families.
- It took into account parental, house, organisation, neighbourhood, or transport circumstances.

- It offered findings pertaining to scheduling or programme availability.
- Qualitative studies, quantitative studies, longitudinal studies, systematic reviews and practice-oriented studies that were relevant were featured in the review.

Evidence analysis

Information was analysed in the selected studies for the following:

- Type of participation barrier found.
- The altitude of the barrier's operation.
- The importance of parenting/caregivers.
- The connection between schedules and other constraints.
- Programme organisation implications.
- There are some limitations of the evidence.

The results were presented in the following five themes:

- Participation inequality due to socioeconomic factors.
- Lack of time and scheduling issues.
- Travel, distance and safety.
- Financial, information and social obstacles.
- Consequences for programme planning.

Analysis was conducted with regard to the relationship between barriers and not treating them as single variables. For instance, a transportation problem could turn into a scheduling problem, if the public transportation is not available after a specific hour. A work conflict can turn to a childcare issue when a child care provider is unable to make it to work to pick up a child.

Relevance and quality assessment

The evidence was evaluated on its relevance to the research questions. The studies were categorized according to:

Direct evidence: research studies that focus directly on the relationships between youth sport and low-income families and participation or access.

Supporting evidence: Research on one or two of those areas that offer findings/evidence. Contextual evidence: Reviews or studies that present theoretical or policy background information that is not directly related to scheduling.

No numerical quality score was created to represent the studies. They were interpreted in terms of their variations in design. A longitudinal study may give evidence related to predictors of participation and drop out; and a qualitative interview study may give detailed information regarding family experiences. The types of evidence address various questions and should not be considered to be the same.

Results and Discussion

Socioeconomic inequality in youth-sport participation

This assumption that organised sport is of equal 'accessibility' once a programme has been made available is not supported by the evidence. Participation reported by children from lower-income families is lower, and they face challenges in participation due to cost, transportation, welcome, and opportunities (Tandon et al.). A study on the exclusion of young people from sporting experiences revealed that there was more to social exclusion than the financial cost. The results indicate that a programme may have an open but may also have an implicit message for young people from lower income groups that they are not expected to develop their sports skills from an early age, as well as lower cultural capital (Kingsley and Spencer-Cavaliere). There is also evidence that participation and dropout occur as a result of a number of levels of influence. Using longitudinal data, Vella, Cliff, and Okely investigated the relationship between organised-sport participation and drop-out, and identified relationships with household income, education of parents, language, parental support, and sporting experiences (Vella et al.). This is a significant finding for the present review as it questions only the motivational explanations. Children's preferences and experiences are important, but are constrained and shaped by opportunities and challenges in families, organisations, neighbourhoods and wider socio-economic contexts.

Time, employment, and scheduling conflicts

Time is an integral component of organised sport. Just the length of a session is not enough to qualify for participation. Families might have to involve in preparation, transportation to venue, waiting, equipment provision during the activity and return home. The demands are apportioned differently among households.A caregiver might want to help a child participate in the program, but may not have the flexibility to drive the child to the program at the same time of the week (van Leeuwen et al., "Facilitating Children’s Club-Organized Sports Participation"). Another cause of scheduling conflicts may be when multiple children schedule an activity that will interact with another child's activity. If there are more than one child in a family, the caretaker may need to make decisions as to which child can go, particularly when activities take place at a similar time or place.But these limitations can make it harder for families to retain even if they register to begin with. A household can attend at one time and then drop out at another time due to work, school, and/or household circumstances. It is possible that drop-out is due to changing feasibility, and not changing interest.A set time schedule can also lead to inequitable effects. A weekday afternoon will work out for children when their caregivers are available to them during the day but not for children when their caregivers work late in the day. For some families, it will be easier to attend the late evening session but not for those that rely on public transportation.The question is when most families would like to attend, not just the question. The real question is which families don't qualify for each proposed schedule.

Transportation, distance, and safety

Scheduling is closely related to transportation. There can be a mismatch between the transport needs of sport involvement and parental transport resources. Barriers identified in the research include the length and/or length of unsafe routes, no private vehicle, limited public transportation, and inadequate social networks for shared transportation (van Leeuwen et al., "Facilitating Children's Club-Organized Sports Participation"). The hours of travel are important. A path that will be OK during the day, may not be OK after dark. Fares and frequencies of public transport may be reduced after a certain time. One day a caregiver might be able to pick up a child, but the next day not because of work or family obligations.This therefore should be done in conjunction with the venue and schedule. Even if a venue is geographically close, it may be unavailable because the path may be unsafe, or because of the end of the public transportation service.The study also showed that lower income parents were more likely to report barriers to sports/recreation facilities. The study also demonstrated that access was not consistently related to income, nor was neighbourhood deprivation, implying a more complex relationship (Harrington et al.). This discovery has repercussions that are of value in practice. Accessibility should not be measured just on the basis of distance. It should also take into account the time it takes for the journey, cost, safety, the necessity of an adult to accompany and the availability of transport at the start and end of the session.

Financial assistance does not remove all barriers

Financial constraints are significant even if the fees to join programmes are lowered or eliminated. Families may be required to cover transportation, equipment, uniforms, food or communication costs.Clark et al. analysed a Canadian program that supported low-income families to participate in sport. They also found that financial support enabled families to support their children's participation; however, there were still multiple challenges such as lack of time, transportation, and enrollment expenses (Clark et al.). This indicates that all access to a programme should not be assessed by the cost of the programme entry fee. The participation burden is the sum of the direct costs, travel costs, time, information and practical work needed by the caregivers.Financial support can also be lacking if families do not know about financial support or are not able to complete application procedures. Van Leeuwen and colleagues reported that information that was understandable, trustworthy, and efficient in the referral process was important in determining whether low-income parents would seek financial support for their child's participation (van Leeuwen et al., "Acquiring Financial Support"). Information is thus a component of the accessibility of the schedules. Families should be aware of session times, how schedules will be changed, the equipment needed to attend sessions, and how to register, as well as if there is transport or financial support for families.

Parent and household responsibilities

Adults are often responsible for organising youth-sport participation. Registration, communication, transport, payment, equipment and attendance may be the responsibility of the

parents/caregivers. These responsibilities can prove to be obstacles if there is a shortage of time or support in the home.The results by Van Leeuwen et al. illustrate that parents' facilitation is not simply an activity. It is composed of several tasks which need to be coordinated.The timing of one child's activity could also impact on other members of the family. During the session a caregiver may be required to supervise siblings, cook food, walk between sites, and do chores around the house. The invisible tasks which can be done in attendance are often not captured.A programme which sees repeated absenteeism as a lack of commitment may thus be misreading the issue. The absence could be a sign that the schedule is not affordable as it requires an amount of unpaid care work that a household is not able to provide.

Barriers accumulate across levels

The evidence indicates that there is no common obstacle to non-participation. Instead, barriers accumulate.

Low income families may face:

- Registration fee or equipment fee.
- Irregular employment.
- Limited transport.
- Extended or unsafe journey times.
- Childcare responsibilities.
- Multiple children's schedules.
- Difficulty accessing information.
- Sense of feeling excluded or of not belonging.
- School / Exam requirements.

These conditions can be reinforcing to each other. Attendance may be delayed due to limited transportation. Travel time may be a barrier to work or child care. Attendance is likely to be irregular due to irregular work. Families might not be able to find alternative programmes, because of lack of information.

Person-environment mismatch concept is helpful in the fact that it does not place all the blame on the family. The difficulty might lie in the fact that the demands of organised sport can not be met by the resources available in the household (van Leeuwen et al., "Facilitating Children's Club-Organized Sports Participation").

It is, therefore, possible for the same programme to be different for different families. For families who work during the day and have their own car, a set time in the evening might be easier. May not be easy for families with shift work, more than one child, limited transportation or safety considerations.

Evidence domains identified in the review

The evidence was organised into five domains shown in Table 1.

Table 1. Evidence domains and implications for youth-sport scheduling

Evidence domain	Main evidence pattern	Scheduling implication	Limitation
Socioeconomic inequality	Lower-income and lower-affluence families report lower participation and more barriers (Tandon et al.; Vella et al.).	Scheduling should be examined alongside income, support, and material access.	The strength of the relationship varies across contexts.
Time and work	Work schedules, single parenthood, multiple children, and household demands affect participation (van Leeuwen et al., "Facilitating Children's Club-Organized Sports Participation").	Programmes should collect information about actual family availability.	Families may have changing or unpredictable schedules.
Transport and safety	Distance, transport cost, route safety, and travel timing affect access (van Leeuwen et al., "Facilitating Children's Club-Organized Sports Participation"; Harrington et al.).	Venue and session time should be assessed together.	Transport conditions differ by location and season.
Financial and informational access	Financial assistance may not remove time, transport, and information barriers (Clark et al.; van Leeuwen et al., "Acquiring Financial Support").	Programmes should assess total participation burden and communication.	Studies use different definitions of financial assistance.
Belonging and exclusion	Material resources, cultural capital, and feelings of welcome influence participation (Kingsley and Spencer-Cavaliere).	Scheduling should be included in a wider inclusion strategy.	Scheduling alone cannot address all forms of exclusion.

Evidence-to-tool development

The planning tool was developed by translating the recurring evidence domains into programme questions. The process involved four steps:

1. Identify barriers reported in more than one source.
2. Determine whether the barrier is individual, household, organisational, neighbourhood, or structural.
3. Identify the scheduling decision affected by the barrier.
4. Convert the finding into a practical question or monitoring indicator.

For example, studies reporting work and family scheduling conflicts support the planning domain "family availability." Studies reporting transport and route difficulties support the domain "transport and safety." Studies reporting financial assistance alongside continued time and transport barriers support a multi-domain approach rather than a fee-only approach.
The resulting tool is not a validated scale. It is an evidence-informed framework for programme planning.

Preliminary Planning Tool

Purpose

The planning tool is designed for community programmes that need a practical method for identifying schedule-related access barriers. It is intended for programmes with limited funding, staff, facilities, and technical capacity.The tool does not assign a numerical effectiveness score. It guides a cycle of information gathering, planning, testing, and revision.

The evidence-informed scheduling planning cycle depicted in Figure 1 is a practical process for community youth-sport programmes. It starts with a family availability map and competing demand identification, and progresses to transportation, safety, financial, and information availability assessment. Programmes can then try out different schedules, observe participation and barriers reported, and adjust their arrangements accordingly. This cycle is iterative as the availability of family and access conditions can change over time. This is not about blame for absence being placed on families but is about determining whether there is a problem in the design of the programme which is creating avoidable barriers.

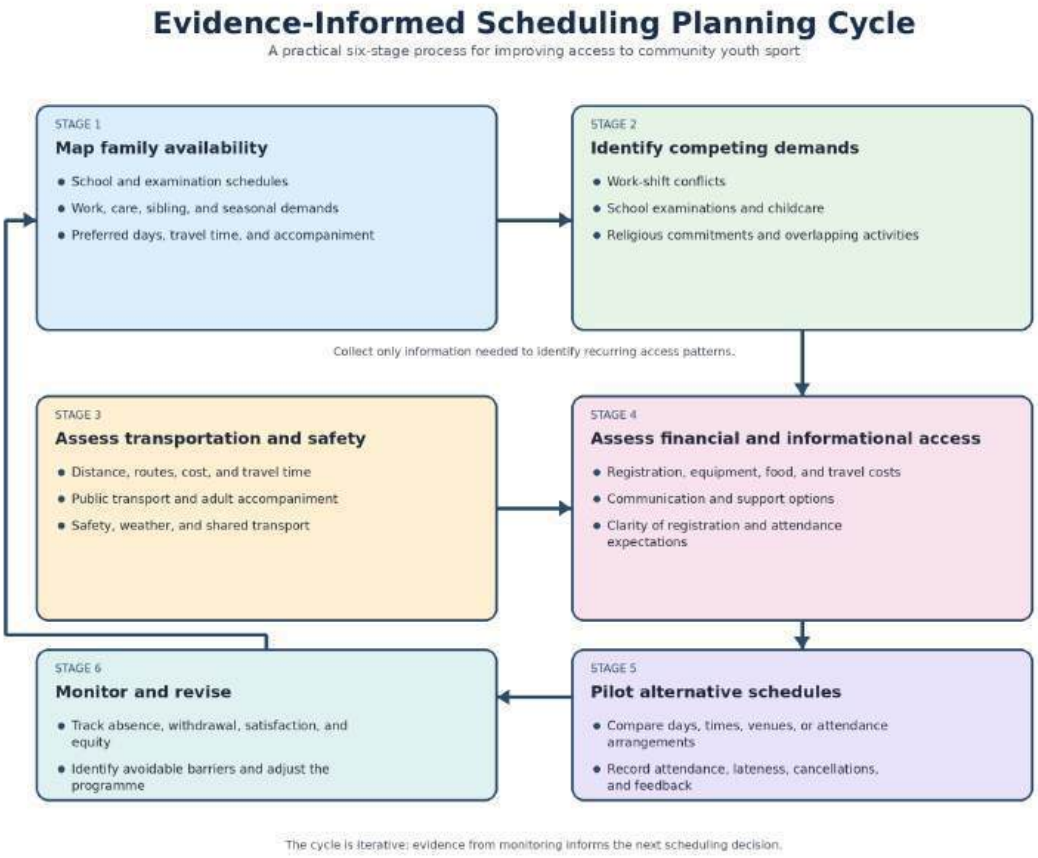


Figure 1. Evidence-informed scheduling planning cycle. The six-stage cycle translates evidence on family availability, competing demands, transport, safety, financial access, information, and participation monitoring into a practical planning process for community youth-sport programmes.

Figure 1 was created using Python and the Pillow software library. The layout and wording were developed with assistance from Anara. No figure was reproduced from another publication.

Table 2. Preliminary planning tool for identifying schedule-related access barriers.

Planning domain	Evidence to collect	Question for coordinators	Possible response
Family availability	Work, school, childcare, and household schedules	When can participants attend consistently?	Offer more than one feasible time where possible.
Competing demands	Employment, examinations, care work, religious activities, and sibling schedules	What regularly conflicts with the proposed session?	Change the time, frequency, or attendance arrangement.
Transportation and safety	Travel time, cost, route, safety, and available transport	Can participants reach the venue at the proposed time?	Change the venue or coordinate shared transportation.
Financial access	Fees, equipment, uniforms, food, and travel costs	Which costs remain after registration support?	Provide equipment, transport, or referral support.
Information and trust	Registration instructions, language, communication channels, and schedule notices	Can families understand and use programme information?	Simplify communication and use multiple channels.
Equity monitoring	Differences by income, gender, disability, family structure, and neighbourhood	Does the schedule favour some groups over others?	Consult affected families and revise arrangements.

Practical Implications

There are implications for programme coordinators, schools, community organisations and local authorities in the use of the planning tool.

- Firstly, scheduling needs to be seen as part of access planning and not as an administrative process that is overlaid on venue availability.
- Secondly, information should be obtained directly from families through programmes. Staff assumptions about availability might not be accurate if they don't share the work, transport and caring situations of the participants.
- Thirdly, a single indicator like attendance in general should not be used in a programme. While there may be no overall changes in attendance, there can be persistent absence or withdrawal of certain subgroups.
- Fourth, if resources permit, programmes need to take flexibility into consideration. This can involve changing session time, having more sessions per week, offering closer

venues, providing shared transport, allowing flexibility of attendance and/or using community organisations or schools.

- Fifth, care should be taken in the interpretation of attendance data in programmes. Frequent absences could indicate that the circumstances with different families change or that there is a transport challenge or simply a scheduling issue – it's not just a lack of motivation.

Last but not least, the tool should be localised. A realistic schedule is one that is realistic for an urban environment where public transport is plentiful, but may not apply in a rural environment. One community may have access to a weekday evening while another community may not have access to a weekday evening or may have unsafe or impractical access to a weekday evening.

Limitations

- First, there are varying definitions of participation, dropout, retention, access and socioeconomic disadvantage in the studies. Because of these differences universal numerical thresholds for effective scheduling cannot be developed.
- Second, the review does not conclude that a reason for dropping out is due to scheduling. It recognises the potential role of scheduling as an important institutional factor, which interacts with income, transport, work, school, childcare, safety and programme quality.
- Thirdly, the planning tool has yet to be tested. It needs to be empirically tested in terms of its feasibility, usefulness and impact on participation.

Future Research

The planning tool should be assessed in actual community sport programmes in the future. A pilot evaluation can be used to gather baseline data prior to the introduction of schedule changes and check the attendance following the evaluation.

Some outcomes may be:

- Attendance rate.
- Repeated non-attendance.
- Late arrival.
- Withdrawal or dropout.
- New registration.
- Family satisfaction.
- Reported issues with transportation and scheduling.
- Differences in participation between social groups.

A mixed methods design would be an appropriate method. Attendance records may reveal trends and patterns and interviews or focus groups may shed light on some of the reasons behind a specific attendance schedule working for some families but not others.

Seasonal variation also should be the focus of future studies. Household schedules might vary with agricultural, rural, migrant or informal-work periods, in the planting, harvesting, examination, weather and employment periods.Feasibility testing of the tool should be carried out. Researchers should be interested in the ability of the coordinator to gather the necessary data, the family's willingness to feed back and the programmes' flexibility to make decisions from the data.Equity needs to be assessed in the tool as well. A schedule can help to boost attendance for everyone, but results in some groups being excluded. Evaluation should, therefore, look at income, gender, disability, family structure, neighbourhood, language and access to transport differences.Children's views should also be taken into account in future research. Parents can say what barriers they see for getting to and from school, and children can say what barriers they see for their involvement in school, in terms of peer relationships, pressure in school, safety, fun, and belonging. Both perspectives need to be taken into account in an inclusive planning process.

Conclusion

There are individual, family, organisational, neighbourhood and socioeconomic factors that influence youth-sport participation. Low-income families face financial, transportation, time, scheduling, information, planning and parental support challenges, all of which have been shown to present a barrier to participation (Tandon et al.). Thus, the meaning of ‘scheduling' must be interpreted as an institutional condition of access, rather than as a neutral administrative decision. Session times may interact with work, school, child care, transportation, safety and household responsibilities. They can be easier for some families and more difficult for others to participate.This structured evidence review created a first draft of an evidence-informed planning tool with sidomains: family availability, competing demands, transportation and safety, financial access, information and trust, and equity monitoring.The tool is not a validated scale and it does not show that changing the time schedule for the classes will guarantee that there would be no dropouts. It is designed to support the community programmes in making more transparent and locally informed scheduling decisions that will be more sensitive to inequalities of access.
The central argument is that policies to increase youth-sport participation will not only need to lower costs of participation or maintain high levels of motivation, but will also need to address these issues. It needs to take into account the practical circumstances which enable families to attend consistently. Scheduling is one of these conditions, and should be considered as part of a wider community sport access agenda.

Code and Data Availability

This article did not generate or analyse a new primary dataset and did not use study-specific computational code. The evidence reviewed in the article is identified in the reference list.

AI Acknowledgement

During the preparation of this work, the author used Sci Space to assist with literature organisation. After using this tool, the author reviewed and edited the resulting content as necessary and took full responsibility for the publication's content.

Acknowledgements

None.

Authors

Ranbir Goel is a high school student at Delhi Public School R.K. Puram. His interests include youth sport, social inequality, community participation, and programme design.

Works Cited

1.Eime, Rochelle M., et al. “A Systematic Review of the Psychological and Social Benefits of Participation in Sport for Children and Adolescents: Informing Development of a Conceptual Model of Health through Sport.” International Journal of Behavioral Nutrition and Physical Activity, vol. 10, 2013, article 98. https://doi.org/10.1186/1479-5868-10-98.

2.Tandon, Pooja S., et al. “Socioeconomic Inequities in Youth Participation in Physical Activity and Sports.” International Journal of Environmental Research and Public Health, vol. 18, 2021, article 6946. https://doi.org/10.3390/ijerph18136946.

3.Clark, Marianne, et al. “Parental Perceptions of a National Program That Funds Sport Participation for Low-Income Children and Youth in Canada.” Leisure Sciences, vol. 44, 2019, pp. 1082–1098. https://doi.org/10.1080/01490400.2019.1700573.

4.van Leeuwen, Lonneke, et al. “Facilitating Children’s Club-Organized Sports Participation: Person–Environment Misfits Experienced by Parents from Low-Income Families.” Children, vol. 9, 2022, article 1746. https://doi.org/10.3390/children9111746.

5.Vella, Stewart A., et al. “Socio-Ecological Predictors of Participation and Dropout in Organised Sports during Childhood.” International Journal of Behavioral Nutrition and Physical Activity, vol. 11, 2014, article 62. https://doi.org/10.1186/1479-5868-11-62.

6.Kingsley, Bethan, and Nancy Spencer-Cavaliere. “The Exclusionary Practices of Youth Sport.” Social Inclusion, vol. 3, no. 3, 2015, pp. 24–38. https://doi.org/10.17645/si.v3i3.136.

7.Harrington, Daniel W., et al. “Parents’ Perceived Barriers to Accessing Sports and Recreation Facilities in Ontario: Exploring the Relationships between Income, Neighbourhood Deprivation, and Community.” International Journal of Environmental Research and Public Health, vol. 14, 2017, article 1272. https://doi.org/10.3390/ijerph14101272.

8.van Leeuwen, Lonneke, et al. “Acquiring Financial Support for Children’s Sports Participation: Co-Creating a Socially Safe Environment for Parents from Low-Income Families.” Children, vol. 10, 2023, article 872. https://doi.org/10.3390/children10050872.

9.Snyder, Hannah. “Literature Review as a Research Methodology: An Overview and Guidelines.” Journal of Business Research, vol. 104, 2019, pp. 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039.

10.Page, Matthew J., et al. “The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews.” Systematic Reviews, vol. 10, 2021, article 89. https://doi.org/10.1186/s13643-021-01626-4.

11.Campbell, Mhairi, et al. “Synthesis without Meta-Analysis (SWiM) in Systematic Reviews: Reporting Guideline.” BMJ, vol. 368, 2020, article l6890. https://doi.org/10.1136/bmj.l6890.

Alzheimer's Disease, Air Pollution, and Biomarkers: A Mechanistic Analysis

By Dhiya Prabhuram[1] and Vanitha Nithianandam, Ph.D.[2]

Abstract

Air pollution has become a major global health concern with increasing discoveries and evidence linked to neurodegenerative diseases, particularly Alzheimer's disease. While existing research and studies have shown to examine the role of individual pollutants such as PM2.5, NO_2, O_3, and their linked pathways, there are still limited studies specifically comparing these individual pollutants and their influence on distinct biological pathways, while linking those to their corresponding biomarkers for a more precise comparison and outcome. Through a secondary data analysis using peer-reviewed studies, air quality and biomarker databases, the levels of each pollutant were compared with and linked to specific changes in biomarkers such as amyloid pathology (Aβ42), tau pathology (t-tau), axonal injury (NfL), and neuroinflammation (YKL-40 / GFAP). Findings were then organized into two tables, Table 1 recording the pollutant and mechanism relationships, and Table 2 recording the strength of Alzheimer's biomarkers. Results from this study indicate that PM2.5 is associated with decreased CSF Aβ42 levels (β= -0.10), and increased NfL levels. NO_2 showed strong associations with tau pathology (t-tau) with increasing value of (β = 0.34), as well as an increase in NfL (β= 0.36). O_3 data showed strong associations with neuroinflammation (YKL-40 / GFAP), neuronal/axonal injury, and some increased levels of amyloid accumulation. However, the data collected for O_3 did not include any statistical values to show their relationship and there were limited studies that used human CSF data. Overall, all three pollutants were not directly measured with neuroinflammation markers, but were all recorded to be linked with increased neuroinflammation levels. The biomarker data analysis showed tau pathology-related biomarkers and NfL biomarkers had the largest effect sizes, which resulted in a stronger association with Alzheimer's disease development, compared to amyloid alone, where it had the smallest effect size, but one of the strongest associations with Alzheimer's disease development. Ultimately, these findings and results emphasize the importance of linking environmental exposure data with biomarker analysis data, when researching its effects on Alzheimer's disease, as it provides a more precise comparison and measurable outcome.

Literature Review

"Air pollution now ranks as the second leading risk factor for death, accounting for 8.1 million deaths globally," according to HEI's new State of Global Air 2024 Report (Health Effects Institute). Air pollution has become one of most prevalent environmental health concerns in the world, in which it's known for its severe effects on the lungs, heart, brain, and overall life expectancy. Over the past several years, researchers have started to connect long-term exposure

[1]Santa Susana High School, Simi Valley, CA
[2]Yale School of Medicine, Stephen & Denise Adams Center for Parkinson's Disease Research, New Haven, CT

to air pollution to neurodegenerative diseases like Alzheimer's disease. Alzheimer's disease, being one of the most common forms of dementia, is caused by the destruction of brain cells ,which can lead to inflammation and cognitive decline. This disease is characterized by cognitive decline and brain damage. From a biological aspect, Alzheimer's disease is indicated and identified by several key biomarkers. Biomarkers are measurable indicators of a certain condition or disease, in this case, Alzheimer's disease. These markers indicate specific processes or mechanisms that eventually lead to the development of Alzheimer's disease (AD). AD is indicated by several key biomarkers such as amyloid beta plaque accumulation, tau protein phosphorylation, axonal injury (or neuronal injury), and neuroinflammation. Additionally, these biomarkers are often measured in a fluid called cerebrospinal fluid (CSF), which is a fluid that surrounds the brain and spinal cord, providing proper insight into biological and chemical changes and processes that occur in the nervous system. These biomarkers represent the underlying mechanisms that are involved in neurodegeneration and are often utilized to track disease progression, as summarized in a recent peer-reviewed neurological research paper (Lane et al.).

Alzheimer's Biomarkers and Disease Progression

The accumulation of amyloid-beta (Aβ42) is one of the earliest and most common changes in Alzheimer's disease. Amyloid-beta is a protein that accumulates in the brain which forms "amyloid plaques," eventually leading to disruption of communication between neurons. Decreased levels of Ab42 in cerebrospinal fluid are often associated with increased plaque accumulation in brain tissue because most of the amyloid-beta proteins deposit inside the brain tissue, moving away from the cerebrospinal fluid, which also makes this a crucial early indicator of the development of Alzheimer's disease (Abukuri). However, emerging research on neuropathology by Kim and other researchers suggests that tau-related biomarkers show stronger associations with Alzheimer's disease development, compared to amyloid-beta, specifically talking about their roles in neuronal injury and cognitive decline (Kim et al.). Additionally, other research has emphasized the significance of key biomarkers which include neurofilament light chain (NfL), which is a structured protein and when damaged, is released into the CSF, and results as a measurable indicator of disease severity and progression. Increased levels of these biomarkers in CSF further contribute to their role in the development of neurodegenerative diseases (Agah et al.). While markers such as GFAP and YKL-40, which Pelkmans and other researchers explain to be elevated in the dementia phase of AD (Pelkmans et al.), associated with neuroinflammation is referred to as the brain's immune response, where cells release chemicals that can signal and cause swelling and damage to neurons, contributing to neuronal damage (Brockmeyer and D'Angiulli). Ultimately, these different biological mechanisms and biomarkers are shown to be interconnected, yet distinct pathways that all lead to the development of Alzheimer's disease at different levels.

Association with Air Pollutants

Air pollution has appeared to be a significant environmental factor contributing to neurodegenerative diseases, which shifts from traditional age or genetically related factors that are normally seen behind the diagnosis of Alzheimer's. Key air pollutants such as particulate matter 2.5 (PM2.5), nitrogen dioxide (NO_2), and ozone (O_3), compared to other pollutants, are well known for inducing neuroinflammation and oxidative stress, as explained by Brockmeyer and other researchers in a peer-reviewed study on the role of neuroinflammation in air pollution. Additionally, Brockmeyer explains the process by which pollutants enter the brain by entering the lungs, crossing and disrupting the protective blood-brain barrier, and eventually leading to overall neuronal damage. (Brockmeyer and D'Angiulli).

Among these air pollutants, PM2.5 has been shown to be the most extensively studied by most environmental science researchers. A large population-based cohort study discovered that increased exposure to PM2.5 may be associated with changes in amyloid-beta related biomarkers (Casey et al.). However, this study reported less consistent findings and results when examining the association between other biomarkers, like tau, which shows that the effects PM2.5 has on Alzheimer's disease development is not uniform across each and every aspect of the disease progression. These findings suggest that PM2.5 effects across different biological pathways are unclear and have not been researched enough. Moreover, nitrogen dioxide, a traffic pollutant, has been examined for its particular role and effects on neuronal damage. Specifically, some studies have shown that NO_2 exposure is possibly associated with neuronal/axonal injury and NfL-related biomarkers (Kim et al.). Similarly, to PM2.5, these findings are not uniform across different biological processes and vary based on the type of biomarker, which is crucial to note here because there is a lack of studies that conduct a comparison between the different effects of these individual pollutants through linking them to their corresponding biomarkers. Lastly, ozone is examined as one of the main environmental factors that contribute to neurodegeneration, although it is less frequently researched in human populations in some published studies. Most research involving ozone and air pollution suggests its effects on oxidative stress and neuroinflammation. However, this pollutant presents a more complex and less consistent association with Alzheimer's disease specifically, compared to the other two pollutants. Many published studies focus on ozone's relationship with dementia or neurodegenerative diseases in a general aspect. Across many studies, this pollutant has been shown to have strong and consistent associations with two key biological pathways, neuroinflammation and oxidative stress. A study on the mechanisms of ozone-induced neurotoxicity specifically shows how this happens, by generating ROS, which stands for reactive oxygen species, which can eventually lead to damage to neuronal cells by disrupting mitochondrial function and lipid peroxidation (Marin-Castañeda et al.). A very recent study, which examines the associations between air pollutants and the markers of neuroinflammation, found no significant association between ozone and core AD biomarkers. However, numerous other studies have shown that ozone contributes to some amyloid beta accumulation, but absolutely no associations with the biomarker tau.

Moving on from the individual effects of these air pollutants, more studies have constantly emphasized the role of neuroinflammation and oxidative stress as they are key

pathways that are linked with Alzheimer's disease. Leading researchers in neuropathology and neurotoxicity departments explain that individuals who are exposed to high concentrations of air pollution show early signs of Alzheimer's pathology. Moreover, in Lane's study, they explain that through sustained inflammation and oxidative stress, air pollution remains to be a major contributor to these neurological processes, reinforcing the idea that air pollution cannot be affected through one biological pathway, but rather multiple interactions with biological processes.

Despite the growing body of research in this field, several limitations remain. Many of these studies focus on an individual pollutant or group of pollutants without comparing their effects on different biological processes in one study. All biological processes are indicated by biomarkers, which are crucial in my study as it provides measurable evidence and a more precise comparison between these air pollutants and biological processes. There are limited studies that focus on and compare the strengths of these pollutants and biological effect relationships, while linking those effects with measurable evidence of their corresponding biomarkers. Therefore, a clear gap exists on comparing how each pollutant influences those key biological pathways from a quantitative data analysis. This leads me to my research question: To what extent do specific air pollutants influence the biological mechanisms associated with the development of Alzheimer's disease as measured through related biomarkers?

Before diving into my methodology, from identifying the gaps in existing literature and forming a research question, I hypothesized that specific air pollutants will differentially influence biological mechanisms associated with Alzheimer's disease while being reflected through its corresponding biomarkers. Specifically, I believe that PM2.5 and NO_2 will have the strongest associations with mechanisms with the strongest correlation with Alzheimer's development, resulting in those pollutants having the most harmful effects compared to O_3.

Methodology

This project uses a quantitative correlational design and secondary data analysis to explore and analyze the relationships between 3 types of air pollutants, PM2.5, NO_2, O_3, and key Alzheimer's biological pathways while linking them to their corresponding biomarkers for a more precise comparison. Rather than collecting primary data, this research design approach is aimed to analyze existing and published studies and databases to properly evaluate how each air pollutant influences these biological processes linked to measurable indicators of Alzheimer's disease.

The primary sources used for analysis and data collection came from academic databases such as PubMed, ScienceDirect, the EPA Air Quality database, the Alzbiomarker database, and academic journals. The independent variables in this research are the long-term exposures to each of the three air pollutants as well as the air pollutants themselves. This exposure data was collected from the EPA (United States Environmental Protection Agency) air quality database. This database collects and monitors exposure data from across the country. The exposure data values were collected to mirror realistic environmental conditions, rather than extreme

experimental conditions. The dependent variables involved in my research are the Alzheimer's biomarkers such as Aβ42 (amyloid pathology), tau-p181 (tau pathology), NFL (neurofilament light chain or neuronal injury), and YKL-40 / GFAP (neuroinflammation). These specific biomarkers were chosen in my study because they correspond to the key biological processes researched in Alzheimer's disease and serve as measurable indicators for these processes, providing a precise comparison for the pollutant and mechanism relationships.

Quantitative data were collected from peer-reviewed studies and academic journals that reported the associations between pollutant exposure and mechanism levels, using statistical values. The data for PM2.5 were collected from large-scale human cohort studies from ScienceDirect by Alemany and other researchers, which specifically looked at the statistical relationships between PM2.5 and NfL and Aβ42. The data for NO_2 was collected from those studies as well, except those studies reported to have associations with NO_2 exposure and tau pathology, using statistical values. The specific biological mechanism data collected for ozone came from studies that used animal experimental models by Zhang and other researchers. Data were also collected from general peer-reviewed studies focusing on the association between ozone and neuroinflammation. However, this pollutant had some limitations due to the limited availability of studies that recorded mechanism and pollutant exposure data from human CSF, as well as specific statistical values to support the strength of the associations.

Biomarker data was collected to assess the strength of its association with Alzheimer's disease. The biomarker data was collected from the Alzbiomarker database which organizes decades on fluid (CSF) biomarkers for Alzheimer's disease. The specific data recorded for each biomarker was compiled of statistical significance (p-values), effect sizes, and the direction of AD.

To properly analyze these findings, they were organized into two tables. Table 1 to record and summarize all of the pollutant-mechanism relationships, including any statistical values that came from the study. While Table 2 further supported those findings by recording all of the appropriate effect sizes and statistical values for each biomarker that was researched in my study and were shown to have an association with the air pollutants. Additionally, this research approach was informed by epidemiological studies and papers, but was instead adapted and refined into a secondary data analysis approach to perform a comparative analysis study. The purpose of creating these two tables is to properly compare the different influences air pollutants can have on Alzheimer's related mechanisms and linking it to how strongly it is associated with the development of Alzheimer's disease.

Results

Table 1: Mechanism Association Between Air Pollutants and Alzheimer's Biological Pathways

Pollutant	Amyloid Pathology (Aβ42)	Tau Pathology (tau-p181)	Axonal Injury (NFL)	Neuroinflammation (YKL-40 / GFAP)
PM2.5	↓ CSF Aβ42 (β= -0.10)	Not reported	↑ NFL (β= 0.31-0.38)	↑ Proposed mechanism (not directly measured)

NO_2	Not reported	CSF t-tau (β = 0.34)	↑ NFL (β= 0.36)	↑ Proposed mechanism (not directly measured)
O_3	↑Aβ accumulation (animal model)	Not reported	↑ Hippocampal neuronal damage (animal model)	↑ TNF-α, ↑IL-6

* Arrows (↓↑) indicate increased or decreased levels.

Table 2: Strength of Alzheimer's biomarkers (AD vs. CTRL)

Biomarker	Pathway	Effect Size	p-value	Direction of AD
Aβ42 (CSF)	Amyloid plaque accumulation	0.552	p<0.0001	↓
tau-p181 (CSF)	Tau phosphorylation/ neurofibrillary tangle formation	1.867	p<0.0001	↑
tau-total (CSF)	Overall neurodegeneration and neuronal damage	2.477	p<0.0001	↑
NFL (CSF)	Axonal injury/neuronal structural damage	1.983	p<0.0001	↑
YKL-40 (CSF)	Neuroinflammation/microglial activation	1.362	p<0.0001	↑
GFAP (Plasma & Serum)	Astrocyte activation/inflammatory response	1.928	p<0.0001	↑

Overview and Analysis

This study examined the extent to which three different air pollutants: PM2.5 (Particulate Matter), NO_2 (Nitrogen Dioxide), and O_3 (Ozone), are all associated with biological mechanisms in Alzheimer's disease. The independent variables in this study included long-term exposure levels (national level) for each air pollutant. The exposure data was collected from the EPA (Environmental Protection Agency) Air Quality database. These typical exposure concentrations were 8-9 ug/m^3 for PM2.5, 40-45 ppb for NO_2, and 0.07 ppm for O_3. Different concentrations included different units based on the physical properties of each pollutant. PM2.5 is a solid particulate matter, so it is measured in micrograms per cubic meter, which represents its mass concentration in the atmosphere. On the other hand, NO_2 and O_3 are gases, so they are measured in parts per billion (ppb) or parts per million (ppm). The dependent variables in this study included amyloid pathology (Aβ42), tau pathology (tau-p181), axonal Injury (NFL), and neuroinflammation (YKL-40 / GFAP).

PM2.5's results showed the strongest correlation with amyloid pathology in human cohorts. At the typical national exposure level of 8-9 ug/m^3, the data reported that higher long-term exposure to PM2.5 resulted in a 0.10 standard deviation (β= -0.10) decrease in CSF Aβ42 levels. Since the decreased CSF Aβ42 levels resulted in amyloid plaque accumulation in the brain, this finding shows that environmentally relevant exposure levels can clearly contribute to signs of amyloid plaque accumulation. The data reported also that PM2.5 was associated with increased levels of NfL (axonal injury) by 0.31-0.38 standard deviations (β= 0.31-0.38). NFL stands for neurofilament light chain, which is a structural protein that is released into the

bloodstream or CSF (cerebrospinal fluid) when neuronal axons are damaged, resulting in structural axonal injury. While neuroinflammation (YKL-40 / GFAP) was proposed as a biological mechanism in the literature and peer-reviewed studies, inflammatory biomarkers were not directly qualified in the studies as well as direct measurements of the neuroinflammation levels. However, the Alzbiomarker database lists YKL-40 / GFAP as one of the biomarkers, which is strongly associated with neuroinflammation. Overall, PM2.5 shows to heavily influence amyloid pathology and axonal injury pathways at its typical exposure levels.

The results for nitrogen dioxide (NO_2) showed to have significant associations with increased levels of tau pathology and axonal injury. At the typical national exposure level of 40-45 ppb, the data reported that NO_2 was positively associated with a 0.34 standard deviation ($\beta = 0.34$) increase in CSF t-tau levels. Additionally, NO_2 was associated with increased levels of NFL (axonal injury) by 0.36 standard deviations ($\beta= 0.36$). From the reported data in the literature, NO_2 was not shown to have as strong of an association with increased Aβ42 levels as PM2.5 did. As with PM2.5, neuroinflammation was a proposed biological mechanism but not directly measured in the human cohort.

The results for ozone O_3 showed strong associations with amyloid pathology, axonal injury, and neuroinflammation. The data reported for O_3 differed from PM2.5 and NO_2, resulting in some possible limitations in this study. The data collected for this air pollutant were derived from a peer-reviewed study that involved using animal experimental models, rather than human CSF studies. The typical national exposure level was 0.07 ppm, but most of the mechanistic data collected from the animal experimental model study used higher exposure levels, which included 0.25 ppm and 1 ppm. At 0.25 ppm, the data reported increased levels of hippocampal neuronal damage in Table 1. This connects to axonal injury as structural neuronal injury mirrors similar processes as biomarkers such as NFL. At 1 ppm, increased TNF-α and IL-6 levels were observed. TNF-α and IL-6 are proinflammatory cytokines which are established indicators of neuroinflammation, aligning with the neuroinflammation column in Table 1. Additionally, significant increases were observed in Aβ42 levels, resulting in Aβ accumulation in brain tissue. Despite the fact that the animal experimental study collected data in a setting using higher O_3 exposure levels (0.25 and 1 ppm), which are significantly higher than the typical exposure levels, 0.07 ppm, it concluded that O_3 levels are directly proportional to amyloid accumulation, axonal injury, and neuroinflammation. Most importantly, data collected directly from human CSF biomarker-associated studies were not reported or found for O_3 in the peer-reviewed literature. Hence, O_3's influential data is supported by animal experimental mechanistic data instead of larger-scale human biomarker data.

The data observed from Table 2 were collected from the Alzbiomarker database. The data from this table quantifies how strongly each biomarker is associated with Alzheimer's disease by comparing individuals who are diagnosed with Alzheimer's disease and those who are not (healthy/control). The effect size column represents the magnitude of that difference. A greater effect size shows that the biomarker is strongly a characteristic of Alzheimer's pathology. Out of all the biomarkers observed in Table 2, total-tau indicates the greatest effect size, 2.477, showing

that the overall neurodegeneration is one of the most prominent biological mechanisms associated with Alzheimer's. Moreover, NfL and tau-p181 also show large effect sizes. NFL's effect size, 1.983, and tau-p181's effect size 1.867, represent significant tau phosphorylation/tangle formations and axonal injury in individuals with Alzheimer's disease. The two biomarkers strongly associated with neuroinflammation had effect sizes such as GFAP as 1.928 and YKL -40 as 1.362. However, Aβ42 (CSF) shows a moderate effect size, 0.552, and is decreased in Alzheimer's. This decrease still means that Aβ42 still has a strong and harmful association with Alzheimer's, indicating that amyloid-beta becomes deposited into plaques in the brain tissue instead of staying in CSF, lowering the Aβ42 (CSF) levels.

Even though amyloid pathology is widely known to be central to the development of Alzheimer's disease, the smaller effect size of Aβ42 (CSF) does not necessarily mean that it is irrelevant. It suggests that amyloid accumulation possibly occurs during the early stages of Alzheimer's, even years prior to diagnosis. By then, individuals are already diagnosed with Alzheimer's disease and biomarkers such as tau-related neurodegeneration and axonal injury might show significantly greater effect sizes between AD patients and the healthy individuals. All the biomarkers listed in Table 2 show a p-value column, where it is observed that they all share the same p-value ($p<0.0001$). This simply indicates that differences between Alzheimer's disease and healthy control groups are statistically reliable. The "Direction of AD" shows whether each biomarker increases or decreases in Alzheimer's disease compared to the control group. This column is relevant in my findings because it helps determine whether air pollution changes biomarkers in the same direction as Alzheimer's disease. Ultimately, this table serves as a comparison between Table 1, to indicate the extent to which air pollutants influence these biomarkers, which differ in association with Alzheimer's disease.

The air pollutants, PM2.5 and NO_2, show strong associations with tau-phosphorylation and NFL, which are the biomarkers observed to have the largest effect sizes in Alzheimer's disease, which possibly indicates an alignment with high-impact neurodegenerative pathways. O_3 shows the activation of inflammatory and amyloid accumulation primarily through animal experimental models, and does not show any association with tau-pathology, which includes biomarkers with the largest effect sizes. Overall, these findings support my hypothesis that individual air pollutants, particularly PM2.5 and NO_2 will differentially influence biological mechanisms and pathways that are strongly associated with the development of Alzheimer's disease.

Limitations

While this research recognizes the significant correlations between particular air pollutants and Alzheimer 's-related biological mechanisms, there are several limitations that should be considered after analyzing the findings. First, the data collection relies on a secondary data analysis method, utilizing published peer-reviewed papers and publicly available online databases such as the EPA Air Quality Index and Alzbiomarker database, rather than collecting data directly from experimental methods, which might show more accuracy. As a result, this

study does not aim to directly identify the cause of air pollution on Alzheimer's disease. Instead, the analyses are established from observed correlations between air pollutants with biological mechanisms/pathways and their associated Alzheimer's biomarkers. While the exposure results that were recorded in this paper show the typical exposure levels nationally, they provide a broader and general reference for air pollution exposure. It is important to consider that this air quality database does not show individual exposure levels, which are reliant on geography, environmental conditions, and other factors. Additionally, a limitation that should be considered is that in some scenarios, it is not clear whether an individual has already experienced symptoms or is in the early stages of Alzheimer's, since Alzheimer's gradually develops over the span of a few years. Other factors such as age, medical history, and genetics, were not taken into account and would make it more difficult to determine how strongly pollutant exposure alone can influence these biological stages. Another limitation that should be considered is that some of the mechanism findings, specifically for the ozone pollutant, come from an animal experimental model study. Even though these animal models show proper insights and findings focusing on biological pathways like neuroinflammatory pathways and oxidative stress, these findings may not align with human-related results. Regardless of these limitations, this paper still provides a useful and insightful understanding of the correlation between pollutants' association with biological mechanisms and the development of Alzheimer's.

Summary and Research Implications

Regardless of these constraints, these findings and results contribute to a growing body of research that focuses on and observes the effects of environmental conditions on the development of neurodegenerative diseases. This study aims to provide an additional understanding of the comparison of different air pollutants and their influence on crucial neurodegenerative pathways. The findings from my data collection, such as PM2.5 and NO_2 which align with neurodegenerative pathways, axonal injury and tau-phosphorylysation, the pathways that are associated with the biomarkers, tau-p181 (CSF), tau-total (CSF), and NFL (CSF), which have the largest effect sizes compared to the other existing biomarkers, meaning that they are strongly associated with the development of Alzheimer's disease. These comparisons serve as a foundation for evaluating and analyzing which environmental exposures may have the greatest biological impact related to Alzheimer's pathology. Public health policymakers could potentially use this research to enforce stricter and safer health policies as well as testing different methods to reduce pollution in areas where exposure levels are highly toxic and with higher rates of neurodegenerative diseases in individuals. Moreover, urban planners and environmental officials could focus on reducing emissions that come from traffic and industrial sources that release or produce these toxic pollutants into the atmosphere, especially PM2.5 and NO_2, according to my findings. For researchers particularly specializing in environmental science and neurobiology, these findings emphasize the significance of focusing on and combining air pollution exposure data with mechanism and biomarker-related research, which not a lot of researchers tend to focus on when researching neurodegenerative diseases.

Future research could build off of this, while using more accurate methods, by recording individuals' air pollution exposure over time and measuring and observing changes in Alzheimer 's-related biomarkers in those same individuals. Ultimately, these findings reveal public health implications, such as recognizing specific environmental factors that contribute to neurodegenerative diseases, which could potentially support efforts to reduce harmful exposure and to take proper precautions when living in geographic areas with higher levels of exposure, as well as educating different strategies and approaches aimed to protect long-term brain health and prevent early stages or symptoms of neurodegenerative diseases.

Works Cited

Abukuri, Daniel Naawenkangua. "Novel Biomarkers for Alzheimer's Disease: Plasma Neurofilament Light and Cerebrospinal Fluid." International Journal of Alzheimer S Disease, vol. 2024, May 2024, pp. 1–15, https://doi.org/10.1155/2024/6668159.

Agah, Elmira, et al. "CSF and Blood Levels of Neurofilaments, T-Tau, P-Tau, and Abeta-42 in Amyotrophic Lateral Sclerosis: A Systematic Review and Meta-Analysis." Journal of Translational Medicine, vol. 22, no. 1, 2024, p. 953, https://doi.org/10.1186/s12967-024-05767-7.

Alemany, Silvia, et al. "Associations Between Air Pollution and Biomarkers of Alzheimer's Disease in Cognitively Unimpaired Individuals." Environment International, vol. 157, Dec. 2021, p. 106864, https://doi.org/10.1016/j.envint.2021.106864.

"AlzBiomarker | ALZFORUM." Www.Alzforum.Org, https://www.alzforum.org/alzbiomarker. Accessed 4 Sept. 2026.

Brockmeyer, Sam, and Amedeo D'Angiulli. "How Air Pollution Alters Brain Development: The Role of Neuroinflammation." Translational Neuroscience, vol. 7, no. 1, Jan. 2016, pp. 24–30, https://doi.org/10.1515/tnsci-2016-0005.

Casey, Emma, et al. "Association Between Fine Particulate Matter Exposure and Cerebrospinal Fluid Biomarkers of Alzheimer's Disease Among a Cognitively Healthy Population-Based Cohort." Environmental Health Perspectives, vol. 132, no. 4, Apr. 2024, https://doi.org/10.1289/ehp13503.

EPA. "Air Quality - National Summary | US EPA." US EPA, 31 July 2018, https://www.epa.gov/air-trends/air-quality-national-summary.

Health Effects Institute. "New State of Global Air Report Finds Air Pollution Is Second Leading Risk Factor for Death Worldwide." Health Effects Institute, 12 June 2024, https://www.healtheffects.org/announcements/new-state-global-air-report-finds-air-pollution-second-leading-risk-factor-death.

Kim, Boram, et al. Alzheimer's Disease Neuropathologic Change Mediates the Relationship Between Ambient Air Pollution and Dementia Severity. July 2024, https://doi.org/10.1101/2024.07.18.24310646.

Lane, Margaret, et al. "The Effects of Air Pollution on Neurological Diseases: A Narrative Review on Causes and Mechanisms." Toxics, vol. 13, no. 3, Mar. 2025, p. 207, https://doi.org/10.3390/toxics13030207.

Marin-Castañeda, Luis A., et al. "Mechanisms of Ozone-Induced Neurotoxicity in the Development and Progression of Dementia: A Brief Review." Frontiers in Aging Neuroscience, vol. 16, Oct. 2024, https://doi.org/10.3389/fnagi.2024.1494356.

Pelkmans, Wiesje, et al. "Astrocyte Biomarkers GFAP and YKL‐40 Mediate Early Alzheimer's Disease Progression." Alzheimer's & Dementia, vol. 20, no. 1, Sept. 2023, pp. 483–93, https://doi.org/10.1002/alz.13450.

Zhang, Xu, et al. "Ozone: Complicated Effects in Central Nervous System Diseases." Medical Gas Research, vol. 15, no. 1, Jan. 2025, pp. 44–57, https://doi.org/10.4103/mgr.MEDGASRES-D-24-00005.

The Final 50: Where Competitive Tiers Clash: Stroke Mechanics By Sophia Chen

Abstract

In this study, I analyze the relationship between athletic performance tier and racing strategies in mid-distance competitive swimming, specifically focusing on the four 200-meter strokes. Using a Proportional Pacing Decay (PPD) as the testing metric, I compiled data from professional swimmers worldwide who competed in international competitions in 2025 and compared them against provincial-tier Canadian athletes. The needed data consisted of 4 tabular datasets, ranging from 8x12, which were, within each dataset, further categorized by gender. This research examines whether elite-tier swimmers showed superior pacing preservation (or lower decay amounts) on the final lap of a race in comparison to provincial-level swimmers. Through evaluating the interactions between stroke mechanics, performance tier, and pulmonary (respiratory) functions, this research paper provides insights into physiological fatigue management and discusses the practical applications of pacing models for high-performance swim coaching.

Keywords

Sports Science, Biomechanics, Proportional Pacing Decay (PPD), Central Governor Model.

Introduction to Competitive Swimming and Performance Tiers

The sport of swimming is regulated by a rigid reality: water is about 800 times denser than air, meaning that drag forces increase quadratically with speed. The 200-meter swimming race, lasting roughly 2 minutes, serves as the intersection between immediate and intense anaerobic power and aerobic endurance. This unforgiving reality pushes athletes into a stage of metabolic fatigue, and as a swimmer moves through the four lengths of a long-course pool, the progressive cumulative muscular burnout can disrupt stroke technique and cause a swimmer's body position to drop in the water, resulting in increased drag. While it is true that elite international competitors possess clear advantages in raw power and speed over provincial athletes (often due to advanced training maturity and advantageous anthropometric traits), a critical question remains: Does being a world-class athlete automatically translate to a swimmer having superior tactical control, especially during times when the body faces the climax of physiological stress? By evaluating both performance tiers, this study investigates whether the physical depletion is universally equal or if elite athletes are uniquely outfitted to resist technical deterioration and deceleration on the final lap.

Introduction to Pacing Models & The Central Governor

While fatigue in the world of swimming has primarily been viewed as a purely mechanical failure of the muscles due to lactic acid accumulation or glycogen depletion, modern

sports science frequently cites the Central Governor Model, an idea proposed by South African sports scientist Tim Noakes and his colleagues in the early 2000s,[1] which advances the idea that the brain subconsciously regulates the intensity of exercise to protect the body from serious bodily damage, resulting in an individual in nearly never hitting their true physical limit.[2,3] According to this model, an athlete's pacing strategy is continuously calculated by the Central Nervous System, which balances the body's sensory feedback with the remaining distance of the race.[4] In a 200-meter swimming race, this manifests in a tactic in three different ways: positive splitting (the first half of a race being faster than the second half), negative splitting (the second half of a race being faster than the first half), and even splitting (the first and second halves of a race being the same speed). Since the first 50 meters of a 200-meter-long course race benefits from a diving start, the second 50-meter lap serves as the optimal baseline for an athlete's mid-race velocity. The final 50 meters, on the other hand, serves as the final test of endurance. By examining the relationship between mid-race and final speed, this paper explores how the protective mechanism of the Central Governor Model operates across different skill tiers, highlighting whether elite-level training allows athletes to bypass neural warning signals more effectively than provincial-level swimmers.

Introduction to the PPD Formulas

To quantify the rate of velocity drop-off during the final length of a 200-meter race, this study utilizes a metric called the Proportional Pacing Decay (PPD). By comparing the data within each swimmer, the formula does not get naturally skewed by the higher raw velocities of elite athletes. The PPD approach ensures that the pacing preservation is compared fairly across both international and provincial performance tiers.

A positive *PPD* indicates a percentage slowdown, or that the pacing has decayed as the 200-meter race went on. Conversely, a negative *PPD* showcases that the athlete negatively-split the back half of the race, or that they went faster on their final length relative to their second length pace.

Introduction to Swim Meets, Rankings, and Sourcing

To construct my dataset, I compiled data from top international aquatic competitions and provincial-level meets, all held during the 2025 swim season (September 2024 - August 2025).[5,6] Top international aquatic competitions included the 2025 World Aquatics Championships, the 2025 Toyota U.S. National Swimming Championships, the FISU World University Games, the 2025 World Aquatics Junior Swimming Championships, the 2025 TYR Pro Series, the 2025 Australian Swimming Trials, and the European Swimming Championships.[7-10] This allowed for the pairing of 4 elite-tier datasets against 4 provincial-tier datasets, which in turn provided a clear view of how performance environment and tier affect pacing in races.

Methodology

To what extent does performance tier (Elite Globally vs. Provincially) impact an athlete's Proportional Pacing Decay (PPD) across the four primary competitive swimming strokes in 200m LCM events, and what does this reveal about the psychological fatigue threshold?

The sample for this study consisted of two cohorts across the four swimming disciplines: an elite international cohort (n = 12 per event) and a provincial cohort (n = 12 per event) who are registered within Canadian provincial databases. Only athletes ranking in the top 25 internationally or within the top 25 provincially in their respective events during the 2025 season were considered for inclusion. Through the filtered sub-population, simple random sampling was used to select n = 12 distinct athletes per cohort for each event (via a random number generator). The independent variable is defined categorically as the performance tier (Elite vs. Provincial), while the dependent variable is the Proportional Pacing Decay (PPD), the percentage score representing the velocity drop-off or the acceleration between an athlete's second and fourth 50-meter laps. By normalizing the data against each swimmer's own second length, the shape of the pacing curve is isolated, which allows the study to disregard the absolute speed differences between the groups and allows for a fair comparison of energy distribution.

All statistical computations were performed using an unpooled, independent, two-sample t-test using a TI-84 calculator. Since independent two-sample t-tests do not assume equal variances between two independent groups, they allowed for a more robust analysis of my data. Since my study hypothesizes that world-class training allows athletes to manage final-lap decay more effectively, a one-tailed test was implemented at a statistical significance (α) = 0.05, where:

Null Hypothesis (H_0): $\mu_{Elite_x} \geq \mu_{Provincial_x}$
(If the null hypothesis is true, the true mean proportional pacing decay of elite swimmers in stroke *x* is greater than or equal to that of Provincial swimmers. Any observed difference in our sample is due entirely to random sampling variation.)

Alternative Hypothesis (H_a): $\mu_{Elite_x} < \mu_{Provincial_x}$
(If the alternative hypothesis is true, the true mean proportional pacing decay of elite swimmers in stroke *x* is significantly less than that of Provincial swimmers, supporting the psychological theory of a higher, more disciplined Psychological Fatigue Threshold in elite athletes.)

Furthermore, my study fulfills the three core statistical conditions, which are all fulfilled:

1. **Randomness**: A random number generator was used to select the 12 swimmers from the filtered top 25 pools, ensuring that every athlete in the sub-strata had an equal probability of selection.
2. **10% Independence Rule**: The sample size of n=12 is less than 10% of the total population of international- and provincial- level swimmers.

3. **Normality Condition**: Considering the small sample size (n=12), the Normality condition was verified by evaluating the distribution characteristics of the split data. These distributions showed neither extreme skewness nor severe outliers, allowing me to justify the Normality of this study.

Results and Discussion

Men's and Women's 200 Meter Freestyle

An independent, two-sample t-test was conducted to analyze the Proportional Pacing Decay (PPD) differences between elite international and provincial-level swimmers in the 200-meter freestyle.

Men's 200 Freestyle

The elite international male cohort, of which n=12, had a sample mean PPD of 1.78% (s_1 = 3.20%), while the male provincial cohort, which is also n=12, had a sample mean PPD of 0.11% (s_2 = 2.35%). In testing the alternative hypothesis (H_a: $\mu_{Elite_x} < \mu_{Provincial_x}$), my calculations yielded a t-statistic of 1.46 and a p-value of 0.920 (df = 20.18). Since the p-value = 0.920 > α = 0.05, the study **fails to reject the null hypothesis** for the male cohort.

Women's 200 Freestyle

The elite international female cohort, of which n=12, had a sample mean PPD of 0.08% (s_1 = 1.62%), while the female provincial cohort, which is also n=12, had a sample mean PPD of 1.32% (s_2 = 3.35%). In testing the alternative hypothesis (H_a: $\mu_{Elite_x} < \mu_{Provincial_x}$), my calculations yielded a t-statistic of -1.16 and a p-value of 0.132 (df = 15.86). Since the p-value = 0.132 > α = 0.05, the study **fails to reject the null hypothesis** for the female cohort as well.

Discussion

The varied statistical patterns across performance tiers in the 200-meter freestyle highlight a unique autonomy in technique: since it is the most drag-minimizing stroke, athletes can manipulate their race strategies. The hydrodynamic efficiency of freestyle allows swimmers to ensure that they do not experience immediate penalties for certain race strategies, such as going out too fast (positive splitting). In the male cohorts, the data imply an unexpected trend: elite global swimmers actually experienced a higher mean pacing decay when compared to the provincial cohort (1.78% to 0.11%), suggesting that in high-stakes racing environments, the preferred strategy is to gain and move through "clean water" earlier on in the race to avoid gravity waves and wake from swimmers in neighboring lanes. As a result, elite, international-level men typically take a higher-risk, faster front-half approach to their race strategy, consciously expending a large portion of their anaerobic glycogen capacity during the first 100 meters of their race and later succumbing to a higher rate of fatigue-induced drop-off,

while provincial-level men, lacking the same type of field pressure, tended to conserve energy earlier on to maintain a nearly even-split pacing strategy, as per the 0.11% mean pacing decay.

Meanwhile, the women's data trended in the hypothesized direction, with elite global women showing an average decay of 0.08% compared with provincial-level women, who had a mean pacing decay of 1.33%. Although there is no statistical significance (p-value = 0.132), this highlights the ability of world-class female athletes to even out their splits, suggesting that they prioritize their distance per stroke (DPS) and can maintain their body position throughout most of the race. By maintaining high stroke efficiency rather than relying on an unsustainable stroke rate, elite female swimmers can preserve their glycolytic reserves and delay the accumulation of blood lactate and hydrogen ions, hence allowing them to have complete control over the Central Governor's inhibitory fatigue commands, which, in turn allowing them to match their mid-race velocities on the final lap with the second lap and resist the deceleration that is observed more obviously in the provincial tier.[11,12]

Ultimately, the 200-meter freestyle race indisputably highlights how different genders correlate to different tactics, and how pacing styles and performance tiers interplay with one another when drag resistance is minimized. In this part of the study, it is evident that while elite men prioritize early race dominance over final-lap energy conservation (which results in a larger pacing drop-off than regional swimmers), elite women do quite the opposite – they rely more on technique and mechanics, working towards even splitting. Ultimately, since neither cohort reached the threshold for statistical significance ($\alpha = 0.05$), the freestyle stroke establishes itself as a stroke where the competitive tier, alone, cannot dictate a universal pacing strategy, but rather provides scope for diverse race strategies and technical execution across different skill levels.

Men's and Women's 200-meter Butterfly

An independent, two-sample t-test was conducted to analyze the Proportional Pacing Decay (PPD) differences between elite international and provincial-level swimmers in the 200-meter butterfly.

Men's 200 Butterfly

The elite international male cohort, of which n=12, had a sample mean PPD of 5.58% (s_1 = 2.45%), while the male provincial cohort, which is also n=12, had a sample mean PPD of 5.74% (s_2 = 2.65%). In testing the alternative hypothesis (H_a: $\mu_{Elite_x} < \mu_{Provincial_x}$), my calculations yielded a t-statistic of -0.16 and a p-value of 0.438 (df = 21.87). Since the p-value = 0.438 > α = 0.05, the study **fails to reject the null hypothesis** for the male cohort.

Women's 200 Butterfly

The elite international female cohort, of which n=12, had a sample mean PPD of 2.96% (s_1 = 1.92%), while the female provincial cohort, which is also n=12, had a sample mean PPD of 5.45% (s_2 = 3.86%). In testing the alternative hypothesis (H_a: $\mu_{Elite_x} < \mu_{Provincial_x}$), my calculations

yielded a t-statistic of -2.00 and a p-value of 0.031 (df = 16.13). Since the p-value = 0.031 > α = 0.05, the study **rejects the null hypothesis** and favors the alternative hypothesis. This means that there is statistically significant evidence to conclude that elite global female swimmers maintain a lower rate of final-lap pacing drop-off than provincial-level swimmers.

Discussion

The 200-meter butterfly event is widely regarded as one of the most physically demanding events in competitive swimming, considering the continuous upper-body pulling motions intertwined with synchronous fly kicks, resulting in a stroke that requires a continuously stable core. Any loss in power or core stability will instantly change the athlete's body position in the water, resulting in a body position that is closer to being vertical, which leads to an instantaneous increase in drag. For the male cohort, the statistics are similar: provincial swimmers had an average PPD of 5.74%, while elite swimmers had an average PPD of 5.58%. This suggests that for an event like the butterfly, there is a universal barrier: since the butterfly requires a simultaneous recovery of both arms out of the water, as metabolic fatigue and hydrogen ion (H+) accumulation peak on the final length of the race, the muscular power required to lift the upper torso clear of the surface declines. The data indicate that this physical limitation impacts both male groups to a similar extent; while elite men generate higher absolute mid-race speeds, their final lap deceleration is relatively similar to that of provincial swimmers due to the muscular requirements of the stroke. On the other hand, the women's data demonstrated a clearer statistically significant split between the elite women's tier, who had a final lap decay of 2.96%, and the provincial tier, who had a 5.45% drop-off. This statistically significant difference can be explained by two major factors: endurance in upper-body power and respiratory efficiency. In the butterfly stroke, it is important to emphasize sustaining a flat, forward-driven breath, as it allows for the head to stay low, which further protects a swimmer's horizontal alignment. When provincial female swimmers face intense fatigue, their stroke rate slows down, and their hips drop, which creates a domino effect that forces their heads higher out of the water to breathe. As a result of the lifting action, the lower body drops, causing resistance to skyrocket. Elite women, in contrast, utilize their superior core stabilization to maintain their distance-per-stroke (DPS). This high efficiency delays the metabolic distress signals that are sent to the central nervous system, and satisfies the brain's subconscious pacing calculations under the Central Governor Model, which prevents the nervous system from triggering waves of inhibitory fatigue commands.13 In addition, at the international level, elite female athletes likely possess much more racing experience in the 200 butterfly, giving them an advantage in knowing how to manage their energy across all four laps; meanwhile, provincial-level swimmers have a larger chance of misjudging their energy output, resulting in the ultimate drop-off, as shown in the data.

In conclusion, the 200-meter butterfly underscores how the underlying physiological demands of a stroke yield completely different pacing outcomes between genders. For men, the absolute anaerobic power that is expended early in the race triggers a substantial final-lap

deceleration that even top-tier training cannot circumvent. This universal slowdown suggests that for male athletes across all tiers, coaching should focus on improving absolute aerobic capacity and metabolic moderating to help sustain the intensive upper-body torque on the final length. For women, however, the gap between elite and provincial performance is visible, suggesting that world-class female flyers can rely more heavily on technical precision and stroke efficiency to hold their body position together under stress, and can overcome the fatigue that causes provincial-level competitors to experience a much more severe drop-off in splits. As a result, this suggests that putting a heavier emphasis on technical swimming in butterfly could have a more sustained impact on a provincial-level swimmer's 200 butterfly.

Men's and Women's 200-meter Breaststroke

An independent, two-sample t-test was conducted to analyze the Proportional Pacing Decay (PPD) differences between elite international and provincial-level swimmers in the 200-meter breaststroke.

Men's 200 Breaststroke

The elite international male cohort, of which n=12, had a sample mean PPD of 3.81% (s_1 = 2.62%), while the male provincial cohort, which is also n=12, had a sample mean PPD of 2.94% (s_2 = 2.47%). In testing the alternative hypothesis (H_a: $\mu_{Elite_x} < \mu_{Provincial_x}$), my calculations yielded a t-statistic of 0.84 and a p-value of 0.795 (df = 21.92). Since the p-value = 0.795 > α = 0.05, the study **fails to reject the null hypothesis** for the male cohort.

Women's 200 Breaststroke

The elite international female cohort, of which n=12, had a sample mean PPD of 3.39% (s_1 = 1.91%), while the female provincial cohort, which is also n=12, had a sample mean PPD of 3.11% (s_2 = 3.22%). In testing the alternative hypothesis (H_a: $\mu_{Elite_x} < \mu_{Provincial_x}$), my calculations yielded a t-statistic of 0.26 and a p-value of 0.601 (df = 17.87). Since the p-value = 0.601 > α = 0.05, the study **fails to reject the null hypothesis** for the female cohort.

Discussion

The statistical analysis of the 200-meter breaststroke shows a structural pattern: the pacing decay scores, for both genders, are considerably tight and consistently elevated, ranging approximately between 2.9% and 3.8% across all four groups. This lack of statistical differentiation indicates that the physical constraints of breaststroke form a universal equalizer that binds both elite and provincial swimmers to similar fatigue levels.

The distinct stroke of breaststroke is the most hydrodynamically inefficient and technically complex in competitive swimming. The body goes through an acceleration during the kick, followed by a deceleration period during the gliding phase, where the knees drop, hence creating a huge drag. To minimize this force, an athlete must be able to utilize a tight and timed

gliding phase. However, similar to the other strokes, as metabolic acidosis and hydrogen ion (H+) accumulation overwhelm the body in the final 50 meters, the timing that the glide requires begins to decay.

The data indicate that when this wall is hit, the Central Governor's inhibitory fatigue commands affect all swimmers, regardless of their tier.[4] While elite swimmers have superior absolute speed, their drop-off parallels that of provincial-level swimmers. This occurs as a result of the breakdown of technique, which is a skill that just conditioning cannot surmount. Furthermore, the slim difference between tiers highlights how the discrepancy in competitive experience or high-level conditioning and training cannot skirt around the technical laws that breaststroke requires. Unlike freestyle or backstroke, where an athlete can adjust their pacing style or avoid fatigue by hiding behind a slower stroke rate, breaststroke is primarily technical. For instance, if either a provincial-level swimmer or a World Championship finalist rushes their stroke on the final lap due to exhaustion, their hips will drop, stalling the gliding phase, which finally causes them to slow down.

So, ultimately, the 200-meter breaststroke event underlies an environment where performance tiers do not separate pacing decay. For both men and women, the foremost way of improving is not so much trying to out-power the water through strength and conditioning; rather, coaches and athletes should emphasize technique that will allow them to mitigate the resistance that comes with swimming breaststroke.

Men's and Women's 200-meter Backstroke

An independent, two-sample t-test was conducted to analyze the Proportional Pacing Decay (PPD) differences between elite international and provincial-level swimmers in the 200-meter backstroke.

Men's 200 Backstroke

The elite international male cohort, of which n=12, had a sample mean PPD of -0.20% (s_1 = 2.04%), while the male provincial cohort, which is also n=12, had a sample mean PPD of -0.72% (s_2 = 2.45%). In testing the alternative hypothesis (H_a: $\mu_{Elite_x} < \mu_{Provincial_x}$), my calculations yielded a t-statistic of 0.56 and a p-value of 0.710 (df = 21.30). Since the p-value = 0.710 > α = 0.05, the study **fails to reject the null hypothesis** for the male cohort.

Women's 200 Backstroke

The elite international female cohort, of which n=12, had a sample mean PPD of 0.64% (s_1 = 2.04%), while the female provincial cohort, which is also n=12, had a sample mean PPD of -0.15% (s_2 = 1.96%). In testing the alternative hypothesis (H_a: $\mu_{Elite_x} < \mu_{Provincial_x}$), my calculations yielded a t-statistic of 0.97 and a p-value of 0.828 (df = 21.97). Since the p-value = 0.828 > α = 0.05, the study **fails to reject the null hypothesis** for the female cohort.

Discussion

The findings of the 200-meter backstroke reveal an interesting trend across both performance tiers: both levels showed a near-zero or negative average pacing decay, meaning that swimmers either evenly-split their race or negative-split their races. This universal ability to resist late-race deceleration highlights the distinct anatomical and sensory advantages of backstroke, as unlike the other three disciplines, the backstroke stroke allows the swimmer to be continuously exposed to the air. This eliminates the respiratory restrictions that impede the other three strokes, and by maintaining oxygen intake, swimmers are able to delay fatigue levels and better regulate their blood pH, which essentially helps suppress early inhibitory neural signals from the Central Governor Model.

Firstly, for both male cohorts, both groups showed a remarkable consistency: elite swimmers increased their speed on their last laps, resulting in a PPD of -0.20%; provincial swimmers did the same, with a PPD of -0.72%. This negative-splitting pattern points towards a shared characteristic of backstroke: since backstroke is a stroke where athletes must continuously gaze upwards, they lack immediate visual feedback regarding their position in a pool or proximity to their competitors. This creates a "blind" effect, forcing swimmers to rely on their individual metronomic paces. Through lacking the disruptive visual triggers that affect other strokes, swimmers are less likely to over-exert early on or be impacted by those around them. As a result, both elite and provincial men can conserve their anaerobic capacities across the middle 100 meters, ultimately allowing them to drop their final lap times without experiencing extreme technical decline.

Similarly, findings for the female cohorts also showed minimal decay. The elite tier had a PPD of 0.64%, and the provincial women had a negative drop-off of -0.15%. Since p-value = $0.828 > \alpha = 0.05$, there is a lack of statistical significance, reinforcing the idea that the design of the stroke naturally allows swimmers to preserve their stroke lengths and body positions. In backstroke, the arms rotate alternately, similar to freestyle, avoiding the resistance spikes that butterfly and breaststroke encompass. Furthermore, the continuous, low-drag nature of the stroke, mixed with the continuous stream of oxygen, nullifies how different competitive tiers are a differentiating factor.

As a result, the 200-meter backstroke event demonstrates how performance tiers do not differentiate race strategies since both tiers deliver similar, almost even-split races. The combination of continuous oxygen intake and lack of forward vision likely forces athletes to follow their own racing strategies rather than getting carried away by the competitor in the neighboring lane. In a more practical sense, these results indicate that while training additional meters or yards and continuous conditioning is advantageous, it could also be beneficial for coaches, across both tiers, to focus on refining the lack of spatial awareness – for instance, optimizing underwater dolphin kicks and stroke count per 50 – to optimize an athlete's absolute velocity within this pacing model, which has already been proven to be efficient.

Conclusion

In this communication, the findings demonstrate that final-lap pacing decay (PPD) in the 200-meter distance is not determined by an athlete's competitive level alone, but rather by how that level interacts with the physical, respiratory, and psychological demands of each stroke. Through the Central Governor Model, the brain's subconscious fatigue commands depend not only on physical exertion but also on environmental cues, such as being ahead of the field in an international final, which would boost physical output, or pressure and negative thoughts, which can negatively impact performance and impede one's total ability to perform.[14] In strokes such as the 200-meter breaststroke for both women and men or the 200-meter butterfly for men, physical constraints are universal, drawing similar fatigue curves across the different tiers. In contrast, low-resistance strokes such as freestyle and backstroke give athletes the freedom to adjust pacing without immediate physical impact. However, a strict limitation of this study is that, since the data were exclusively gathered from the top-25 rankings within each tier, it would not be accurate to generalize this to a wider swimming population. In short, heavy water resistance forces a similar drop-off across both groups, whereas low-drag strokes offer more tactical choice. This leaves the female 200-meter butterfly as the only event where elite status provided a clear statistical advantage ($p = 0.031$) against final-lap fatigue.

AI Acknowledgement

During the preparation of this work, the author used Google Gemini to refine sentence clarity and format mathematical equations. The underlying data collection, statistical computations (via TI-84 calculation), and thematic analysis were independently executed by the author. After using this tool, the author reviewed and edited the content as needed and took full responsibility for the content of the publication.

Acknowledgements

None.

Author

Sophia Chen is a rising high school senior and competitive swimmer who plans to pursue a double major in Statistics and Creative Writing. Driven by an interest in sports analytics, she enjoys combining quantitative research with clear storytelling to explore human performance and biomechanics in the water.

Works Cited

1.Central governor and fatigue – neuromechanics of human movement: For learners by learners. Central Governor and Fatigue. (n.d.). https://pressbooks.calstate.edu/humanneuromechanics/chapter/central-governor-and-fatigue/

2.The mind is greater than the muscles: What the central governor theory – bloc coaching & race team. (2025, October 8). https://www.bloc-raceteam.com/blogs/news/the-mind-is-greater-than-the-muscles-what-the-central-governor-theory-teaches-us-about-ironman-performance

3.Inzlicht, M., & Marcora, S. M. (2016). The Central Governor Model of Exercise Regulation Teaches Us Precious Little about the Nature of Mental Fatigue and Self-Control Failure. Frontiers in Psychology, 7, 656. https://doi.org/10.3389/fpsyg.2016.00656

4.St Clair Gibson, A., & Noakes, T. D. (2004). Evidence for complex system integration and dynamic neural regulation of skeletal muscle recruitment during exercise in humans. British journal of sports medicine, 38(6), 797–806. https://doi.org/10.1136/bjsm.2003.009852

5.Swimcloud. (n.d.). https://www.swimcloud.com/

6.Swimrankings - swimming rankings and results - worldwide, International Swim Sport. Swimrankings. (n.d.). http://swimrankings.net

7. Omegatiming. (n.d.-a). https://www.omegatiming.com/File/00011900010204EE0104FFFFFFFFFF01.pdf

8.FISU World University Games 2025 results - schedule. (n.d.-a). https://rhineruhr2025.com/en/schedule

9.Meet results: 2025 Australian Swimming Trials. 2025 Australian Swimming Trials. (n.d.). https://liveresults.swimming.org.au/sal/2025Trials/

10. World Aquatics Junior Swimming Championships 2025 - Romania | World Aquatics Official. (n.d.-d). https://www.worldaquatics.com/competitions/4759/world-aquatics-junior-swimming-championships-2025

11. Psychology and the central governor model with Professor Tim Noakes | EP#43 - YouTube. (2017, July 23). https://www.youtube.com/watch?v=PmJA17IKhJ0

12. Noakes, T. D., St Clair Gibson, A., & Lambert, E. V. (2005). From catastrophe to complexity: a novel model of integrative central neural regulation of effort and fatigue during exercise in humans: summary and conclusions. British journal of sports medicine, 39(2), 120–124. https://doi.org/10.1136/bjsm.2003.010330

13. Noakes T. D. (2012). Fatigue is a Brain-Derived Emotion that Regulates the Exercise Behavior to Ensure the Protection of Whole Body Homeostasis. Frontiers in Physiology, 3, 82. https://doi.org/10.3389/fphys.2012.00082

14. Marcora, S. M., & Staiano, W. (2010). The limit to exercise tolerance in humans: mind over muscle?. European journal of applied physiology, 109(4), 763–770. https://doi.org/10.1007/s00421-010-1418-6

Frozen in Time: A Modern Look at Icing the Kicker By Rohan Thakare

Abstract

Icing the kicker, the strategy of calling a timeout immediately before a high-stakes field goal attempt, has become a well-known strategy used throughout the NFL. While research ruled in favor of icing in its early years of use, modern research has conveyed a change in effectiveness. This paper examines whether icing the kicker remains an effective strategy in the modern NFL by analyzing evidence across three key areas: historical trends in icing effectiveness, environmental factors affecting field goal outcomes, and individual kicker ability. Across these three points, research and evidence have shown the outcome of icing favoring the kicker rather than the defense. Firstly, kickers have adapted to the strategy, and coaches implement this last second freeze into practices, allowing kickers to mentally prepare and neutralize the tactic entirely. However, a deciding factor that should be accounted for more than icing is the environment. Environmental conditions such as wind and precipitation introduce variables that no timeout can replicate; additionally, the individual skill level of modern NFL kickers has risen to the point where a brief stoppage does little to alter the result. Taken together, these findings suggest that icing the kicker is not the reliable disruptive force coaches have long believed it to be. In a sport where every timeout carries real strategic weight, burning one on a tactic that the evidence does not consistently support is a decision that deserves far more scrutiny than it typically receives on the sideline.

Introduction

In American football ("football"), the art of kicking is vital. Game-winning field goals (three points per successful kick) and extra-points (one point per successful kick) not only decide the outcome of a single game, but can have implications for a team's entire season, from missing the playoffs to being in contention for the sport’s ultimate prize in its biggest stage, a trip to the Super Bowl for the Lombardi trophy. Most recently, the stakes of a single kick were on full display in the final game of the 2025 National Football League (“NFL”) season when Baltimore Ravens’ rookie kicker Tyler Loop pushed a 44-yard attempt wide right as time expired in a winner-take-all Week 18 matchup, sending the Pittsburgh Steelers to the playoffs with a 26-24 victory and eliminating Baltimore from postseason contention. Essentially, one missed kick ended the Ravens' season on the spot (11).

Scenarios like these have led coaches to find any competitive advantage they can utilize, especially when nearly a quarter of NFL games in the last twenty-five seasons (2000 - 2025) are decided by three points or less (5). One such strategy that has become a staple of late-game coaching decisions is icing the kicker. "Icing the kicker" is when a defensive team calls a timeout immediately before a game-winning or game-tying field goal attempt in the closing seconds of regulation or overtime. The goal is to disrupt the kicker's rhythm and focus during high-pressure moments, such as a tie game or one-score deficit. Although icing the kicker may influence some attempts under particular conditions, the available evidence does not show that it is a consistently

reliable strategy. Distance, weather, venue, experience, and individual kicker ability appear to be more dependable predictors of field-goal success..

History of Icing

The history of icing the kicker originated in the late 1970s when coaches began looking for competitive advantages in high-pressure situations late in games. Coaches observed that a timeout could influence some kickers to tense and overthink the attempt, which gradually made the tactic more appealing and widespread: coaches observed that a timeout could influence some kickers to tense and overthink the attempt, which gradually made the tactic more appealing and widespread. Ryan Dahnke, a sports data analyst whose research focuses on NFL game strategy, notes in his study "Breaking the Ice: Investigating how Timeouts Shape Game-winning Kicks in the NFL" that "since 2000, icing frequency has held steady at around 36% of ice-able attempts, albeit with some variation year-over-year (4).”

Rise and Fall of Icing

Early research suggested that icing the kicker might reduce success in certain situations. When coaches first consistently iced the kicker, it proved an effective strategy. Scott Berry and Craig Wood, statisticians writing for Chance, a peer-reviewed journal published by the American Statistical Association, conducted one of the earliest data-driven analyses of the practice. In their 2004 study, Berry and Wood defined "pressure kicks" as attempts made within three minutes or less in regulation or overtime that would result in the game being tied or the kicking team taking the lead, and found that icing the kicker was associated with the percentage of successful attempts in the 40-55 yard range to drop by about ten percent [data obtained from 1991-2002] (2). This implies that a simple change such as a timeout called before a game winning kick attempt was able to greatly affect the outcome, causing many coaches to take notice and adapt the practice. As a result, icing became increasingly common, and what was once a rare situational tactic evolved into a widespread strategy deployed in close games in the NFL.

However, as icing became so pervasive, so did preparation against it. Coaching personnel and kickers began simulating icing-scenarios in practice. This led to kickers developing mental strategies and warm-up routines to stay composed under pressure, leading to kickers overcoming the monumental pressure that icing adds. Per a 2011 Dan Graziano (ESPN Reporter) article, ESPN’s Stats & Information department found that over a decade period from 2001-2011, kickers made 83.0% of attempts when iced compared to 70.2% when they were not iced in the final in the final seconds of the fourth quarter (6). One possible explanation is that teams began incorporating timeout interruptions and pressure situations into practice. This led to kickers developing mental strategies and warm-up routines to stay composed under pressure, leading to kickers overcoming the monumental pressure that icing adds.

An ESPN statistics study of NFL games found a complete change in the data compared to that a decade prior, stating “some modern analyses even suggest that icing may benefit kickers.” Additionally, a May 2025 analysis by the Sports Analytics Group at Berkeley, covering 25 years

of NFL data through the 2024 season, found no statistically significant icing effect after controlling for distance and kicker experience, confirming that the ten percent drop icing once produced has effectively vanished (3). These findings suggest that icing is not a statistically reliable general strategy once other factors affecting field-goal success are considered.

Environmental Contribution

Furthermore,environmental factors also have an influence on the outcome of a field goal attempt. For example, wind is an extremely volatile natural element that can benefit kickers or put them in precarious situations. Wind alone has multiple implications that determine a kick's outcome such as direction, speed, and chill. To demonstrate the impact of wind, consider a kicker attempting a sixty-yard field goal with a strong wind at his back along with the added force that can provide the extra distance needed to make the kick. However, if the wind shifts even slightly off-target, the ball can be pushed off its intended path, so the kicker must change his alignment to compensate. A wind gust blowing directly toward the kicker will slow the ball as it travels against it, effectively making a sixty-yard attempt feel more like a seventy-yard attempt. Depending on the speed of the wind, these effects can be subdued or amplified, making it one of the most unpredictable variables a kicker faces. Noted and graphed by Best Ball Stats, "Windy conditions make kicking field goals much more difficult than in ideal conditions. Kickers have to account for the direction and strength of the wind in order to be sure the wind does not take the ball on an unintended path. However, different wind speeds have various effects on the difficulty of field goal attempts (1) ." As illustrated in Figure 1, field goal conversion rates decline and grow increasingly unpredictable as wind speed rises, with the most dramatic drop in consistency occurring at speeds between 20 and 30 MPH.

Figure 1: Effect of Wind Speed on Field Goals (1)

In conclusion, a timeout does nothing to change the direction of the wind or reduce its speed, it simply pauses the game while the same conditions remain. Therefore, icing does not reduce the physical difficulty created by unfavorable weather.

Precipitation similarly presents a significant challenge that no timeout can replicate or counteract. Best Ball Stats, a sports analytics publication, notes in their article, "In addition to

wind speed, the second important weather factor for outdoor kicks is precipitation. An accumulation of rain or snow during a game can make it difficult for the holder to set the kick and for kickers to kick the ball without slipping (1)." As shown in Figure 2, field goal conversion rates drop noticeably in the presence of precipitation, falling from roughly 84% in dry conditions to below 80% when precipitation is a factor.

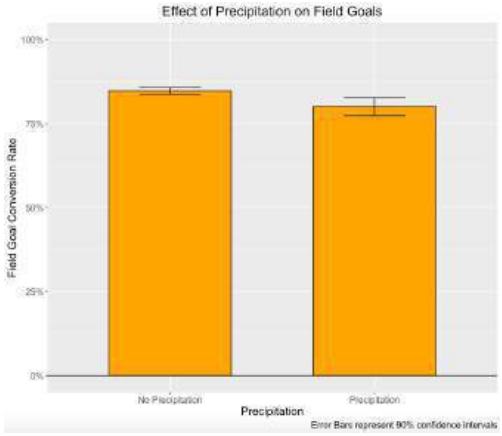


Figure 2: Effect of Precipitation of Field Goals (1)

Beyond its direct effect on kicking, precipitation reshapes the entire game. Snow and rain force coaches to abandon pass-heavy schemes, slow the pace of play, and alter every strategic decision that leads to that moment, making it the most disruptive variable on the field. Although teams can practice in difficult conditions, precipitation introduces physical uncertainty that a timeout cannot eliminate.These are variables entirely outside of anyone's control, and they have a direct and measurable impact on the outcome of a kick. A coach who chooses to burn a timeout on icing in the middle of a downpour is essentially ignoring the most important factor in the game at that moment. Using a timeout to ice a kicker in severe weather may therefore provide little additional disruption beyond the conditions already affecting the attempt.

Player Profile

A notable predictor of a field goal's outcome is the kickers'profile themselves. Career accuracy, distance range, and individual consistency are what ultimately determine whether a kick goes through the uprights, not a timeout called by the opposing coach. NFL kickers have become measurably more accurate at every distance since 1999, with kickers in 2024 performing twenty percentage points better on kicks from 50-54 yards and forty percentage points better on kicks from 55-59 yards compared to kickers from 1999-2003(8).

The league average field goal percentage now sits above eighty-eight percent , and the NFL has made clear it expects kickers to meet that standard. Former Atlanta Falcons head coach Mike Smith put it plainly: "If you can't make 80 to 85 percent, you're probably not going to be a kicker in the NFL (10)." The careers of Justin Tucker, who converted over 90% of his career

field goal attempts, and Harrison Butker, who posted a historic 94.3% accuracy rate in 2023, illustrate what clearing that bar looks like (7).

The NFL has equally little patience for those who fall short: Dan Bailey, long-stead kicker for the Dallas Cowboys and at one point the second-most accurate kicker in NFL history, was released by the Minnesota Vikings after his accuracy dropped to a career-low 68% in 2020 (9); and Robbie Gould was cut by the Chicago Bears despite being the franchise's all-time leader in field goals and points (10). The reality is simple: if a kicker is not accurate enough to keep his job, a timeout is not what costs him the kick. These examples demonstrate the importance of long-term accuracy and consistency. Although a timeout may affect a particular attempt, individual ability remains a more stable predictor of a kicker’s overall performance.

Conclusion

Icing the kicker is a strategy rooted in an era when it could actually make a difference. Modern studies present a more complicated picture. The data tells a different story today. Kickers have adapted to it through rigorous mental and physical preparation, rendering the psychological disruption it once caused largely obsolete. Environmental factors such as wind and precipitation exert a far greater influence on the outcome of a kick than any timeout could, and the individual skill level of modern NFL kickers has risen to a point where a brief stoppage does little to alter the result. Coaches who burn a timeout on icing are not gaining an edge; they are sacrificing a valuable resource for a tactic that the game has moved past. In a sport where the margin for victory is razor thin and a single kick can send a team home, strategic decisions must be grounded in what actually affects outcomes. Icing the kicker does not.

Works Cited

Batra, Ayush. “Finding the NFL’s Best Kickers.” Best Ball Stats, 23 Feb. 2021, https://bestballstats.com/2021/02/23/finding-the-nfls-best-kickers/. Accessed 3 Aug. 2026.

Berry, Scott M., and Craig Wood. “A Statistician Reads the Sports Pages: The Cold-Foot Effect.” CHANCE, vol. 17, no. 4, 2004, pp. 47-51. https://doi.org/10.1080/09332480.2004.10554926.

Bukovec, Zack. “Does Icing the Kicker Even Work?” Sports Analytics Group atBerkeley, 27 May 2025, https://sportsanalytics.studentorg.berkeley.edu/articles/icing-the-kicker.html. Accessed 3 Aug. 2026.

Dahnke, Ryan. “Breaking the Ice: Investigating How Timeouts Shape Game-Winning Kicks in the NFL.” Ryan Dahnke, Apr. 2025, https://ryandahnke.com/icing-the-kicker/. Accessed 3 Aug. 2026.

“Frequency of NFL Regular-Season Scores, 2000-2025.” Cleanup Hitter, https://cleanuphitter.com/nfl/stats/nfl_common_scores.php. Accessed 3 Aug. 2026.

Graziano, Dan. “Does Icing the Kicker Really Work?” ESPN, 12 Dec. 2011, https://www.espn.com/blog/nflnation/post/_/id/49988/does-icing-the-kicker-really-work. Accessed 3 Aug. 2026.

Loring, Scott. “Stats Don’t Lie: Why Harrison Butker Is the NFL’s Best Kicker.” Arrowhead Addict, 30 Mar. 2024, https://arrowheadaddict.com/posts/stats-dont-lie-why-harrison-butker-is-now-nfl-best-kicker-01hscpr8bw8g. Accessed 3 Aug. 2026.

McLaughlin, Conor. “Visualizing NFL Kicker Accuracy Trends (1999-2024).” Conor McLaughlin, 5 Jan. 2025, https://conormclaughlin.net/2025/01/visualizing-nfl-kicker-accuracy-trends-1999-2024/. Accessed 3 Aug. 2026.

“Minnesota Vikings Release Kicker Dan Bailey.” CBS News Minnesota, 9 Mar. 2021, https://www.cbsnews.com/minnesota/news/minnesota-vikings-release-kicker-dan-bailey/. Accessed 3 Aug. 2026.

“NFL Kickers Up Up and Away.” Pro Football Hall of Fame, 18 Oct. 2018, https://www.profootballhof.com/news/nfl-kickers-up-up-and-away. Accessed 3 Aug. 2026.

Simmons, Myles. “Sunday Night Football: Steelers Clinch AFC North, Final Playoff Spot with Wild 26-24 Victory.” NBC Sports, 4 Jan. 2026, https://www.nbcsports.com/nfl/profootballtalk/rumor-mill/news/sunday-night-football-steelers-clinch-afc-north-final-playoff-spot-with-wild-26-24-victory. Accessed 3 Aug. 2026.

The New Wave Palestinian Music and Palestinian Identity in the Midst of Occupation and Displacement By Yusra Khalil

Abstract

Since 1948, the Palestinian people have been facing decades of occupation, displacement, and the erasure of their identities. In the midst of this there has been an emergence of music – whether by Palestinian artists or others – engaging with themes of Palestinian identity amid its ongoing erasure. Existing studies have examined the role of music in cultural and diasporic identity formation in general terms. However, few studies actively bring together lyrical, visual, and testimonial evidence on the new wave of Palestinian music together within a single frame. In this paper, I ask: what role does the new wave of Palestinian music play in the preservation of the Palestinian identity in the midst of occupation and displacement? To answer this, I analyse four songs, looking at lyrics, music videos, and where relevant, artist interviews, that tackle recurring themes such as homeland, memory, resistance, exile, return, belonging, language, and cultural continuity. This study argues that music representing the Palestinian identity sustains shared national identity in the midst of erasure by preserving collective memory and cultural narratives for generations of Palestinians and activists. This work's significance relies on the analysis of contemporary Palestinian music's significance; including its rise in popularity and impact in recent years. This is particularly relevant at the moment given the new international visibility of Palestinian music alongside the simultaneous rise in censorship and platform removal of Palestinian cultural content; a contradiction that reflects how precarious and contested visibility actually is. By analysing this newer generation of Palestinian artists, this study explores how an existing tradition of musical resistance is being changed and amplified through this new wave of Palestinian artists.

Keywords collective identity, collective memory, cultural identity, diaspora, resistance, erasure, ethnomusicology, homeland, cultural continuity.

Introduction

> "I hate that I'm more comfortable in English
> But every day I tell myself that it's a privilege
> 'Cause I can talk about the way my people living
> And I'll do it on this beat 'cause I know you'll wanna listen"
>
> – On This Land by Saint Levant (ft. Sol Band)

From the moment Palestinian artist Saint Levant first made his appearance on Coachella, his career rapidly gained recognition. His performance remained on Coachella's YouTube channel; a concoction of Palestinian flags, Palestinian beats and lyrics about liberty, anti-censorship, and the personal stories behind the conflict. After a few weeks, it was wiped out

and nowhere to be found. Only a few clips remained, filmed from the crowd's perspective. No official recording remained publicly accessible. Despite only being a performance, this erasure is a clear simplified representation of what has been happening for years. While the removal of this performance cannot be equated to the historical experiences Palestinians have been facing, it reflects the fragility of cultural visibility in modern media and starts a conversation revolving around the representation and erasure of it. As part of this greater censorship wave, there has been a rise in efforts to suppress pro-Palestinian advocacy in recent years, including through arrests, detentions, and deportation proceedings. A clear example of this is the arrest and attempted deportation of Columbia student activist Mahmoud Khalil (Chater). This recent wave of censorship is not something new; since 1948, the people of Palestine have been facing decades of occupation, displacement, and censorship progressively getting worse war by war. In order to understand the scale of this erasure, it is useful to consider it alongside other forms of settler-colonial control. In *What Israel Fears from Palestine?*, Raja Shehadeh effectively compares many aspects of Israeli policies towards Palestinians to South African apartheid. He argues that both systems make use of segregation, dispossession, and the restriction of fundamental rights while also attempting to normalise these conditions in everyday life (Shehadeh). This comparison brings to light the idea that the struggle is not only political but also cultural; the erasure of identity usually results in the erasure of land and community. Faced with this, cultural preservation becomes extremely important as it could be one of the few remaining means that maintain collective memory, history, and national identity across generations despite continued occupation and displacement. Through lyrics, visual imagery, and the effective circulation of songs across borders, music allows Palestinians (both located nationally and internationally) to preserve shared memories, express experiences of occupation, and confirm the collective sense of belonging. Therefore, this study explores the role of music in preserving the Palestinian identity during occupation and displacement. To place this argument within existing scholarship, it is useful to first establish the theoretical and empirical foundations of the study.

Existing literature has created theories from identity and colonial power generally; some others examined the role music plays in diasporic identity creation. Stuart Hall's theory on cultural identity as always in progress and Edward Said's work on representation and imperialist narratives in *Culture and Imperialism* create a greater understanding of identity and colonial power. Aimé Césaire's work on colonialism and decolonisation further reflects the way reclaiming cultural voice directly challenges representations set by the colonising powers. Within this framework, theorists such as Lidskog and Hesmondhalgh have shown how music helps diasporic communities sustain memory and express their identities across generations. However, very few works bring in lyrical, visual, and testimonial evidence on Palestinian music together in a single frame; particularly during the recent period of intensified erasure and international attention. Even fewer reflect on this new wave of Palestinian artists who approach the musical resistance differently from previous generations. This study addresses that gap by drawing on Hall's theory of identity to show how Palestinian music actively preserves identity under

occupation and displacement. Specifically, I argue that this new wave of Palestinian music both preserves and reworks Palestinian identity through four main mechanisms: an emotional relationship to the homeland, a reminder of memory in recurring objects and images, heritage being portrayed as something that is continuously remade instead of being passed down intact, and resistance expressed through the refusal to let Palestinian self-representation be faded. For the empirical analysis, I will be analysing four songs' lyrics, visual imagery (music videos), and artist interviews (when relevant) for contextual understanding. I will be looking at recurring themes relating to homeland, memory, resistance, exile, return, belonging, language, and cultural continuity. Across these three forms of understanding I will show how they interact together to actively construct and preserve shared Palestinian identity. The findings will then be interpreted through the theoretical frameworks presented in the literature review to examine how music functions as a tool for identity preservation under conditions of occupation and displacement.

Having laid out the study's central argument and its place in the literature, the rest of the paper continues as follows. This paper is organised in five main sections. The first section reviews existing literature on Palestinian identity and the role of music plays in its preservation and identifies the gap this paper addresses. The second section provides the historical context going back to the political history of the Palestinian struggle since 1948. The third section presents the theoretical framework by drawing on Hall's theory of cultural identity, with Said and Césaire as secondary frameworks that focus on the analysis of narrative and decolonisation. The fourth section is the methodology which introduces the logic behind the selection of songs analysed. The findings section then discusses by theoretical ideas the dataset and explores recurring lyrical, visual, and testimonies that play a part in the preservation and evolution of Palestinian identity.

Literature Review: Music, Identity, and the Gap in Palestinian Studies

Within Palestinian studies, many have explored and documented the historical and political aspect of the struggle, however not many have done so specifically through music. Khalidi argued that the Palestinian identity itself "emerged in response to many stimuli"; through historical, political, and social changes (Khalidi 36). He further argued that neighbouring Arab countries, like Syria and Lebanon, were also beginning to think of themselves as distinct nations during the late 19th and early 20th centuries. Therefore, Palestinian identity was not simply a reaction to Zionism but part of a broader progression of national identity formation across the region (Khalidi). This allows for situating the Palestinian identity historically but says little about the way that identity is actually sustained or communicated after being formed; something this study looks at through music. No population could go through decades of erasure, oppression, and false hope without it impacting part of their identity. This reflects the idea that identity and politics are inherently inseparable (Zaid et al.). In their article on the nation branding of Palestinian music, Zaid et al. quote Palestinian DJ Sama' Abdulhadi: "I never chose to be political. I'm just Palestinian, so politics is part of your name" (qtd. in Zaid et al. 681). A thought similarly shared by one of our later case studies, Saint Levant, where he expressed "my existence

as a Palestinian individual is political in its nature, so therefore any work that I do is also political" (Cambridge Union). Both artists present this politicisation as an unavoidable condition of being Palestinian which contradicts readings of politically engaged art as a deliberate stance rather than an inherit way of being. Faced with this level of displacement and occupation, the "Palestinian identity is often defined by shared experience of displacement, yearning for homeland, and the longing to return… [and] being surrounded by patriotic and folklore music" (Zaid et al. 681). Existing studies on Palestinian music have been mainly focused on earlier periods of oppression. Massad highlights the way revolutionary and nationalist songs track the political history of Palestine from the Nakba onward. On the other hand, Chosy portrays the way musical resistance has evolved. However, neither of these accounts address what happens to that resistance once it goes from being aimed at a primarily Palestinian or Arab audience to reaching international listeners with no lived connection to the occupation at all. Despite the fact that this study centres on music produced largely for and within a Palestinian or Arab audience, less has been said about the more recent wave of Palestinian and diaspora artists who are reaching international audiences and changing how Palestinian identity is viewed from outside. Furthermore, rapper Safa Hathoot expressed the internal identity confusion faced by Palestinian youth in Israel by rapping "we say we're Palestinian but… outside they treat you as Israeli" (Maira and Shihade 18).

Having established how Palestinian identity itself has been theorised, it is necessary to turn into how music specifically has been theorised as a means for that identity. Ethnomusicology situates music as the intersection of culture and identity (Đorđević). Lidskog explains that music helps diasporic communities in two ways: (1) passes memories and culture through generations and (2) it allows people to express their own identity and tell their own stories through it. This framework, however, follows the idea that the transmission of culture and self-expression is generally available to diasporic communities. It does not account for what happens when the act of passing on memory or telling a community's story is itself under threat as it is for Palestinians. Hesmondhalgh cautions that this relationship should not be understood as solely positive, although music can help people express and preserve who they are, its impact is still influenced by unequal media and political structures. Within Palestinian music studies specifically, Massad explores how songs about Palestine have tracked the political history of the struggle since 1948, while Maira and Shihade and Zaid et al. show how more recent Palestinian music continues this work of identity construction and resistance under conditions of occupation and erasure. Across the three, music tends to be read as evidence of identity or resistance that is already established instead of being something actively changing and evolving. These existing studies, mainly address music produced for and circulating within a Palestinian, Arab, or diasporic audience before the era of global steaming and social media. As a result, fewer studies have looked at the current generation of Palestinian and diaspora artists that are actively recreating the traditional musical resistance for a global stage. This is the gap this study addresses.

Given all of this, why music specifically? As Brehony put it, “music is not a universal language but a site of confrontation” (Brehony 12). Barghouthi argues that during the Intifada, collective folksongs “played a high activist role and added much fuel to the Intifada fire” (Barghouthi 154) and generally that Arab folk traditions “contributed immensely and effectively to the highlighting and maintaining of the Palestinian national identity” (Barghouthi 151). However, this function of music has been mainly explored in a context of earlier periods of the occupation and within a national context. This study further contributes to the existing works by focusing on the current more global wave of Palestinian music.

Occupation, Erasure, and the Conditions for a New Wave

Before turning to the songs themselves, it is necessary to trace the historical conditions that made this new wave of musical resistance possible. In 1948, the Nakba, translated as ‘catastrophe’ (Shehadeh), accompanied the creation of the state of Israel and caused mass displacement of the Palestinian population. Several towns and villages were depopulated and renamed under Hebrew place names. This has been documented as part of the erasure of the pre-1948 Palestinian landscape (Chosy). Palestinian refugees displaced in 1948 and their descendants have since been denied the right of return under Israeli law (Shehadeh). Israel’s founding narrative has historically drawn on both biblical claims to the land and the framing of the 1948 war as a defensive war of independence (Shehadeh). Throughout the years, Palestinians have had glimpses of a possible ending to the occupation and displacement and of a potential return to their lands. These glimpses have come and gone through: the 1948 Nakba, the 1967 Six-Day War (after which Israel occupied the West Bank, Gaza, and East Jerusalem), the Oslo Accords of 1993 - 1995 (which limited Palestinian self-governance without resolving core issues of the state, borders, or refugees), and the 2023 Gaza War, caused by the October 7 Hamas attack and Israel’s military campaign (Shehadeh). The current new wave of Palestinian music has emerged more recently to a greater audience. Two conditions specifically could help us explain why it is happening now. First, efforts to suppress Palestinian advocacy and resistance have intensified since the genocide on Gaza began in 2023 (“Israel Has Committed Genocide”). This is reflected through the censorship of performances, the surveillance and prosecution of activists such as Mahmoud Khalil (Chater), and the restrictions on pro-Palestinian expression. This has left music as one of the only remaining means for open self-expression because it is less easily policed than public speeches or assembly and it allows for political meaning to be held while moving across borders and platforms. Second, the new generation of Palestinian and diaspora artists has grown up with a stronger ambition for freedom and self-representation than previous generations. These artists increasingly work explicitly to challenge Western-imposed narratives about Arab and Palestinian identity (Reda). Artist Zeyne further expressed this idea when she mentioned she feels the newer generation is less afraid of speaking up when it comes to their music (afikra - عفكرة).

Identity, Narrative, and Decolonisation

Building on the historical and political conditions behind this new wave, this section turns to the theoretical framework used to examine it. This study draws primarily on Stuart Hall's theory of cultural identity as its main analytical theory. Throughout, Said's work on narrative and representation in *Culture and Imperialism* and Césaire's work on culture's role in decolonisation in Discourse on Colonialism are referenced as secondary theoretical frameworks.

Hall distinguishes two ideas of cultural identity: (1) a stable, shared 'true' identity rooted in a common history and (2) identity as a product that is "never complete, always in process" (Hall 222). This study makes use of the latter definition. Hall also frames diaspora identity through hybridity and a recoverable homeland, which is a lens this thesis uses to read homeland in Palestinian music as constructed and relational rather than simply nostalgic while remaining attentive to the specific conditions of displacement and denied return that shape the Palestinian case (Shehadeh). This understanding of the homeland fundamentally challenges the Western Enlightenment conceptions that prioritise state sovereignty, national borders, and property ownership. While Hall shows how identity is continuously produced, Said explains who controls the terms of that production.

Said's *Orientalism* shows how representation can itself be a representation of power by constructing a people as 'Other' (Said, Orientalism). In *Culture and Imperialism*, Said argued that "the power to narrate, or to block other narratives from forming and emerging, is very important to culture and imperialism" (Said, Culture and Imperialism xiii) and that "national culture organises and sustains communal memory" under displacement (Said, Culture and Imperialism 215). This study takes Said's emphasis on narrative control as one of its central theoretical claims: that Palestinian music resists erasure by asserting the right to narrate.

This raises the question of what cultural identity becomes under colonialism. Colonial regimes of representation had the power to "make us see and experience ourselves as 'Other'" (Hall 225). Furthermore, even after independence, "institutions of the now former colony often still continue to benefit the former coloniser and subjugate the formerly colonised subjects" (qtd. in Okazaki et al. 93). Through this political, social, and economic presence, culture and identity become subject to change. This does not mean that identity becomes foreign or one of the coloniser's but that it gains added complexity and a colonial history to it.

Finally, Aimé Césaire's studies on culture and decolonisation more specifically through his work alongside Léopold Sédar Senghor in the concept of 'négritude' add a valuable different angle to this framework. In *Discourse on Colonialism* (Césaire), he argues that colonialism works by "thingifying" the colonised community. This is done through stripping them of their own history and culture and reducing them to objects viewed from the outside (Césaire). Decolonisation therefore requires the colonised to take back their voice and culture instead of accepting the identity created for them. Négritude essentially treats cultural self-expression as a political act instead of solely an aesthetic one. Beyond Hall and Said, Césaire's work allows an understanding of cultural production as a way through which an occupied and colonised people reclaim their own narrative and identity in the midst of erasure. These three frameworks work together: Hall establishes that identity is continuously produced, Said shows that this production

is directly linked to who gets to narrate it, and Césaire highlights that reclaiming voice and culture is itself a political act. Applying this to Palestinian music, this combination makes it possible to treat the songs analysed simultaneously as acts of identity construction and as acts of reclaiming control over their own narratives. With this theoretical framework set, the following section explains how the songs analysed here were selected and interpreted.

Methodology

This study uses qualitative thematic analysis, which allows meaning to be drawn from language, image, and narrative in order to consider their independent contexts. The songs were selected according to three criteria: (1) the artist identifies as Palestinian or Palestinian adjacent, (2) the material spans a range of genres and locations, including the occupied territories, Palestinian citizens of Israel, and the diaspora, to capture the Palestinian identity in all its complexity, and finally (3) songs that engage explicitly with themes of homeland, memory, exile, resistance, or return. Once a song met these criteria, it was then analysed through three types of evidence: lyrics which show how identity is narrated, the accompanying music videos which show how identity is visualised through places, objects, and imagery, and where relevant, artist interviews which reveal the intentionality behind those choices. Analysing all three together allows themes to be explored in their different shapes – textual, visual, and testimonial evidence – avoiding the risk of reading lyrical meaning without knowing artist intent, or treating visual symbolism as self-evident without lyrical context. The reading of each song then led into a structured coding process. The coding process involved: (1) becoming familiar with the full dataset and (2) making a list of initial codes from recurring words, images, and phrases, then (3) grouping those codes into themes such as homeland, memory, resistance, exile, and belonging, and finally, (4) a review of these themes against the dataset as a whole to confirm they held across the material. Some limitations of this study could be: (1) where songs are performed in Arabic, analysis of translated lyrics risks losing meaning that does not survive translation, and (2) the samples, while deliberately varied, remain a small selection that cannot and should not represent Palestinian music as a whole. Despite these limitations, analysing a varied sample through lyrics, imagery, and testimonials together makes it possible to trace how Palestinian identity is constructed and sustained across different forms, thereby directly addressing the research question.

Reclaiming the Narrative

The dataset analysed here currently includes four songs carefully selected to capture a range of genres, production contexts, and generational voices within contemporary Palestinian music: (1) *Deira* by Saint Levant featuring MC Abdul, a hip-hop track built on a sample of Algerian chaabi and filmed in Jordan, (2) *Ghosn Zeytoun (Olive Branch)* by Elyanna, an Arabic language ballad staged around archival home video footage, (3) *7arrir 3aqlak / Asli Ana* by Zeyne, filmed across Morocco and Jordan and blending Levantine pop with explicit commentary on Westernisation, and (4) *SEEDS* by Szaba, which intercuts English and Arabic verses with an

archival interview featuring Ghassan Kanafani from 1970 (Anti Imperialist Front TV). Despite this variety of genre, language, and setting, the coding process revealed recurring concerns. These included: an emotionally loaded relationship to homeland and possibility for return, the encoding of memory in recurring motifs of symbolic objects, heritage understood as something actively remade and preserved, and resistance expressed as refusal to let Palestinian self-representation be erased or renamed. The following sections take each theme in turn drawing on lyrics, visual imagery, and artist interviews together.

Hall's theory of cultural identity gives an interesting point of view for how homeland works across the different songs. None of these songs treat it as a settled occupiable place, they instead construct it around the longing and the promise of return. This reflects the non-territorial conception of homeland introduced in the theoretical framework which challenges Western Enlightenment ideas rooted in sovereignty and property. In *Deira*, this construction is reflected through emotional impact instead of geographical location, where Saint Levant sings "oh, my homeland / joy and sorrows / oh, my precious homeland / the beloved Palestine is calling me". Through these lines, we see themes of memory and grief in the same thought which not only describes Palestine but actively creates it as a site of identification, the roots, the motherland. This song carries even more weight when the song title itself is explained by Saint Levant. "Deira" was the name of the hotel Saint Levant's parents built in Gaza in 2000 after moving there. As Levant states, "On July 16, 2014, four children playing ball out front of the hotel were killed by Israeli rocket fire," he then added that "most recently the hotel was totally destroyed by bombing three months ago." (Bain). Therefore, the building of Palestine as a place of identification that carries great weight reinforces Hall's point that identity through representation circles back to the historical roots of it, adding to its complexity (Hall). On the other hand, Elyanna's *Ghosn Zeytoun (Olive Branch)* reflects a similar idea visually rather than lyrically. In her performance, she sings around overgrown grass reflecting a land that waits for the return of its people and a white cloth hung resembling a shroud, wings, and sail at once. This reinforces the idea that the homeland doesn't belong in a single settled image, it remains contested and incomplete; the very same way Hall describes identity as something always in process. In Elyanna's own words in the interview, "Palestine is me. It's my identity" (#ABtalks), closing the gap between the person and the place altogether. Both these songs suggest that homeland functions exactly as Hall's theory would predict: not a simple geographical place but a part of their identity that they are actively and continuously making.

The way homeland functions as an emotional anchor in these songs, memory is instead reflected more strongly through recurring physical motifs: keys, seeds, Jaffa oranges, old photographs. Memory through this lens is sustained not only by individuals but also by objects and places that hold a community's shared past; it's a reminder of the place they hope to return to. The significance of these objects becomes more impactful once linked with the erasure already discussed in this paper's historical section, where Palestinian towns and villages were depopulated and renamed under Hebrew place names after 1948 (Shehadeh). Bearing in mind that history, these recurring objects look like substitutes for a landscape that can no longer be

visited or named as it was. Keys appear multiple times throughout the visuals of *7arrir 3aqlak / Asli Ana* (sometimes held, most times worn with traditional clothing during a Dabke performance), and in *SEEDS* lyrically where Szaba raps "took his keys and they told him "Pack up" / So I'ma take those same keys, open that shit back up". The key is one of the more famously recognised symbols of the Palestinian right of return serving as a reminder of the actual house keys many families carried with them when they were expelled or fled in 1948, keeping the hope that return will happen. The two songs use this same object to carry a similar message but through different means: one lyrically, the other visually. The keys within the Dabke performance in *7arrir 3aqlak / Asli Ana* turns what is a private family object into a collective one; it then becomes something the whole group visibly carries instead of a personal heirloom. It essentially turns a personal act of memory into a public, communal act of continuity. Szaba's line does the opposite by turning the moment of expulsion ("they told him 'Pack up'") into a declaration of return ("I'ma take those same keys, open that shit back up"). Across these two representations, the keys stop being evidence of what was lost and become a promise of what is still owed. Furthermore, both are paired with the idea of continuity: Szaba raps "we were the seeds they left behind" while Zeyne, once again, conveys the same idea visually, through the depiction of a grandmother taking jewellery off herself to hand it to both Zeyne and a little girl, presenting memory as something handed down bodily between women, across generations. Here the keys carry significant meaning as they make reference to the many displaced families kept from homes they were never allowed to return to, by pairing this with the idea of continuity; it creates an active reminder of and testament to continued longing for a return. *SEEDS* reflects this intergenerational transmission further by sampling an archival interview audio from 1970, in which Ghassan Kanafani describes his people being "killed and starved for 20 years" (Anti Imperialist Front TV). By importing a 1970 testimony directly into a present-day track, SEEDS makes memory literally intergenerational, not solely thematic. The choice of interview adds depth as Kanafani was assassinated in 1972 which means his voice, brought back into a 2024 track, is reviving a silenced witness decades after his death. Following Said's framework, this narrative reclamation directly hands the floor to Kanafani instead of speaking on his behalf. This essentially collapses the distance between a contemporary artist describing the past and the past testifying for itself. The Jaffa orange motif carries a similarly layered history, tied to Jaffa's pre-1948 agricultural economy, the displacement of its population in the Nakba and the fruit's survival today as a marker of a place many of the artists' families cannot visit. Even more understated visual details are the black and white home video footage in *Ghosn Zeytoun* or even the goat hair textiles glimpsed in *Deira*'s rural scenes. They work similarly as objects that already carry memory before song adds anything to them. Building on Hall's point that identity suggests that memory in these songs is doing more than longing nostalgically, it is actively building the continuity of Palestine in the midst of erasure.

As mentioned, memory is about carrying the past towards a hopeful future, heritage in this dataset is Hall's production of identity turned towards tradition. It is not something that is preserved intact but rather something that throughout generations is reworked, mixed, and

reclaimed into the present (Hall). This is seen in *Deira* through its very composition, Saint Levant has explained that the track fuses Algerian chaabi, sampled from Dahmane El Harrachi and representing his mother's heritage, with Palestinian lyrics and imagery representing his father's, describing it as a result of a mix of both cultures (Bain). Heritage in this sense is explicitly hybrid, constructed, and refined instead of being inherited as a whole, which is what Hall's theory would predict. Zeyne makes a related but more explicit political claim about Dabke in an interview by calling it "the dance that is connected to the land" and describing her generation as "finally reclaiming who we are" against what she frames as a Western imposed narrative about Arabic identity (afikra - عفكرة). Elyanna frames her own creative process in similar language describing it as "reusing old poetry and transforming it into my own style and my own world" (#ABtalks), treating tradition explicitly as raw material for reinvention instead of something inherited and left untouched. Zeyne's mention of "reclaiming who we are" mirrors Césaire's 'négritude' through which shared Black identity fights against colonial racism, her claim on Dabke reclaims Palestinian identity, which is made more significant through her mention of a Western-imposed narrative (afikra - عفكرة). Together, these three artists suggest heritage functions as Hall theorises identity more broadly; never a fixed thing to be recovered and passed down but something that is continually produced and repositioned by the people making and remaking it.

While memory and heritage reflect identity being reworked, the final theme concerns who holds the authority to narrate that identity. Said's work in *Culture and Imperialism* on culture under imperialism brings up the importance of narrative control: "the power to narrate, or to block other narratives from forming and emerging, is very important to culture and imperialism" (Said, Culture and Imperialism xiii) and that national culture sustained through story, song, and symbol "organises and sustains communal memory" even under conditions of dispossession (Said, Culture and Imperialism 215). This makes the simple right to narrate a form of resistance. Across this dataset, Palestinian music is refusing to let Palestinian self-representation be erased, renamed, or narrated by others. In this context, cultural representation becomes what Said identifies as the "real struggle" over political interest (Said, Culture and Imperialism 57). This also reflects Césaire's idea of colonialism as a process of "thingifying" the colonised (Césaire), reducing an entire people to an object to be described. Through this lens, the songs analysed stand against that reduction by insisting that Palestinians narrate their own history and identity. This is where memory and resistance overlap, the same erasure of village names after 1948 that gives memory objects their significance is what these songs are narratively pushing back against. In *SEEDS*, the watermelon appears as a direct visual reminder of a historical restriction on displaying the Palestinian flag (Sella), and its appearance here alongside the kufiyah and a literal 'Peace for Gaza' remembrance parcel draws that history of censorship into a contemporary visual vocabulary. Objects that once served as substitutes for banned symbols are now reused as signs of continuity. The same track's archival interview clip, in which Ghassan Kanafani is dismissed for saying his people have been "killed and starved" and interviewer Richard Carleton argues that being banned from using their own name is preferable

to being killed, links the erasure of Palestinian identity directly to physical survival (Anti Imperialist Front TV). *Deira*'s line "I promise you the caged bird sing in the wintertime" rapped by MC Abdul, reflects a quieter version of the same refusal, caged but not silenced, an image that may be drawing on the Palestinian liberation song tradition and, in English, on Maya Angelou's caged bird, though a stronger source is needed to confirm those connections (Genius). Zeyne highlights the underlying logic most directly in an interview, describing her generation as "rewriting our narrative, the narrative that has been imposed on us by the West" (afikra - عفكرة). Read against Said's work, this narrative control is stated as a goal of the artists themselves. Across *SEEDS*, *Deira*, and Zeyne's own words, we see resistance take the forms Said and Césaire describe; Said's insistence on narrating one's own story and Césaire's reclaiming of a voice suppressed for too long. Together, these artists make narration itself an act of resistance.

Conclusion

This study sets out to ask what role this new wave of Palestinian music plays in preserving Palestinian identity under conditions of occupation and displacement. This study argues that this new wave of Palestinian music both preserves the Palestinian identity and actively constructs and reworks that identity using memory, homeland, and resistance as a lens through which identity is remade under occupation and displacement. Through the artists' own imagery, memory, and language, they are actively standing against the erasure, redefinition, and silencing that have for too long contributed to the struggle and oppression of the Palestinian people since the Nakba. Looking through the lens of Hall's identity theory, the four songs analysed here: *Deira, Ghosn Zeytoun (Olive Branch), 7arrir 3aqlak / Asli Ana,* and *SEEDS* actively work to build on the existing Palestinian identity through longing for homeland, objects that carry memory across generations, and heritage that is continuously reworked and changed instead of being passed down intact. Through Said's lens, "the power to narrate… is very important to culture and imperialism" (Said, Culture and Imperialism xiii), these songs can also be seen as narrative control, refusing the erasure and renaming that has shaped the Palestinian experience since the Nakba (Shehadeh). The removal of Saint Levant's Coachella performance with which this paper opened is only one angle of the same struggle over who gets to narrate Palestinian identity. Music's continued circulation despite such erasure is itself evidence of its role in sustaining that identity. This dataset remains small, and as noted in the methodology, analysis of translated Arabic lyrics risks losing meaning specific to the original language. Future work could extend this analysis to a larger, more genre-diverse sample, or turn to audience reception instead of artist output alone. What this study suggests is that in the continued absence of political resolution, music remains one of the few spaces in which Palestinian identity can be actively preserved, remade, and passed on.

Works Cited

#ABtalks. #ABtalks Music With Elyanna | Vol. 11 | اليانا مع. YouTube, 2025, https://www.youtube.com/watch?v=tqAl5cypTqU. Video.

afikra - عفكرة. On Reclaiming Our Narrative & the Power of Dabke | Zeyne at XP Air. YouTube, 2026, https://www.youtube.com/watch?v=ddY3zXEmjOY. Video.

Anti Imperialist Front TV. PFLP Ghassan Kanafani, Richard Carleton Interview COMPLETE. YouTube, 2020, https://www.youtube.com/watch?v=Veoy32G7trY. Video.

Bain, Katie. 'Gaza Rapper MC Abdul & Saint Levant Release Music Video for "Deira"'. Billboard, 24 Feb. 2024, https://www.billboard.com/music/music-news/gaza-rapper-mc-abdul-saint-levant-release-music-video-deira-1235614217/.

Barghouthi, Abdullatif. 'Arab Folksongs and Palestinian Identity'. Journal of Mediterranean Studies, vol. 6, no. 1, 1996, pp. 147–72.

Brehony, Louis. Palestinian Music in Exile. American University in Cairo Press, 2023.

Cambridge Union. Saint Levant | Cambridge Union. YouTube, 2024, https://www.youtube.com/watch?v=e9x9mf4D4N0. Video.

Chater, James. 'US Judge Orders Deportation of Palestinian Activist Mahmoud Khalil'. BBC, 18 Sept. 2025, https://www.bbc.com/news/articles/cj9zv8xnml8o.

Chosy, Mackenzie. 'Palestinian Musical Resistance : From NAQBA to Now'. University of Missouri--Columbia, 2024, https://doi.org/10.32469/10355/104791.

Césaire, Aimé. Discourse on Colonialism. 1950. Monthly Review Press, 2000.

Elyanna اليانا. Elyanna - Olive Branch (غصن زيتون). YouTube, 2023, https://www.youtube.com/watch?v=zymQkDlMbi4. Video.

Hall, Stuart. Cultural Identity and Diaspora. 1990.

Hesmondhalgh, David. 'Towards a Critical Understanding of Music, Emotion and Self‐Identity'. Consumption Markets & Culture, vol. 11, no. 4, Nov. 2008, pp. 329–43, https://doi.org/10.1080/10253860802391334.

'Israel Has Committed Genocide in the Gaza Strip, UN Commission Finds'. Ohchr, 2025, https://www.ohchr.org/en/press-releases/2025/09/israel-has-committed-genocide-gaza-strip-un-commission-finds.

Khalidi, Rashid. The Hundred Years' War on Palestine : A History of Settler Colonialism and Resistance, 1917-2017. Picador, 2020.

Lidskog, Rolf. 'The Role of Music in Ethnic Identity Formation in Diaspora: A Research Review'. International Social Science Journal, vol. 66, nos 219–220, Mar. 2017, pp. 23–38, https://doi.org/10.1111/issj.12091.

Maira, Sunaina, and Magid Shihade. 'Hip Hop From "48 Palestine'. Social Text, vol. 30, no. 3, 2012, pp. 1–26, https://doi.org/10.1215/01642472-1597314.

Massad, Joseph. 'Liberating Songs: Palestine Put to Music'. Journal of Palestine Studies, vol. 32, no. 3, 2003, pp. 21–38, https://doi.org/10.1525/jps.2003.32.3.21.

Reda, Latife. ‘Sensing Palestine: Global Decolonial Aesthetics in Latin American–Palestinian Musical Collaborations’. Alternatives: Global, Local, Political, Feb. 2026, https://doi.org/10.1177/03043754261422801.
Said, Edward. Orientalism. Pantheon Books, 1978.
Said, Edward W. Culture and Imperialism. Vintage, 1993.
Saint Levant. Saint Levant - Deira Ft. MC Abdul (Official Video). YouTube, 2024, https://www.youtube.com/watch?v=vRTtwtjNLdU. Video.
‘Saint Levant & MC Abdul – Deira’. Genius, 2024, https://genius.com/Saint-levant-and-mc-abdul-deira-lyrics.
Sella, Adam. ‘The Fruits and Vegetables of Palestine and Their Significance’. Www.Aljazeera.Com, 31 Aug. 2023, https://www.aljazeera.com/news/2023/8/31/the-fruits-of-palestine-and-their-symbolism.
Shehadeh, Raja. What Does Israel Fear From Palestine? Profile Books, 2024.
SZABA. Szaba & L Don - SEEDS (Official Music Video). YouTube, 2024, https://www.youtube.com/watch?v=pcxC5Y84YAM. Video.
Zaid, Bouziane, et al. ‘Palestinian Music and the Potential for the Nation Branding of Palestine’. Middle East Critique, vol. 33, no. 4, Oct. 2024, pp. 673–90, https://doi.org/10.1080/19436149.2024.2438517.
zeyne زين. Zeyne - 7Arrir 3Aqlak / Asli Ana | زين - حرر عقلك / اصلي انا (Official Music Video). YouTube, 2024, https://www.youtube.com/watch?v=xhH6CON3SD8. Video.
Đorđević, Ana. ‘A Consideration of the Relationship Between Music and Collective Identity: From the Viewpoint of Cultural Psychology of Music’. Etnoantropološki Problemi / Issues in Ethnology and Anthropology, vol. 15, no. 3, Sept. 2020, https://doi.org/10.21301/eap.v15i3.11.

Digital Navigational Aid for the Visually Impaired using Edge AI: Experimentally-Supported Advantages of Incorporating Safety-Oriented Deterministic Algorithms By Rudra Patel

Abstract

For blind or low-vision individuals, the challenges in navigating complex indoor and outdoor spaces has stimulated the development of AI-powered digital aids that visually interpret the user's surroundings. However, current solutions are relatively narrow-scoped and computationally expensive to run on budgeted hardware and network resources. This paper evaluates the effectiveness of a new mobile app architecture designed to be lightweight, versatile, and highly responsive to immediate hazards. With limited resources available on the target device, the prototype demonstrated a strong ability to detect objects accurately, generate useful verbalized guidance, and appropriately alert the user of various safety hazards.

1. INTRODUCTION

1.1 Motivation

For visually impaired individuals navigating unknown environments, hearing, touch, and spatial senses are key resources for staying safe, wayfinding, and maintaining environmental awareness. Established aids that address these barriers typically include long white canes for tactile feedback and guide dogs. However, users of these products may still feel the need to limit their independent travel due to other persistent limitations - for instance, canes are mainly designed to warn the user of ground-level hazards immediately upon contact, which does not account for overhanging obstacles or quickly moving hazards. These challenges have spurred exploration of high-tech AI digital aids as a supplement to typical forms of navigational tools.

A digital system that interprets visual cues in a way useful to humans will involve computer vision and recognition, depth reasoning, distinguishing safety-critical observations from non-essential ones, prioritizing details, language manipulation, and tracking change over time. This makes it a promising application for artificial intelligence.

An example of such a digital aid would take a visual assessment of the surrounding environment with a camera and a vision model to determine the user's best next steps, then somehow communicate that information to the user - perhaps by spoken language generated by a language model. However, language-model inference can impose substantial computational and energy costs, and commonly used sampling-based decoding methods intentionally introduce variability into generated outputs, posing safety risks when quickly detectable hazards must travel through the non-vision pipeline and be conveyed in an inappropriate way. Instead, priorities shift to low latency, semantic precision, and high conciseness to give the user adequate notice.

1.2 Literature

Independent navigation remains a significant challenge for people who are blind or have low vision because many environments communicate essential information primarily through visual signs, markings, landmarks, and changes in spatial layout (Başgöze et al., Bamdad et al.). Traditional aids such as the aforementioned white canes and guide dogs provide indispensable support, but they do not communicate every type of environmental or route information. Studies of navigational-aid use and user needs show that travelers select and combine resources according to the environment, task, and perceived risk (Başgöze et al., Bamdad et al.). El-Taher et al. found that existing outdoor applications often provide insufficient information about obstacles, intersections, and traffic signals, while Kameswaran et al. observed that users commonly combine navigation applications with conventional mobility aids, orientation-and-mobility skills, and other technologies. These findings suggest that an electronic navigation aid should complement rather than replace established mobility practices and should communicate information that is both relevant to the immediate task and unavailable through the user's other senses or aids.

Consequently, researchers have developed a broad range of electronic travel aids using smartphones, wearable cameras, depth sensors, positioning technologies, and machine-learning models (Real/Araujo, Plikynas et al., Budrionis et al.). Reviews by Real/Araujo and Plikynas et al. classify these systems according to functions such as localization, environmental perception, obstacle avoidance, and user feedback. These reviews also identify recurring limitations involving reliability, cognitive load, portability, infrastructure requirements, and insufficient evaluation with intended users. The systematic review by Budrionis et al. similarly found a mismatch between the capabilities emphasized in smartphone computer-vision prototypes and the needs identified by visually impaired users. Collectively, this literature indicates that successful object recognition or localization alone does not constitute a complete navigation aid. A practical system must convert uncertain sensor observations into timely, understandable, and appropriately prioritized guidance while operating within a mobile device's computational and physical constraints.

Indoor-navigation research demonstrates both the potential and the limitations of current engineering approaches. Ahmetovic et al. evaluated Bluetooth beacons for smartphone localization and exposed a trade-off between localization accuracy, beacon density, and installation effort. NavCog3 demonstrated that smartphone guidance with semantic environmental information could support visually impaired travelers in a large shopping mall, providing an important example of evaluation beyond a small laboratory environment (Sato et al.). Other systems have explored different feedback modalities: Barontini et al. combined RGB-D perception with wearable haptic feedback through a user-centered design process, while EchoSee used on-device three-dimensional reconstruction and spatial audio to represent nearby obstacles (Schwartz et al.). These systems demonstrate that useful navigation information can be delivered through several modalities, but they also reveal continuing dependencies on installed infrastructure, specialized sensors, user training, or hardware available only on certain devices. They therefore support further investigation of comparatively infrastructure-light,

smartphone-based systems while cautioning that laboratory functionality does not by itself establish real-world navigational effectiveness.

A related engineering problem is deciding which detections require intervention and how the system should behave when its perception is uncertain. Islam et al. identified 90 objects particularly relevant to blind and low-vision navigation and found that widely used computer-vision datasets represent only a subset of them. This suggests that conventional object-detection metrics may not reflect whether a model recognizes the hazards and landmarks that matter most during travel. Moreover, navigation-system errors differ substantially in consequence. Abdolrahmani et al. found that users' acceptance of errors depended on the error type and environmental context; some inaccuracies were manageable, whereas others could interfere with safety or orientation. Recent wearable research illustrates the performance demands involved: Gao et al. reported collision-avoidance, latency, power, and operating-duration results for a system combining video, depth sensing, smartphone processing, custom hardware, and multimodal alerts. Although these results cannot be compared directly with a phone-only prototype, they demonstrate that responsiveness, reliability, and resource consumption must be considered together. This evidence motivates a decision architecture that prioritizes navigation-specific hazards and handles urgent actions predictably, rather than treating every model output as equally reliable or important.

Generative artificial intelligence offers new opportunities to translate technical sensor outputs into more natural verbal guidance, but it also introduces risks that matter especially in a navigation context. NaviGPT combines LiDAR sensing, vibration feedback, image understanding, and large-language-model responses in a unified mobile prototype, making it a close conceptual predecessor to systems that combine environmental perception with generated instructions (Zhang et al.). However, studies of multimodal AI used by visually impaired participants have identified hallucinations, failures to understand user intent, inaccurate spatial descriptions, and inappropriate assumptions about users' abilities (Xie et al., Chang et al.). Chang et al. also found that live video AI performed better with static scenes than with the continuously changing information needed in dynamic environments. These weaknesses make an unconstrained language model unsuitable as the sole authority for time-critical safety decisions. A more defensible role is to use language generation for communication—such as producing concise, natural phrasing—while assigning hazard classification, state transitions, and urgent warnings to deterministic logic whose behavior can be inspected and tested.

1.3 Research Gaps

The literature identifies an opportunity to evaluate a hybrid mobile architecture that separates safety control from language generation. In the proposed mobile app architecture, a deterministic state engine interprets structured perception results, assigns safety states, and initiates urgent responses, while a small local language model is limited to producing noncritical explanatory or navigational guidance. This approach does not claim to solve every component of indoor and outdoor navigation; instead, it experimentally examines whether this division of

responsibility produces correct and timely safety behavior and whether the resulting language is clear, concise, and helpful. Because previous projects such as VIS4ION emphasize extended use, acceptability, generalizability, and participant-based evaluation (Beheshti et al.), conclusions from an unfinished or limited prototype should remain appropriately scoped. Initial engineering tests can establish functional feasibility and characterize system behavior, but claims concerning everyday safety or effectiveness would require subsequent evaluation with visually impaired participants in representative environments.

Additionally, constraints on network usage, transportability, accessibility, and other mobility factors will limit the practicality of an advanced AI system. This is a strong argument to consider applications of AI on a smaller scale - inferences run locally on small, isolated, environmentally aware devices that have the potential to be both resource-efficient and functionally effective. However, this requirement limits model size and creates tradeoffs among accuracy, latency, energy consumption, and maintenance. Backward device compatibility for mobile app solutions is also important for accessibility, which forces potential products to use resources even more efficiently.

1.4 Contributions

In this paper, we will investigate a dual-mode mobile app-based digital aid that uses a local-inference vision model (VM) and small language model (SLM) to interpret a camera scene and generate navigational guidance in both indoor and outdoor environments.

To reduce latency in safety decisions, a deterministic algorithm will be inserted between the vision and language layers that skips the SLM entirely upon immediate hazard detection.

Other questions to be explored will include:

- The extent to which the restricted model size hinders the system's effectiveness;
- The capability of a language model to provide clear, concise, and useful imperatives towards its purpose;
- The strengths and challenges of the system's indoor-aid function.

2. METHODS

2.1 Product Overview

The product developed as the primary focus of this paper is a prototype Android application known as "NavAssist." This app converts a live video feed and pedestrian navigational data into real-time verbal guidance using fully bundled vision and language models, with a deterministic safety pathway that responds faster to immediate hazards. It is a two-mode app designed to work not only outdoors as described, but also indoors without navigational data.

This prototype is not designed to replace any existing aids for visually impaired navigation. We strongly advise using this and similar products in practice as a supplement to more common aids, under the full supervision and oversight of a sighted human guide.

2.2 Logical Architecture

As shown in Fig. 1, NavAssist's decision-making architecture extends beyond a standalone vision-language model that converts images directly into tokens; instead, the vision and language models are separated and collaborate with deterministic algorithms that handle safety decision-making and route-informed guidance.

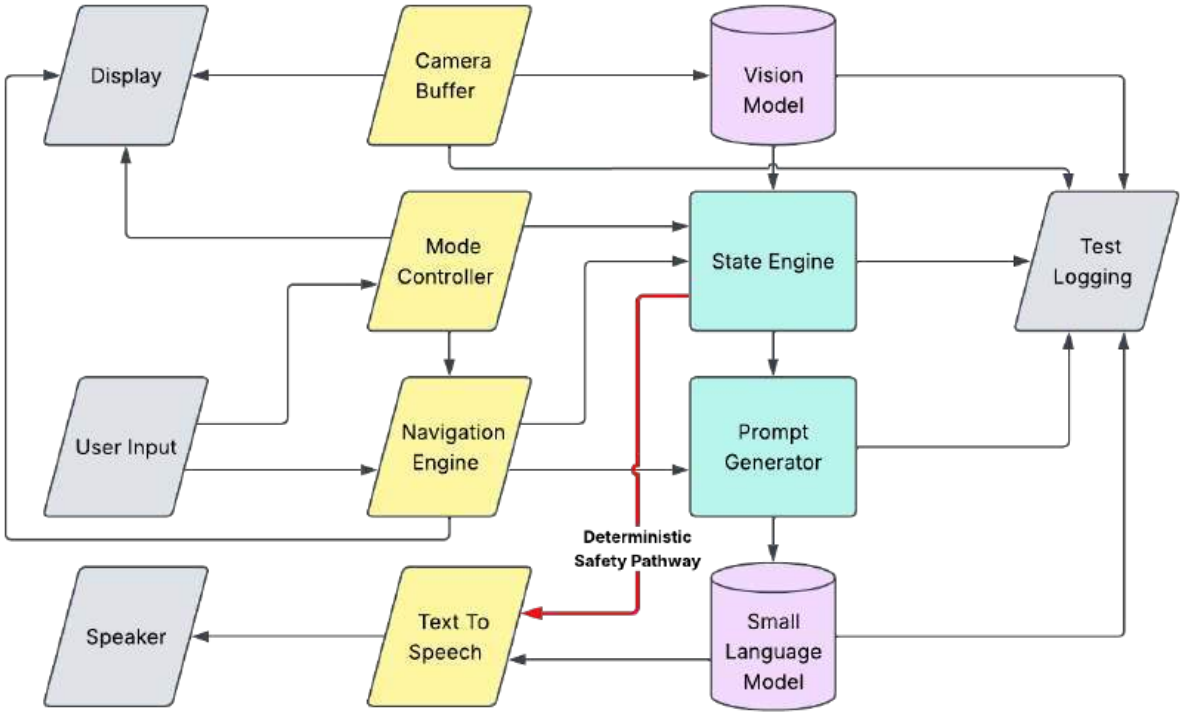


Fig. 1: A simplified diagram of the pipeline architecture of NavAssist.

A simple binary mode controller set by the user determines whether the app is in Route mode or Free Roaming mode. Route mode uses a routing & navigation API that calculates and displays pedestrian paths to a user-specified destination, similar to Google Maps, and supplies the guidance engine with route progress and positional information.

The first stage of the guidance pipeline is the vision layer, which consists of two separate vision models that work together to analyze the camera scene: an object detection model and a depth estimation model. The object detector outputs several label-confidence pairs for objects in the scene, while the depth estimator segments the field of view into generic regions and classifies each region's depth qualitatively rather than with exact numeric distance values.

Then, the state engine, a ruleset-based algorithm rather than an AI model, gathers information from multiple input components to decide whether the app must stay silent, generate SLM guidance, or take a shortcut to direct speech output with formula guidance. In addition to data from the vision layer, it also considers information from the navigation engine in Route mode to detect whether the user has a next route step to complete, is straying off their path, a particular route step is risky (i.e., crossing a street), or similar situations.

For every accepted perception observation, one for every frame, the state engine evaluates these rules in order:

1. Reject unusable observations, which may fit any of these criteria:
    a. Older than 2.5 seconds;
    b. Has the same or an older timestamp than the last processed observation.
2. Check for a detected center-path hazard. A detection qualifies only when:
    a. Its class is on the obstructive allowlist.
    b. It is in the center third of the image.
    c. Object confidence is at least 0.50.
    d. Its box reaches at least 55% down the image.
    e. Its box height covers at least 20% of the image.
    f. FastDepth marks the center region NEAR.
    g. Center-depth confidence is at least 0.75.
3. If multiple objects qualify, the highest-confidence one wins. The engine immediately produces “Stop. <object> ahead.” This bypasses the SLM.
4. Check for an unlabeled depth hazard, which will again bypass the SLM and produce “Stop. Obstacle ahead.” This catches obstacles that EfficientDet does not recognize. It requires:
    a. Center depth is NEAR.
    b. Center-depth confidence is at least 0.75.
    c. Relative depth exists for left, center, and right.
    d. Center depth is no more than 65% of the nearer side depth.
    e. This condition persists for two consecutive accepted frames.
5. Apply an urgent-warning cooldown. Each object label has its own deduplication key, while generic depth hazards use center-depth. Suppress the same warning for 4 seconds.
6. Decide whether ordinary scene guidance is useful, based on whether an object has been detected with a confidence of at least 0.6. If no relevant objects exist, or if that signature matches the last spoken scene, the engine stays silent. New scene guidance also has a 6-second cooldown.

A separate process within the state engine coordinates route maneuvers. A changed maneuver can request an SLM response even without a new relevant visual object. Rerouting and arrival are deterministic spoken events, and an urgent visual warning cancels pending SLM generation.

If the state engine determines SLM guidance is necessary, it compiles all scenic and navigational data into a prompt that follows the structure outlined in Table 1. The language model then processes the prompt and converts it into clear, concise, useful guidance. This guidance is spoken out loud by a text-to-speech engine.

Context	You provide brief spoken orientation for a visually impaired pedestrian. Reply with exactly one short English sentence of at most 20 words. Use only the supplied facts. Do not invent objects, distances, hazards, or route steps. Describe useful objects and their left, center, or right position. Do not say stop; handle urgent hazards separately.
Route mode appended text (with example data)	mode=route reason=route_maneuver_update objects: - person; position=center; confidence=0.82 depth_bands: - center=near; confidence=0.91 - left=medium; confidence=0.84 - right=far; confidence=0.88 route_status=navigating route_instruction=Turn right onto Main Street route_remaining_meters=125 route_remaining_seconds=90 Provide the single spoken orientation sentence now.
Free Roaming mode appended text (with example data)	mode=free_roaming reason=relevant_scene_change objects: - person; position=left; confidence=0.85 - chair; position=right; confidence=0.75 depth_bands: - center=medium; confidence=0.80 Provide the single spoken orientation sentence now.
Additional notes: - Objects and depth regions are sorted to make serialization deterministic. - Object bounds and raw depth values are not sent to the SLM. - Only labels, screen regions, confidence values, depth bands, and selected route facts are included. - Route instructions are limited to 160 characters and stripped of line breaks and special-token-like sequences. - Generation is greedy, limited to 32 new tokens, and has a 12-second deadline. - Output is rejected if blank, multiline, over 240 characters, contains special-token markers, or begins with an urgent stop command.	

Table 1: The structure of SLM prompts within NavAssist.

2.3 Software Dependencies

Table 2 lists the software used to develop the app. Note that the text-to-speech baseline is a remote engine that depends on network connectivity for critical function. While the app's intended vision is to reduce reliance on wireless communication as much as possible, this component is comparatively trivial, so using an API is appropriate for a first-pass implementation.

Component	Dependency
Camera preview and analysis	Android CameraX 1.6.1 (core, camera2, lifecycle, & view packages)
Synchronized video capture	CameraX 1.6.1 video / VideoCapture / Recorder
Object detection neural network	EfficientDet-Lite0 int8 from the MediaPipe model bucket
Depth estimation neural network	FastDepth pruned MobileNet-NNConv5 skip-add float16 conversion
Vision inference runtimes	MediaPipe Tasks Vision 0.10.35 for EfficientDet, LiteRT 2.1.6 for FastDepth
Language model	HuggingFaceTB/SmolLM2-135M-Instruct, deployed as QuantFactory/SmolLM2-135M-Instruct-GGUF/SmolLM2-135M-Instruct.Q4_K_M.gguf
Language-model runtime	llama.cpp v0.3.0 at commit c1d0e7a004015f23bc0233470b747b596f29b264, built into the app through Android NDK/CMake and JNI
Route/navigation provider	Google Navigation SDK for Android 7.7.0 only
Positioning and heading	Single app-owned adapter backed by Navigation SDK
Destination input	Accessible manual latitude/longitude entry, converted to a Navigation SDK Waypoint
Speech output	Android TextToSpeech platform API
Structured persistence	App-private per-session manifest plus append-only JSONL; no Room
Detailed log storage	noBackupFilesDir/field_logs/v1, JSONL + MP4, Storage Access Framework ZIP export
Architecture/dependency injection	Activity-scoped ViewModel plus manual constructor injection and the app-owned seams already present
Concurrency	Existing CameraX analysis executor plus a dedicated, single-request language-generation worker
Serialization	Versioned deterministic P3 text plus navassist-field-log-v1 JSONL/manifest
Active-session execution	Visible-activity, UI-bound session; no foreground service in the first pass

Table 2: A list of software dependency baselines upon which NavAssist has been developed.

2.4 Study Design

The app's performance will be evaluated through qualitative observations from field testing in various outdoor environments over 5 days, totaling 12 hours of trialing across 4 location types and various weather conditions. These variables cannot be controlled with respect to each other. Table 3 lists all trials performed.

3. RESULTS

3.1 Product in Action/Informal Testing

In the end product, the two modes are represented by two tabs showing mode-specific UIs, as shown in Figs. 2 and 3. Route mode includes a Google Maps-style navigation UI along with the camera view and allows the user to enter a destination as latitude and longitude coordinates. The destination-setting infrastructure can be easily improved in implementations beyond this prototype.

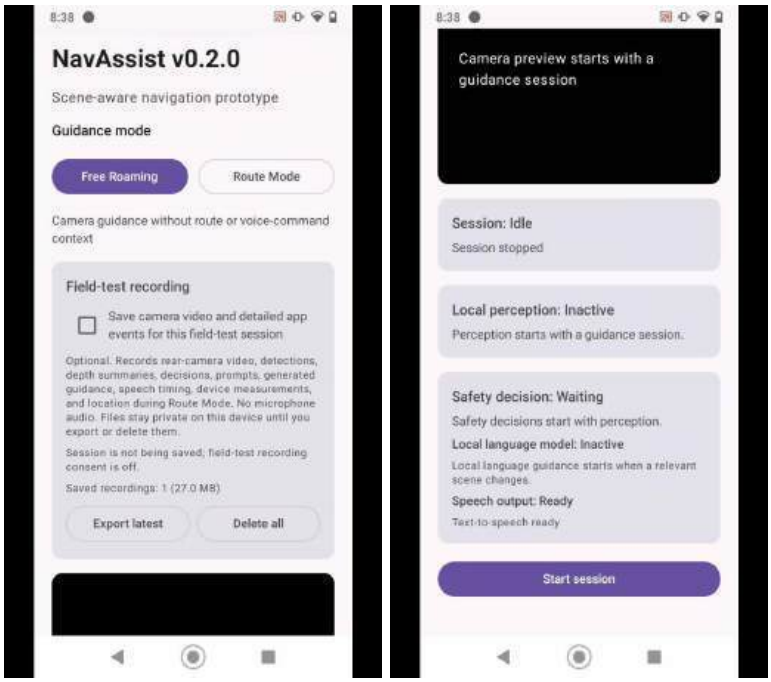


Fig. 2: Snapshots of the Free Roaming mode UI, arranged from left to right in top-to-bottom order.

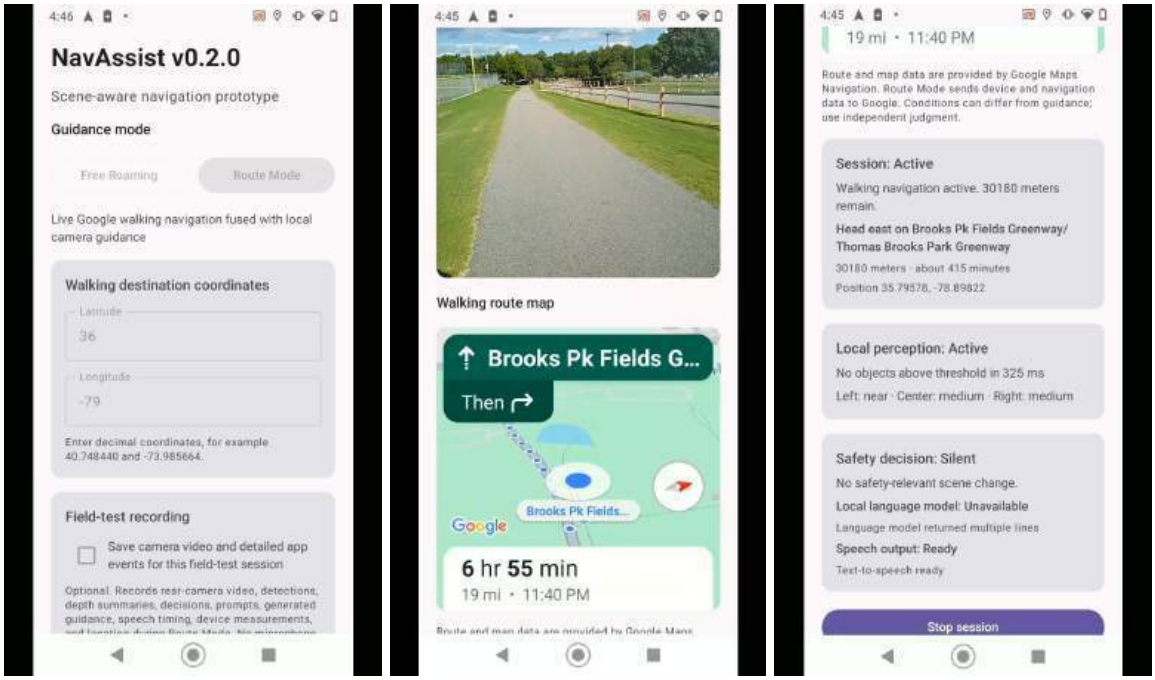


Fig. 3: Snapshots of the Route mode UI in an active session. arranged from left to right in top-to-bottom order.

When a detected object is determined to be a safety hazard, the app will speak the stop message out loud and also leave relevant messages on-screen, as shown in Figure 4:

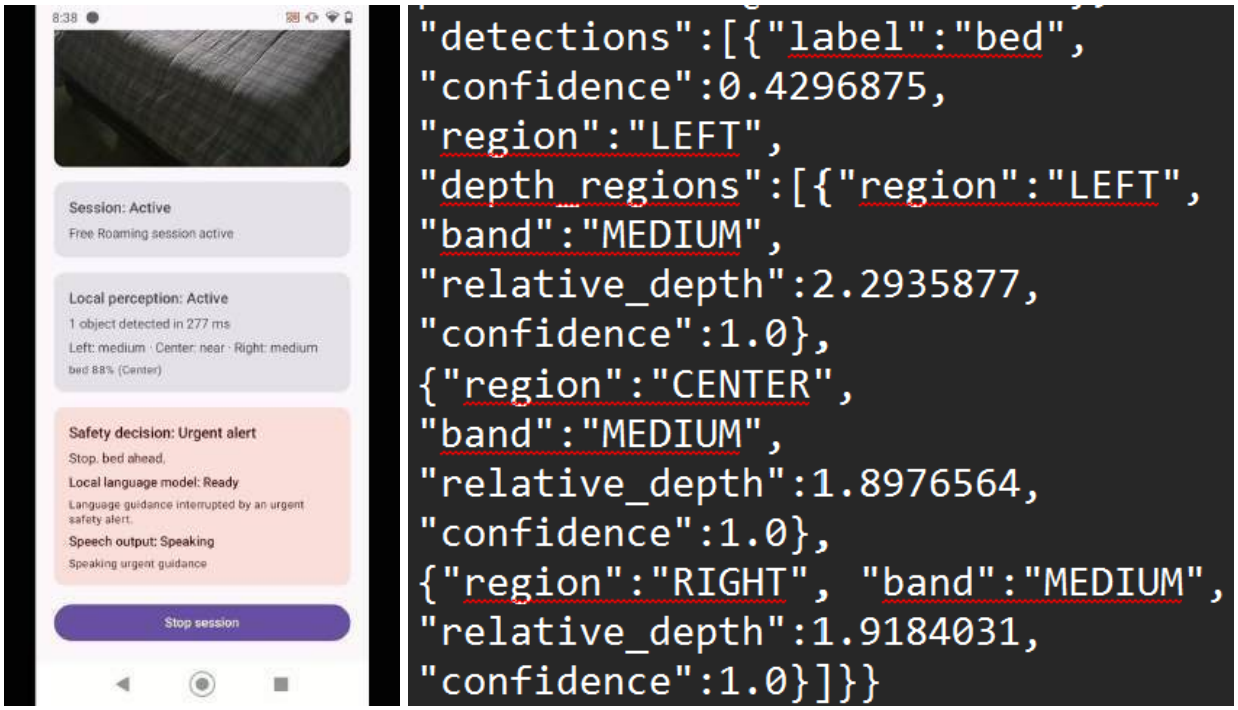


Fig. 4: Snapshots of the app UI and debugging logs detecting a bed and warning the user to stop moving.

To get an early idea of whether the product's abilities would align with its future scope, less mature iterations of the prototype were lightly tested in convenience-sampled environments - namely, rooms of a two-story house. Note that results from these informal tests do not carry the same weight as those from experimental conditions.

One preliminary iteration showed the EfficientDet-FastDepth model pair was effective at detecting people and furniture, among other standalone objects. However, it failed to identify solid walls. This is understandable, since these are foreground features explicitly omitted by an object detection model such as EfficientDet, and FastDepth would not have enough information to determine depth from a single, flat-colored frame. We can also learn this by examining EfficientDet's default label set, as summarized in Section 3.2.

3.2 Field Performance Observations

Table 3 lists qualitative observations from official field tests. Collected data also includes the quantitative statistics summary in Table 4, evaluating metrics relating to pipeline latency, model accuracy, and more. Human experts determined the ground truth for model accuracy metrics.

Conditions	Notes
4-7 PM, sunny, unpopulated soccer park, Free Roaming mode	- Rare SLM guidance responses - Generally low activity, only 14 total for the trial - Late reaction times to 57% of detectable hazards
2-4 PM, overcast, unpopulated soccer park, Free Roaming mode	- Rare SLM guidance responses - Generally low activity, only 14 total spoken outputs - Late reaction times to 28% of detectable hazards - Significantly higher object identification accuracy (83%)
7-9 PM, dark w/ artificial light, heavily populated soccer park, Free Roaming mode	- Lower object identification accuracy (77%) - Frequently missed reactions to detectable obstacles - More frequent SLM guidance than usual
2-4 PM, partially cloudy, downtown city sidewalks, Route mode	- Very concise guidance, consistently 4-8 words - Almost no directional imperatives were inaccurate (5%) - Minor misuses of "slight" and "sharp"
9-12 AM, sunny, empty parking lot with props, Free Roaming mode	- Highest detection accuracy (93%) of all trials - Detects obstacles without providing sufficient information to avoid them - No SLM guidance for a perfectly clear path

Table 3: The most important qualitative observations collected during field tests identified by input conditions.

VM labelling mistakes	9%

Obstacle false positives	3%
Late critical warnings	5%
Missed critical warnings	9%
SLM guidance mistakes	12%
Total state engine activity	28%
Total SLM activity	7%
Average VM latency	287 ms
Average total vision-language pipeline latency	962 ms

Table 4: Quantitative statistics from all field testing.

4. DISCUSSION

4.1 Successes

As shown in Table 4, the object detector model has a very high accuracy rate, and the state engine excels at accurately detecting immediate obstacles. The pre-trained labels were sufficient to cover most potential obstacles in everyday outdoor walking, but with custom benchmarked VMs in future work, usefulness may improve further.

Additionally, the SLM provided clear, concise, and useful navigation guidance. The spatial, navigational, and environmental information it received was rarely lost through shortening.

4.2 System Limitations

We can infer that the narrowed scope of detectable objects stems from the limitations of the vision models themselves. This conclusion is well supported by informal testing from 4.1. They fulfill their purpose well, but the models used are not well suited to the task. EfficientDet merely detects objects and attaches labels, without supporting background features important for traversability or providing location information.

Additionally, we have evidence of a bias against SLM activation in the state engine. To understand why this may happen, we have to consider what an ideal algorithm would look like. Without the ability to detect far-off obstacles, paths, or other non-immediate traversability restrictions, the language layer has few reasons to speak up. The only things the VM system can detect in the first place are immediate obstacles.

4.3 Future Work

As established earlier, we have strong evidence that an object detection model is not entirely practical for our use case, since interpreting potential safety hazards involves more than

simply commenting on the absence or presence of an object. A segmentation model may be preferable instead, as it splits the scene into locatable, identifiable scene regions containing objects rather than only listing the objects seen.

Even after finding better-suited models, pre-trained parameters may not be enough to ensure ideal performance right away. Allowing models to be easily benchmarked, even through separate infrastructure from the app, would enable specially tailored models that produce more useful results. Though the prototype uses hardcoded file paths and checksums for now and doesn't make swapping easy, later implementations can add swappability with simple app-owned model adapters.

The navigation engine's prompt-injection strategy for influencing LM guidance output raises the question of whether Free Roaming mode could also be modified to produce goal-oriented guidance. A useful function could be for the app to identify user-specified visual points of interest. For example, if the user is looking for a blue door leading to a bathroom, the app could identify a blue door down a hall on the right. A promising implementation could be to allow the app to accept voice input in Free Roaming mode, where the text is directly injected into the SLM prompt generator. Then, it depends on whether the language layer can connect the vision model output to the injected user request.

CONCLUSION

This study evaluated whether a resource-constrained mobile system could combine local environmental perception, deterministic safety decisions, and small-language-model guidance for visually impaired pedestrians. Within the tested scenarios, NavAssist achieved promising object-detection accuracy and reliably converted recognized center-path hazards into immediate safety alerts. The deterministic state engine prevented urgent warnings from depending on language-model completion, while the successfully generated SLM outputs generally expressed concise directional imperatives. These results provide preliminary support for separating safety-critical decisions from generative language rather than allowing a language model to control the complete perception-to-guidance process. Nevertheless, the results establish prototype feasibility under the tested conditions, not general navigational safety.

The perception results closely reflected the capabilities of the selected software dependencies. EfficientDet successfully recognized common classes represented in its default training labels, such as people, vehicles, and several household objects. However, its predefined label space does not adequately represent navigation-specific features such as walls, curbs, poles, stairs, and drop-offs. FastDepth, executed through LiteRT, supplemented these detections with monocular relative-depth estimates, allowing the state engine to determine whether the center region was comparatively near. This combination supported reliable hazard detection without requiring a recognized object in every case, but FastDepth’s low-resolution, three-region depth summary did not identify object boundaries, walkable floor space, or a safe route around an obstacle. This inability to detect or distinguish structural and ground-level hazards reflects both

the limitations of the underlying models and the reduced spatial representation adopted for on-device efficiency.

The same distinction applies to the language results. SmolLM2-135M-Instruct, quantized and executed locally through llama.cpp, demonstrated that a small offline model could transform structured scene information into short directional language. Its successful responses were assisted by a restrictive prompt, greedy decoding, a 32-token limit, and validation rules designed to prevent unsupported or urgent commands. However, some prompts produced no usable SLM response. These failures may have resulted from the small model's capacity and quantization, local CPU generation time, the generation deadline, cancellation by newer or urgent events, or rejection by the output validator. Consequently, report response absence as an end-to-end completion failure unless the session logs identify the exact stage. Moreover, the lack of detailed avoidance instructions was not simply a language-generation failure: the SLM received only object labels, coarse left-center-right positions, depth bands, and route information. Because the prompt correctly prohibited invented facts, the model had no verified representation of which neighboring space was safe to enter.

Future development should therefore improve the information supplied to the decision engine before expanding the language model’s authority. A navigation-specific detector or semantic-segmentation model could add walls, curbs, stairs, poles, doors, and other accessibility-relevant classes. Higher-resolution or temporally stabilized depth estimation, floor-plane detection, object tracking, and free-space analysis could then support a deterministic maneuver planner that selects actions such as stopping, moving left, or moving right. Only after calculating such a path should the SLM verbalize the maneuver. Language reliability could also be improved through additional task-specific training, constrained decoding, and deterministic template fallbacks when llama.cpp generation times out or produces invalid output. These additions would introduce new memory, latency, power, and thermal costs that should be measured alongside accuracy. With broader environmental testing and eventual evaluation involving visually impaired participants, the architecture could progress from a successful proof of concept toward a navigation aid whose safety behavior, response availability, and practical limitations are experimentally determined.

Acknowledgments

This work was supported and advised by Dr. Joe Xiao. We are grateful for his continuous guidance and mentorship throughout this research-writing process.

Works Cited

Abdolrahmani, Ali, William Easley, Michele A. Williams, Stacy M. Branham, and Amy Hurst. "Embracing Errors: Examining How Context of Use Impacts Blind Individuals' Acceptance of Navigation Aid Errors." Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2017, pp. 4158–4169. https://doi.org/10.1145/3025453.3025528.

Ahmetovic, Dragan, Masayuki Murata, Cole Gleason, Erin Brady, Hironobu Takagi, Kris M. Kitani, and Chieko Asakawa. "Achieving Practical and Accurate Indoor Navigation for People with Visual Impairments." Proceedings of the 14th Web for All Conference on the Future of Accessible Work (W4A '17), Association for Computing Machinery, 2017.

Bamdad, Banafshe Marziyeh, Manuel Günther, and Alireza Darvishy. "Navigating Independence: A Survey of Visually Impaired People's Experiences and Needs." ResearchGate, Apr. 2026. https://doi.org/10.13140/RG.2.2.22355.34084.

Barontini, Federica, Manuel G. Catalano, Lucia Pallottino, Barbara Leporini, and Matteo Bianchi. "Integrating Wearable Haptics and Obstacle Avoidance for the Visually Impaired in Indoor Navigation: A User-Centered Approach." IEEE Transactions on Haptics, vol. 14, no. 1, 2021, pp. 109–122. https://doi.org/10.1109/TOH.2020.2996748.

Başgöze, Z., et al. "Navigational Aid Use by Individuals with Visual Impairments." Journal on Technology and Persons with Disabilities, vol. 8, 2020, pp. 22–39.

Beheshti, Mahya, Tahereh Naeimi, Todd E. Hudson, Chen Feng, Pattanasak Mongkolwat, Wachara Riewpaiboon, William Seiple, Rajesh Vedanthan, and John-Ross Rizzo. "A Smart Service System for Spatial Intelligence and Onboard Navigation for Individuals with Visual Impairment (VIS4ION Thailand): Study Protocol of a Randomized Controlled Trial of Visually Impaired Students at the Ratchasuda College, Thailand." Trials, vol. 24, 2023, article 169. https://doi.org/10.1186/s13063-023-07173-8.

Budrionis, Andrius, Darius Plikynas, Povilas Daniušis, and Audrius Indrulionis. "Smartphone-Based Computer Vision Travelling Aids for Blind and Visually Impaired Individuals: A Systematic Review." Assistive Technology, vol. 34, no. 2, 2022, pp. 178–194. https://doi.org/10.1080/10400435.2020.1743381.

Chang, Ruei-Che, Rosiana Natalie, Wenqian Xu, Jovan Zheng Feng Yap, and Anhong Guo. "Probing the Gaps in ChatGPT Live Video Chat for Real-World Assistance for People Who Are Blind or Visually Impaired." Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25), Association for Computing Machinery, 2025, 14 pp. https://doi.org/10.1145/3663547.3746319.

El-Taher, Fatma El-Zahraa, Luis Miralles-Pechuán, Jane Courtney, Kristina Millar, Chantelle Smith, and Susan McKeever. "A Survey on Outdoor Navigation Applications for People With Visual Impairments." IEEE Access, vol. 11, 2023, pp. 14647–14666. https://doi.org/10.1109/ACCESS.2023.3244073.

Gao, Yun, Dan Wu, Jie Song, Xueyi Zhang, Bangbang Hou, Hengfa Liu, Junqi Liao, et al. "A Wearable Obstacle Avoidance Device for Visually Impaired Individuals with

Cross-Modal Learning.” Nature Communications, vol. 16, 2025, article 2857. https://doi.org/10.1038/s41467-025-58085-x.

Islam, Md Touhidul, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. “Identifying Crucial Objects in Blind and Low-Vision Individuals’ Navigation.” Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS ’24), Association for Computing Machinery, 2024, article 80. https://doi.org/10.1145/3663548.3688538.

Kameswaran, Vaishnav, Alexander J. Fiannaca, Melanie Kneisel, Amy Karlson, Edward Cutrell, and Meredith Ringel Morris. “Understanding In-Situ Use of Commonly Available Navigation Technologies by People with Visual Impairments.” Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS ’20), Association for Computing Machinery, 2020. https://doi.org/10.1145/3373625.3416995.

Plikynas, Darius, Arūnas Žvironas, Andrius Budrionis, and Marius Gudauskis. “Indoor Navigation Systems for Visually Impaired Persons: Mapping the Features of Existing Technologies to User Needs.” Sensors, vol. 20, no. 3, 2020, article 636. https://doi.org/10.3390/s20030636.

Real, Santiago, and Alvaro Araujo. “Navigation Systems for the Blind and Visually Impaired: Past Work, Challenges, and Open Problems.” Sensors, vol. 19, no. 15, 2019, article 3404. https://doi.org/10.3390/s19153404.

Sato, Daisuke, Uran Oh, Kakuya Naito, Hironobu Takagi, Kris M. Kitani, and Chieko Asakawa. “NavCog3: An Evaluation of a Smartphone-Based Blind Indoor Navigation Assistant with Semantic Features in a Large-Scale Environment.” Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS ’17), Association for Computing Machinery, 2017, pp. 270–279.

Schwartz, Broderick S., Seth King, and Tyler Bell. “EchoSee: An Assistive Mobile Application for Real-Time 3D Environment Reconstruction and Sonification Supporting Enhanced Navigation for People with Vision Impairments.” Bioengineering, vol. 11, no. 8, 2024, article 831. https://doi.org/10.3390/bioengineering11080831.

Xie, Jingyi, Rui Yu, He Zhang, Syed Masum Billah, Sooyeon Lee, and John M. Carroll. “Beyond Visual Perception: Insights from Smartphone Interaction of Visually Impaired Users with Large Multimodal Models.” Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2025, article 62. https://doi.org/10.1145/3706598.3714210.

Zhang, He, Nicholas J. Falletta, Jingyi Xie, Rui Yu, Sooyeon Lee, Syed Masum Billah, and John M. Carroll. “Enhancing the Travel Experience for People with Visual Impairments through Multimodal Interaction: NaviGPT, a Real-Time AI-Driven Mobile Navigation System.” Companion Proceedings of the 2025 ACM International Conference on Supporting Group Work, Association for Computing Machinery, 2025, pp. 29–35. https://doi.org/10.1145/3688828.3699636.

Krabbe Disease: The Current Understanding of Pathology and Therapeutic Approaches
By Raymond Fadous

Abstract

Krabbe disease is a rare lysosomal storage disorder characterized by galactocerebrosidase deficiency and progressive neurological deterioration. This review outlines the clinical presentation and epidemiology of Krabbe disease, before examining its pathology. This includes galactocerebrosidase structure, trafficking, and function, alongside disruptions of sphingolipid and myelin homeostasis due to pathogenic galactocerebrosidase variants. Therapeutic strategies are then evaluated according to their intervention in krabbe disease pathology, and their mechanistic, translational, and clinical strengths and limitations. Presently, hematopoietic stem cell therapy is the principal clinically established intervention, although its efficacy is constrained by treatment timing and incomplete correction of pathology. Current evidence suggests that emerging therapies show varying potential to address distinct components of disease pathogenesis. However, challenges involving treatment delivery, toxicity, variant-specific applicability, and clinical validation continue to restrict translation. Collectively, the reviewed evidence suggests that no single therapeutic modality fully addresses the multifactorial pathology of Krabbe disease. Multimodal strategies combining mechanistically synergistic therapies may address the breadth of Krabbe disease pathogenesis while also compensating for the limitations of individual interventions. Future progress will likely depend on earlier disease detection and clinical research on rational combinations capable of addressing the primary enzymatic defect and its downstream neurological consequences.

Introduction

Krabbe disease (KD) is a rare genetically inherited disorder caused by deficiency of the enzyme galactocerebrosidase (GalC). Without sufficient GalC activity, sphingolipid metabolism is disrupted, leading to the accumulation of toxic metabolites and the progressive destruction of myelin in the nervous system, known as demyelination (Feltri et al.). Demyelination results in impaired neuronal transmission, which gives rise to spasticity, seizures, loss of visual fixation, gait disorder, and more (Krieg et al.).

There are three onset subtypes for KD patients: infantile-onset, juvenile-onset and a rarer adult-onset. The infantile subtype, accounting for 85% to 90% of patients, is most severe, with a mortality rate as high as 90% in the first two years of life (Jain and De Jesus). Later-onset forms progress more slowly, with milder symptoms and longer survival than infantile KD.

KD is inherited in an autosomal recessive manner, meaning that affected individuals must inherit a pathogenic GalC allele from both parents. KD demonstrates a genotype-phenotype correlation, in which the functional severity of the two inherited GalC alleles influences the onset and rate of disease progression. Infantile KD results from alleles containing GalC deletion variants and/or severe pathogenic variants that cause a severe loss of GalC function (GeneReviews). Later-onset forms are more frequently associated with at least one GalC allele

that preserves partial enzymatic expression. This is because a hypomorphic allele (one that retains partial GalC expression) could delay the pathological threshold for demyelination. Nevertheless, genotype is not fully predictive of phenotype, as patients carrying similar GalC variants may differ substantially in clinical presentation. Thus, GalC genotype can support broad classification into infantile or later-onset KD but cannot independently provide a definitive prognosis.

A disease is conventionally defined as rare when the number of patients is <1 in 2,000 in the European Union and affecting less than 200,000 people in the United States (Danese and Lippi)(FDA). KD is defined as a rare disease, with an incidence typically cited at approximately 1 per 100,000 live births across Europe and 0.4 in 100,000 in the United States (Barczykowski et al.). Outside of these two regions, there is limited epidemiological data, except for similar incidence rates reported in Japan and Australia (Figure 1).

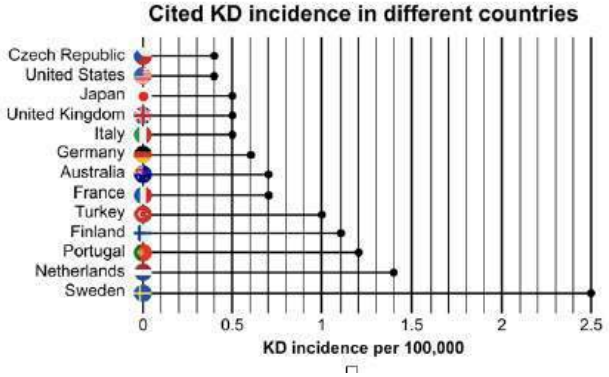


Figure 1. The incidence rates per 100,000 for KD in 13 different countries with epidemiological data on KD. The mean incidence across the 13 countries is ~0.88 per 100,000

Due to its autosomal recessive nature, KD is a major burden within endogamous populations, where individuals are more likely to share ancestry and therefore share pathogenic GalC variants. Among the Druze community in northern Israel, incidence has been reported at ~600 per 100,000 live births (Zlotogora et al.). That is over 650 times greater than the average KD incidence in Figure 1.

The development of effective treatments for KD relies upon a robust understanding of pathological mechanisms, and the successful translation of preclinical findings to clinical intervention. Despite its devastating clinical consequences, KD faces constrained research efforts due to its rarity. Limited patient accrual complicates the design and execution of clinical trials, slowing down the translation of preclinical findings to clinical intervention. Moreover, the rarity of KD delays diagnosis, narrowing the therapeutic window for crucial early intervention (Bascou et al.). The following section explores the pathology of KD, providing a critical foundation for evaluating current and future therapeutic approaches.

Pathology

Myelin and demyelination

Myelin is a protective substance made of sphingolipids and specialized proteins. Sphingolipids like galactocerebroside are a major component in the myelin sheath that act as an electrical insulator for neurons, allowing rapid saltatory conduction of action potentials. The myelin sheath surrounds the axon of neurons, at non-continuous, regularly spaced intervals termed nodes of Ranvier (Figure 2). Myelin electrically insulates the axon, allowing action potentials to "jump" between the nodes of Ranvier.

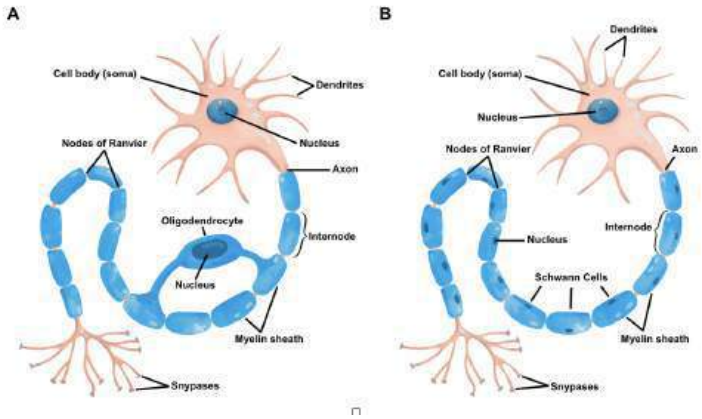


Figure 2. Myelinated neuron diagrams (A) of the central nervous system and (B) of the peripheral nervous system

Neuron myelin is formed by specialized glial cells: oligodendrocytes, in the central nervous system (CNS), and Schwann cells in the peripheral nervous system (PNS). In the CNS (Figure 2A), the neuron is myelinated by the segmented, multilamellar extension of the oligodendrocyte membrane, forming compact internodes (Simons and Nave). In the PNS (Figure 2B), each Schwann cell forms a single internode. Schwann cells have a prominent outer cytoplasmic layer (neurilemma) and are surrounded by a basal lamina.

GalC is a lysosomal hydrolase enzyme that catabolizes sphingolipids, primarily galactocerebroside and psychosine. Galactocerebroside is a major constituent of the myelin sheath (Kohlschutter). Psychosine, on the other hand, is a cytotoxic metabolite that normally exists only at very low concentrations because it is rapidly degraded by GalC (Jain & De Jesus). In the lipid homeostasis of myelin, sections of the membrane invaginate to form vesicles that are rich in galactocerebroside, which require continuous remodelling due to oxidative damage. These galactocerebroside-rich vesicles fuse with early endosomes, which mature into late endosomes before fusing with lysosomes to undergo lipid metabolism.

Depending on the mutation, GalC may reach the lysosome but exhibit little catalytic activity, fail to traffic correctly to the lysosome, or be degraded before reaching its destination. The functional absence of GalC disrupts the lipid homeostasis of myelin, as galactocerebroside cannot be catabolized. Additionally, it leads to psychosine accumulation, causing glial cell death and demyelination due to cytotoxicity. Demyelination disrupts the normal distribution of voltage-gated ion channels along the axon, impairing efficient saltatory conduction (Grider et al.). For this reason, deficient GalC activity results in the symptoms of KD.

Lysosomal trafficking

GalC is one of the 50 known hydrolytic enzymes that participate in the turnover of macromolecules in lysosomes (Saftig and Klumperman). Following synthesis in the ribosomes of the rough endoplasmic reticulum, GalC undergoes a series of post-translational modifications that allows it to be trafficked to their destination in the lysosome (Alroy and Lyons). GalC contains an N-terminal signal peptide that directs it into the lumen of the rough endoplasmic reticulum, where it undergoes N-linked glycosylation of specific asparagine residues. After the signal peptide is cleaved, GalC is transported to the lumen of the Golgi apparatus to acquire a terminal mannose-6-phosphate (M6P), which serves as a recognition marker for the M6P receptor. Once M6P is recognized, GalC is sorted into a transport vesicle to be transferred to an endosomal/lysosomal pathway, maintaining its position on the luminal side until arriving at its destination in the lumen of the lysosome (Figure 3).

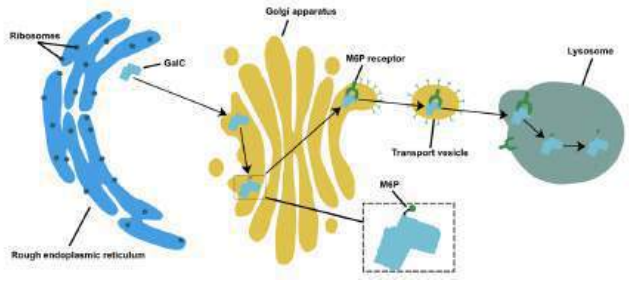


Figure 3. Schematic representation of GalC trafficking from ribosomal synthesis to its final destination in the lysosome via the M6P pathway

GalC structure

The crystal structure of GalC is composed of 668 amino acids (Deane et al.). Three domains are created after folding: residues 25-40 and 338-452 form the β-sandwich domain, residues 41-337 form the triose-phosphate isomerase (TIM) barrel domain, and residues 472-668 form the lectin domain (Figure 4).

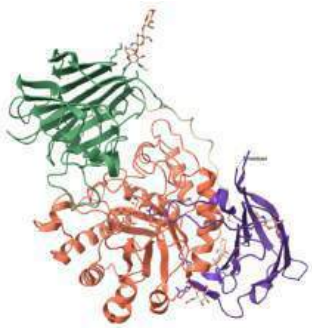


Figure 4. Mouse GalC ribbon structure. Lectin domain colored in green, TIM barrel domain colored in orange, and β-sandwich in purple. Molecular visualization of GalC generated using UCSF ChimeraX (Pettersen et al.). GalC obtained from RCSB protein data bank code 3ZR6.

Mutations of the GalC gene lead to the changes in GalC structure, which hinders its activity in the nervous system. The majority of pathogenic alleles are missense mutations, which cause the replacement of an amino acid with another amino acid, resulting in misfolding and the impairment of lysosomal trafficking (Hill et al.). These are both detrimental enzymatic changes, as they prevent GalC from carrying out the important catalysis of galactocerebroside and psychosine. Less common are null mutations, which completely eliminate the function of a gene product. In those rare KD cases, there is little or no detectable GalC activity. Patients with null pathogenic alleles generally require GalC-restorative therapies. KD missense mutations can affect GalC amino acid structure in all three enzymatic domains (Table 1).

Table 1. Examples of identified missense mutations associated with KD.

Missense mutation*	Domain affected	Description
P150L (Ullah et al.)	TIM barrel	Alters backbone geometry by replacing a rigid proline with a more flexible leucine. Thus, impairs the activation and deactivation of the protein.
S257F (Nasir et al.)	TIM barrel	Replaces small polar serine with bulky hydrophobic phenylalanine, causing severe misfolding and the early degradation of GalC.
I289V (Xu et al.)	TIM barrel	Substitution between two hydrophobic amino acids, but still reduces GalC activity.
L364R (Nasir et al.)	β-sandwich	Substantial physiochemical change, markedly impairing GalC function.
R380W (Spratley et al.)	β-sandwich	Catalytic dysfunction despite having shown localization in lysosomes and undergoing proteolytic processing. This is due to a disruption in the enzyme-substrate interaction, as R380W sits in the active site.
W410G (Nasir et al.)	β-sandwich	Substitutes key stabilizing residue, reducing enzymatic stability and function.
D528N (Zayed)	Lectin domain	Negatively charged aspartic acid replaced with neutral asparagine. Loss of charge disrupts folding and intermolecular interactions, reducing GalC stability and activity.
I583S (Zayed)	Lectin domain	Introduces polarity in a region that relies on hydrophobic interactions for stability. Thus, impairs protein folding and stability.

*Missense mutation is denoted as: (Amino acid code - Residue number - Mutated amino acid code).

Metabolism of galactocerebroside and psychosine

Galactocerebroside is normally degraded in lysosomes by one of two pathways. The first pathway involves GalC, hydrolyzing galactocerebroside to produce ceramide and galactose. The

second pathway involves acid ceramidase (ACDase), which deacylates galactocerebroside to form psychosine and a fatty acid. In the absence of GalC, the second pathway persists, producing psychosine. GalC is also normally responsible for psychosine's catabolism, preventing cytotoxic accumulation (Figure 5).

GalC hydrolyzes both galactocerebroside and psychosine (Figure 5, pathway (1)); galactocerebroside is hydrolyzed to galactose and ceramide, while psychosine is hydrolyzed to galactose and sphingosine. Psychosine is derived from two possible pathways: deacylation of galactocerebroside by ACDase (Figure 5, pathway (2)), and the glycosylation of sphingosine and galactose by ceramide galactosyltransferase (CGT) (Figure 5, pathway (3)) (Li et al.). Galactocerebroside derives only from the glycosylation pathway, where galactose and ceramide are glycosylated by CGT. In this pathway, both the galactose and ceramide can be obtained after psychosine hydrolysis. Psychosine hydrolysis results in galactose that can be used further downstream, and a sphingosine that can be acylated with a free fatty acid to create ceramide.

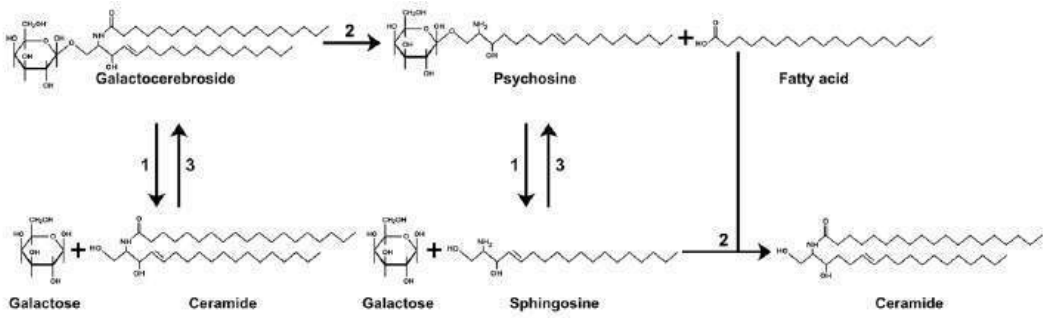


Figure 5. The glycosylation, hydrolysis, and deacylation pathways of galactocerebroside and psychosine.

Psychosine toxicity

Psychosine has a chemistry that enables it to diffuse into the myelin membrane, altering membrane organisation (Figure 6). However, psychosine is normally absent in the myelin membrane, due to GalC activity (Figure 6A). In KD, psychosine accumulates and can exit the acidic lysosome, entering a neutral cytosol. The neutral pH of the cytosol alters psychosine's protonation state, reducing its charge and increasing its lipophilicity, its affinity for non-polar environments. This enhances the ability of psychosine to associate with and disrupt cellular membranes.

Psychosine's role in demyelination involves several interconnected mechanisms. The first is its direct impact on myelin sheaths, where it diffuses into, and alters the integrity of the myelin membrane (Figure 6B). The presence of psychosine in the myelin membrane reduces membrane fluidity resulting in breakage of the myelin sheath, shedding of myelin debris, and demyelination (D'Auria et al.). The second mechanism follows the death of glial cells, as the accumulated psychosine in oligodendrocytes and Schwann cells results in apoptotic cell death. The formation of new myelin halts, causing demyelination.

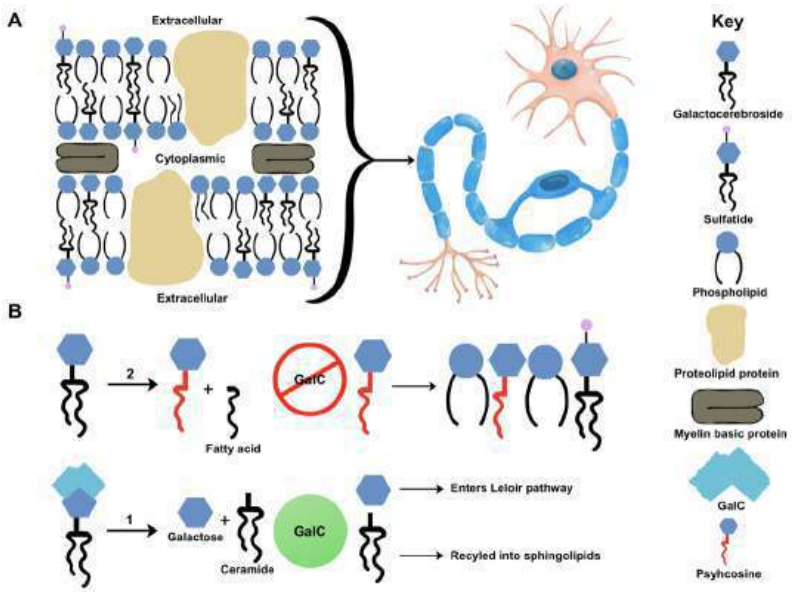


Figure 6. Myelin membrane in presence and absence of GalC.

(A) Normal myelin membrane composition, with phospholipids and sphingolipids (galactocerebroside and sulfatide). (B) The catabolism of galactocerebroside in the presence and absence of GalC. In the absence, psychosine is formed, diffusing into the myelin membrane. In the presence, the galactose and ceramide enter recycling pathways.

Secondary inflammation follows the initial two mechanisms, whereby psychosine causes oligodendrocyte damage, which triggers the activation of microglia and inflammatory signaling that further exacerbates demyelination (Potter and Petryniak). When activated, microglia release different cytokines, which are small signalling proteins. Pro-inflammatory cytokines activate astrocytes, and a special subset of cytokines called chemokines attract macrophages. Activated astrocytes also release chemokines, and pro-inflammatory cytokines that further stimulate the microglia, creating a positive feedback loop. Both the activated microglia and the recruited macrophages phagocytose myelin debris, becoming lipid-laden. These lipid-laden cells can fuse to form a multinucleated globoid cell. It is hypothesized that the globoid cell is also involved in demyelination, enhancing the process.

There are two hypotheses that describe the cascade of these demyelinating events: the psychosine hypothesis and the microglial hypothesis (Figure 7). The psychosine hypothesis proposes that psychosine first disrupts the myelin sheath and kills the myelin-forming cells, so myelin debris is the signal that results in inflammation (Figure 7A) (Nicaise et al.). The microglial hypothesis proposes that the presence of psychosine activates the inflammatory process, triggering the formation of globoid cells which, later, causes demyelination (Figure 7B).

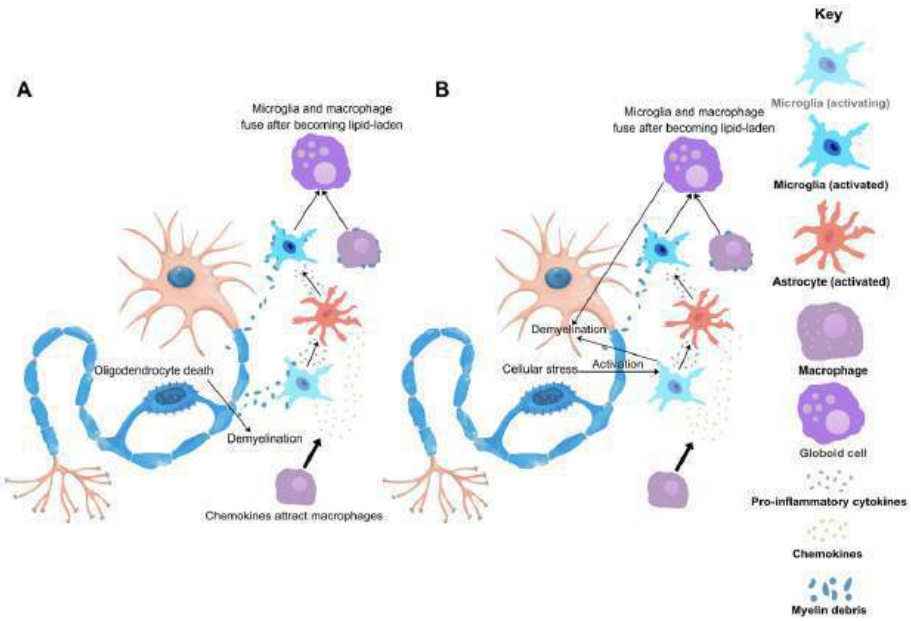


Figure 7. Schematic of the two hypotheses of psychosine toxicity events.
Depicted are the psychosine hypothesis (A) and the microglial hypothesis (B).

The psychosine hypothesis has the greater body of evidence, as studies have shown that psychosine alone is sufficient to induce oligodendrocyte dysfunction, and eventual apoptosis (Won et al.). Evidence also suggests that microglial activation occurs early in disease progression and may actively contribute to pathology rather than simply responding to myelin damage. This would deem the hypotheses complementary rather than mutually exclusive, with microglial activation amplifying demyelination.

After determining the correct pathogenic mechanism, the design of therapeutic strategies can target the greatest pathological burden at the metabolic level. If psychosine accumulation is established as the sole initiating event, therapies that restore GalC activity or reduce psychosine synthesis are more likely to modify disease progression. Still, recognizing that microglial activation amplifies pathology supports the combination of anti-inflammatory therapies with psychosine-targeted ones.

Therapy

The therapeutic management of KD involves targeting one or more of six pathological stages (Figure 8). These studied KD therapies are each driven by their own biological mechanism, ultimately aiming to increase GalC activity, decrease psychosine accumulation, or modulate downstream demyelination and neuroinflammation.

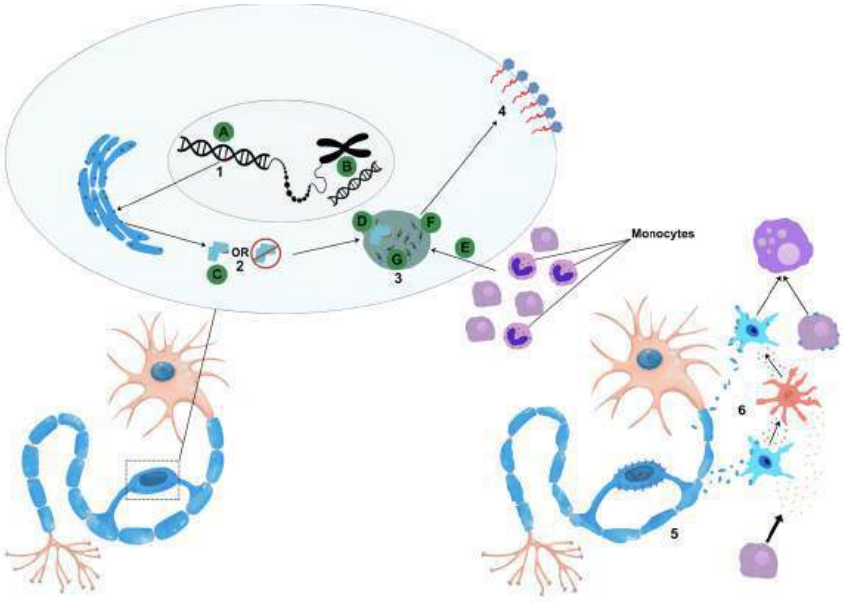


Figure 8. A schematic representation of the pathological progression of KD. (1) The pathogenic GalC gene variant causes (2) defective GalC production/folding/trafficking (3) and reduced lysosomal GalC activity, if any. (4) Psychosine accumulation results from reduced GalC activity, leading to (5) demyelination/neurodegeneration and then (6) secondary inflammation, as per the psychosine hypothesis. Therapeutic intervention points follow the stages of KD pathology, including (A) gene editing, (B) gene therapy, (C) pharmacological chaperone therapy, (D) enzyme replacement therapy, (E) hematopoietic stem cell therapy, (F) substrate reduction therapy, and (G) autophagy activators.

No therapy discussed primarily targets demyelination itself (stage 5); instead, the approaches target upstream causes or downstream inflammatory consequences.

Targeting the first pathological level: Gene-based therapies

Focusing on the root of KD, the mutated GalC gene, is a therapeutic management strategy that aims to restore normal GalC translational synthesis, and thereby slow the demyelinating and neuroinflammatory cascade.

Gene editing

Gene editing aims to fundamentally restore the function of the mutant GalC gene at its original genomic location by correcting the mutated DNA sequence. Preclinical studies on gene editing utilize base editors (BEs), specifically the adenine base editors (ABEs) that mediate an A-T to G-C conversion (Nam et al.). BEs are modified CRISPR systems, a form of gene editing derived from a bacterial immune defense that uses a guide RNA to direct a nucleus to a specific DNA sequence to cut or modify the genome. BEs, however, do not cause harmful chromosomal translocations associated with conventional CRISPR-Cas9 (Liu et al.).

Nam et al. delivered different ABEs to twitcher mice with a G-to-A substitution mutation at codon 355, testing whether the mutant genome could be corrected and how that influenced KD pathology. These ABEs achieved 0.5% correction at the genomic DNA level, subsequently reducing psychosine accumulation by 47% and increasing mean lifespan by ~20% (Nam et al.). Further, the paper reports about 5% correction at the mRNA level and GalC activity restored to ~5% of wild type. This raises the possibility of why even with low DNA editing rates, gene editing can produce measurable biochemical effects and therapeutic benefits.

It must be considered however that BEs are mutation dependent, making human gene editing difficult as BEs must be tailored to the various pathogenic alleles responsible for KD. Perhaps gene editing should first be tailored to the most common variants, once adopted clinically. In future developments, increasing the modularity and editing scope of BEs would solve this issue, expanding the range of correctable GalC mutations and enabling a larger therapeutic platform for KD patients (Rodriguez and Yokota).

Ultimately, gene editing is a potentially powerful therapeutic strategy, once it is able to safely correct the root cause of KD to prevent early-onset pathology while preserving the developing brain. An active research project at University College London is exploring the possibility of correcting the mutant GalC gene before birth to achieve a complete therapeutic effect (University College London).

Gene therapy

The patient genome does not fundamentally require correction if another means of GalC synthesis is introduced. Gene therapy is an experimental strategy that aims to restore functional GalC synthesis and activity by delivering a correct copy of the GalC gene into affected cells via adeno-associated-virus vectors (AAVs). The AAVs transduce CNS cells, delivering the functional DNA episomally (outside of chromosomes, but inside the nucleus) (Wang et al.). The episomal delivery allows expressed GalC to enter the lysosomes. AAVs comprise a broad range of naturally occurring and engineered single-stranded DNA viruses that serve as important platforms for gene delivery and gene therapy applications (Wang et al.).

Animal studies like *Rafi et al.* studied AAV-mediated gene therapy on mice, finding decreased psychosine levels and prolonged survival (Rafi et al.). Since, AAV-mediated gene therapy has entered clinical investigation, but has yet to conclusively present prolongation of survival. Forge Biologics had presented promising clinical findings early on in their study, including increased GalC activity, decreased psychosine, and normalized motor development, before the program was discontinued (Forge Biologics).

The preliminary and mixed findings of clinical studies may partly reflect dose limitations; normalized dosing used in mouse studies are much greater than can be used in clinical studies because of adverse effects like immunotoxicity associated with high dosing of AAVs (Metovic et al.). Indeed. High-dose intrathecal AAV treatment produced dramatically better outcomes than lower-dose treatment in canine KD. AAV vectors transduce vulnerable tissues like liver tissue, in which they can express at the same rates as the targeted tissue and cause the adverse effects (Li

et al.). The future of gene therapy would benefit from suppressing gene expression of AAV vectors in vulnerable off-target tissues. For example, the efficacy of low dose gene therapy could be improved through transient immunosuppression and less immunogenic capsid engineering.

Targeting the second pathological level: Pharmacological chaperone therapy

If the mutated GalC gene is not targeted, the mutant GalC enzyme synthesized or its consequences must be targeted instead to achieve effective therapeutic intervention. Pharmacological Chaperone therapy (PCT) aims to 'fix' mutant GalC during synthesis by correcting the defective GalC folding and lysosomal trafficking using small molecules called chaperones. These chaperones are active-site inhibitors, binding to the GalC in the endoplasmic reticulum to stabilize its native conformation (Hill et al.). Once in the lysosome, chaperones are outcompeted by GalC-substrates due to the pH change, restoring GalC mediated metabolism. Various chaperones have been studied preclinically, demonstrating the ability to restore some residual GalC activity in human cells (Berardi et al.).

Precision is both PCT's greatest strength and greatest limitation. It can restore mutant GalC structure and function, but not universally. PCT is mutation-specific, as GalC mutations alter the enzymatic structure differently. This limits its therapeutic applicability. For example, patients with full null mutations that result in complete catalytic loss are unlikely to benefit from PCT. This restricts its therapeutic value to a genetically defined subset of KD, like missense mutations that result in partial misfolding and trafficking defects. Further, the balance between stabilization and inhibition remains a major therapeutic challenge: dosing must maximize GalC correction while minimizing residual inhibition. Future developments should focus on discovering chaperones with improved specificity to the active site and compatibility with a broader range of mutations.

Targeting the third pathological level: Restoring GalC's presence

Rather than correcting the underlying mutation or rescuing GalC synthesis, third-stage-therapies intervene by introducing functional agents to restore lysosomal sphingolipid metabolism. These therapies aim to prevent the accumulation of psychosine and the ensuing pathological cascade.

Enzyme replacement therapy

Enzyme replacement therapy (ERT) addresses GalC deficits by supplementing recombinant GalC directly, aiming to normalize the degradation of galactocerebroside and psychosine. This recombinant GalC, however, possesses difficulties in crossing the blood-brain barrier (BBB), because GalC is a large and hydrophilic enzyme.
For this reason, the method of injection is an important focus in experiments, guiding recombinant GalC's presence across the nervous system, and its efficacy in restoring GalC activity. Preclinical studies have injected functional GalC directly to the CNS or outside of the CNS (peripherally). CNS injections simply bypass the BBB, resulting in greater GalC delivery to

the brain and survival (Table 2). However, the CNS injections require repeated invasive procedures that carry the risk of introducing infection, among other adverse effects (Slavc et al.). Peripheral injections are considered clinically safer than CNS injection, and more practical for repeated ERT dosing.

Table 2. Injection methods studied in ERT for KD

Injection	**Location**	**Study**	**Key findings**
Intraperitoneal (Peripheral)	The abdominal cavity	Del Grosso et al.	Brain GalC restored to ~42% of wild-type activity
Intravenous (Peripheral)	The bloodstream	Matthes et al.	14-16% reduction of psychosine in the PNS and CNS No amelioration of lifespan
Intracerebroventricular (CNS)	The brain ventricles	Lee et al.	16.5% reduction of psychosine Treated mice survived up to 51 days compared to the control survival of day 40/42
Intrathecal (CNS)	The cerebrospinal fluid	Qin et al.	Significant reduction of psychosine in the brain Prolongation of median lifespan from ~38 days for untreated mice, to 50 days for treated mice

Future strategies are increasingly focused on combining the advantages of both injections; *Del Grosso et al.* have developed a BBB-shuttling nanovector that enables the delivery of recombinant GalC into the CNS without direct injection. Further, they have shown that recombinant GalC can be dose-optimized and successfully incorporated into these nanovectors (Del Grosso et al.).

It must be considered as ERT develops that the production of recombinant GalC dictates the success of ERT. GalC's low natural abundance, complex structure, and susceptibility to denaturation make recombinant GalC production difficult (Del Grosso et al.). Fortunately, there are human-derived cells called HEK cells capable of improving recombinant human GalC yield. HEK cells also have the ability to perform post-translational modifications, beneficial for GalC glycosylation and M6P tagging for lysosomal targeting (Thomas and Smart).

Hematopoietic stem cell transplantation

Hematopoietic stem cell transplantation (HSCT) replaces the patient's hematopoietic system with donor-derived hematopoietic stem cells that differentiate into macrophages and microglia-like cells. Once these donor cells enter the CNS, they are capable of providing two pathological benefits. They are able to partially contribute GalC to surrounding tissue through cross-correction, and provide an immunomodulatory effect by ameliorating secondary inflammation and eliminating globoid cell formation (Aisenberg et al.). In other words, HSCT

acts across both the third and sixth pathological levels of KD, accounting for the cause of demyelination and the consequence of demyelination.

HSCT is the only disease-modifying treatment currently used clinically for eligible patients. *Greco et al.* evaluated the outcomes of HSCT for 41 children with early infantile KD (0-12 months), finding that symptomatic patients presented a median survival of 5 years while asymptomatic patients presented a median of 15.5 years (Greco et al.). That is a 2.3 and 7.0 times increase in median survival respectively, attributed to the transplanted patients' increased GalC activity and decreased psychosine burden. For late infantile cases (12-36 months of age) and juvenile cases (3-16 years), similar lifespan increases have been shown, with a great improvement of cognitive and language function (Yoon et al.) (Aisenberg et al.). For adult cases, a 41 year old patient studied by *Sharp et al.* and a 42 year old patient studied by *Laule et al.* both underwent HSCT after symptom onset, and showed normalization of GalC activity and improvement in gait and balance (Sharp et al.) (Laule et al.).

Although HSCT prolongs survival and preserves neurological function, it's incredibly time-sensitive. Its therapeutic value operates only before major psychosine accumulation, as the therapy takes time to engraft and establish normal GalC activity. Consequently, clinical outcomes are highly contingent on intervention rapidly after diagnosis. This limitation may be addressed through expanding newborn screening and quicker referral to transplant centers with an established donor-matching framework, in which the donor cells are prepared and administered to patients in a timely manner.

However, HSCT's proven clinical benefit simply validates GalC restoration as a viable clinical therapeutic strategy. HSCT targets two stages of KD pathology, but it influences each stage at a lesser efficacy than therapies that solely target that level. More direct GalC-restorative therapies like gene therapy, gene editing, and ERT aim to achieve the correction of GalC activity earlier in KD pathology. Moreover, therapies that target the sixth pathological level account for the further demyelination caused by secondary inflammation. For this reason, HSCT holds a unique role in future KD therapy: perhaps as the interim between a focused first-level therapy, and a focused later-level adjunct therapy.

Targeting the fourth pathological level: Substrate reduction therapy

Substrate reduction therapy (SRT) is reciprocal to ERT in how it relates to KD pathology. The premise of SRT is to reduce the synthesis of psychosine, preventing accumulation, and thus the cytotoxic effects. This reduction does not involve the restoration of GalC, but is achieved by using small-molecules to inhibit other enzymes involved in the metabolic network regulating galactocerebroside and psychosine levels (Figure 9).

CGT is responsible for the production of both psychosine and galactocerebroside, meaning that it can be inhibited to reduce psychosine synthesis directly and indirectly (by reducing galactocerebroside synthesis and subsequent psychosine synthesis) (Figure 9A). ACDase can also be inhibited to prevent psychosine synthesis directly, as the enzyme is responsible for the production of psychosine from galactocerebroside (Figure 9B).

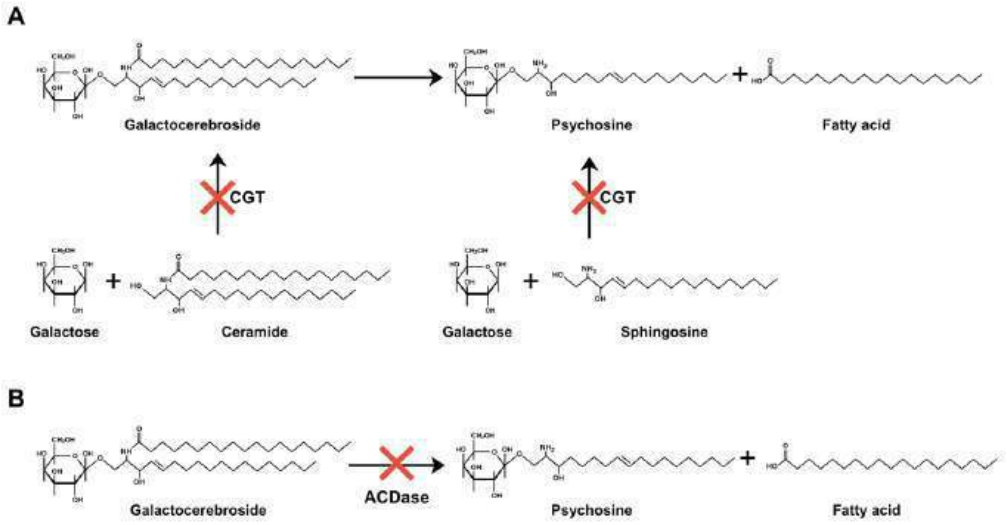


Figure 9. Prevention of psychosine synthesis via SRT. (A) Ceramide galactosyltransferase (CGT) inhibition reduces galactocerebroside and psychosine production, and (B) Acid ceramidase (ACDase) inhibition also reduces psychosine production.

Babcock et al. have developed a potent and selective small molecule inhibitor called S202, able to dose-dependently inhibit CGT. In their preclinical study, S202 reduced galactocerebroside and psychosine levels in both the CNS and PNS, with lower doses prolonging survival by ~50% (Babcock et al.). Interestingly, higher doses of S202 were less effective, with the highest dose tested reducing lifespan by ~16%. This may suggest that chronic CGT inhibition adversely affects healthy tissues, necessitating careful dose optimization.

To inhibit ACDase, *Li et al.* tested the small-molecule drug Carmofur, demonstrating that Carmofur inhibition could prolong the survival of twitcher mice by ~16%, and even decrease psychosine accumulation in the cells of a KD patient (Li et al.). Unfortunately, Carmofur was only effective in twitcher mice that were also heterozygous for ACDase deficiency, which implies that the drug is not sufficient to inhibit the full complement of ACDase activity. While this diminishes Carmofur's therapeutic value, it prompts examination into optimizing both SRT drug choice and dosage based on the genetic background of every patient.

L-cycloserine is a broader SRT agent that inhibits serine palmitoyltransferase (SPT). SPT catalyzes sphingolipid synthesis; inhibiting SPT reduces downstream production of ceramide-derived lipids like psychosine and galactocerebroside. Still, L-cycloserine is not approved for human use, with elevated levels posing toxic effects due to its potency. D-cycloserine is an FDA approved drug that is being considered for repurposing as a less potent alternative to L-cycloserine for KD SRT. Thus, D-cycloserine may achieve a safe degree of SPT inhibition, increasing the likelihood for human studies (Levine and Tsau). While SPT inhibition reduces multiple pathogenic sphingolipid substrates at once, its broad action may compromise sphingolipid homeostasis, and inevitably generate adverse effects.

What the three SRT agents share is their dose-dependent success, each with a threshold at which psychosine is reduced enough to suppress demyelination, without jeopardizing the

sphingolipid homeostasis required for myelin maintenance. Ultimately, SRT cannot correct the primary enzymatic defect in KD, meaning that its value in future treatment of KD is likely as an adjunct therapy.

Targeting the sixth pathological level: preventing further KD progression

Therapies targeting the sixth pathological level prevent further demyelination, modifying downstream cellular pathology to prevent KD worsening. They are adjunctive to therapies that aim to address the progression of KD, rather than fixing GalC defects and its direct consequences.

Autophagy Activators

Autophagy is the natural cellular recycling process that delivers cellular material, like sphingolipids, to the lysosome to be degraded and reused. The impairment of autophagy is characteristic of KD, which creates cellular stress, releasing various inflammatory stimuli that cause microglial activation and secondary inflammation (Papini et al.). Autophagy activators are small-molecule pharmacological agents that enhance autophagy, and in turn, improve the myelinating glial cell's ability to tolerate GalC dysfunction and prevent further demyelination. Currently, there are three autophagy activators studied pre-clinically: Lithium, Resveratrol, and Rapamycin.

Lithium functions by inhibiting an enzyme called inositol monophosphate, promoting autophagy (Sarkar et al.). For KD, however, Lithium's function as an autophagy activator has not been found *in vivo*. Rather, it can partially normalize sciatic nerve-cell signaling (Del Grosso et al.). Resveratrol, is a naturally occurring compound found in foods like grapes and berries, thought to work as an autophagy activator through cellular stress-response pathways. Its usefulness in KD however, seems to be limited, as it cannot restore autophagy marker levels (proteins that indicate autophagy) (Del Grosso et al.).

Rapamycin directly targets one of the main pathways that suppresses autophagy, which makes it the most-supported autophagy activator in KD models. Findings from *Del Grosso et al.* demonstrate the ability to partially restore autophagy markers in the twitcher mouse model, modulating disease progression (Del Grosso et al.). Now, its existing clinical formulations could facilitate future translational investigation.

Targeting multiple pathological levels: Combination therapies

The wide pathological spread of KD may be targeted by administering two or more therapies at once, with each targeting a different level of KD to prevent its cause and effect. The HSCT-gene therapy combination can address two major pathological mechanisms of KD more effectively than either alone; HSCT improves the inflammatory environment, while gene therapy directly boosts GalC activity. The HSCT and intravenous AAV-mediated gene therapy combination extended canine survival beyond one year, compared to approximately 16 weeks for HSCT alone; at the higher AAV dose, neurological function remained normalized to the

pre-determined two year endpoint (Bradbury et al.). Importantly, this canine model is more translationally relevant than the mouse model, providing greater insight into the efficiency of the HSCT/gene therapy combinations in a clinical setting.

The HSCT-ERT combination applies the same logic, through direct enzyme supplementation, without the complexity of viral gene delivery. *Qin et al.* tested the combination in Twitcher mice, resulting in an increase of median survival by nearly 50%, and reduced secondary inflammatory pathology (Qin et al.). While the findings suggest that this combination can be as effective as the HSCT/gene therapy combination, it lacks a consistent and prolonged supplement of GalC. Indeed, the study found that at 36 days, GalC was undetectable in the brains of treated Twitcher mice.

Unlike either of the above combinations, the HSCT-SRT combination does not restore GalC directly. Rather, it aims to reduce psychosine's accumulation and secondary neuroinflammation, slowing KD progression. *Biswas and Le Vine* combined HSCT with L-cycloserine for Twitcher mice, improving survival more than either therapy alone (Biswas and Le Vine). This suggests that SRT can enhance the value of HSCT in Twitcher mice. However, that is the limit of this combination, as it does not fully address the core GalC defect.
A multimodal therapeutic approach may overcome the limits of these dual combinations, aiming to target levels one, four, and six. *Hawkins-Salsbury et al.* studied this multimodal approach in Twitcher mice with KD, simultaneously treating the mice with gene therapy, SRT (L-cycloserine), and HSCT. The combination boosted GalC activity, reduced psychosine accumulation, and decreased neuroinflammation. Indeed, they obtained profound synergistic effects, resulting in an unprecedented increase in median life span from 39.5 days in untreated mice to 298.5 days in triple-treated mice (Hawkins-Salsbury et al.). The multimodal approach utilizing L-cycloserine has its side effects, as the mice were sensitive to the SRT in the weeks following gene therapy and HSCT. This was attributed to the alteration of sphingolipid homeostasis that L-cycloserine causes. More selective inhibitors such as S202 may reduce this limitation, although this remains to be established experimentally in combination regimens. Based on the study's findings, it could be argued that SRT is indeed better as an adjunctive therapy, but should be carried out via a more selective drug like S202 as L-cycloserine's adverse effects are bolstered in the presence of other sphingolipid-modifying therapies like HSCT and ERT. In other words, therapies in combination should maintain a specific role, targeting one of the pathological levels without influencing sphingolipid homeostasis.

Every therapy discussed targets a different point of KD pathology, each experiencing unique research developments to mitigate its clinical weakness while maximizing its metabolic strength (Table 3.)

Table 3. Summary of research and developmental extent of therapies discussed above

Therapy	**Primary target**	**Evidence stage**	**Main strength**	**Main limitation**
Gene editing	GalC mutation	Preclinical	Corrects root mutation	Mutation-specific
Gene therapy	GalC activity	Preclinical, with clinical investigation	Broad GalC restoration	Delivery/toxicity
PCT	GalC folding	Preclinical	Saves misfolded endogenous GalC	Mutation-specific
ERT	GalC deficiency	Preclinical	Direct GalC replacement	Bypassing BBB/duration
SRT	Psychosine production	Preclinical	Reduces pathological substrate burden	Metabolic disruption
HSCT	GalC deficiency + Secondary inflammation	Clinical	Established clinical evidence	Timing/partial correction of targets
Autophagy activators	Downstream pathology	Preclinical	Adjunct potential	No GalC restoration

After all, KD represents a complex, multi-level pathology, making it difficult to address in its entirety. Continued development of early, multimodal therapeutics that specifically target the first levels and its subsequent levels adjunctly offers the greatest potential in combating its multifaceted pathology. Given the rarity of KD and the challenges this creates for patient identification and clinical investigation, broader collaboration and improved screening efforts will be integral in the translation of therapeutic advancements into patient survival.

Acknowledgments

I attest that the ideas, graphics, and writing in this paper are entirely my own. I would like to thank Dr. Victoria Junghans for her guidance and mentorship throughout the development of this review. Her expertise and thoughtful suggestions were integral in refining the review's scope and direction.

Works Cited

Aisenberg, William H., et al. "Direct Microglia Replacement Reveals Pathologic and Therapeutic Contributions of Brain Macrophages to a Monogenic Neurological Disease." Immunity, vol. 58, no. 5, May 2025, pp. 1254–68.e9, https://doi.org/10.1016/j.immuni.2025.03.019.

Alroy, Joseph, and Jeremiah A. Lyons. "Lysosomal Storage Diseases." Journal of Inborn Errors of Metabolism and Screening, vol. 2, Mar. 2014, p. 232640981351766, https://doi.org/10.1177/2326409813517663.

Babcock, Michael C., et al. "Substrate Reduction Therapy for Krabbe Disease and Metachromatic Leukodystrophy Using a Novel Ceramide Galactosyltransferase Inhibitor." Scientific Reports, vol. 11, no. 1, July 2021, https://doi.org/10.1038/s41598-021-93601-1.

Bascou, Nicholas, et al. "A Prospective Natural History Study of Krabbe Disease in a Patient Cohort with Onset between 6 Months and 3 Years of Life." Nih.Gov, 9 Aug. 2018, https://pmc.ncbi.nlm.nih.gov/articles/PMC6083585/.

Berardi, Anna Sara, et al. "Pharmacological Chaperones Increase Residual B-Galactocerebrosidase Activity in Fibroblasts From Krabbe Patients." Molecular Genetics and Metabolism, vol. 112, no. 4, Aug. 2014, pp. 294–301, https://doi.org/10.1016/j.ymgme.2014.05.009.

Biswas, Sangita, and Steven M Le Vine. "Substrate-Reduction Therapy Enhances the Benefits of Bone Marrow Transplantation in Young Mice With Globoid Cell Leukodystrophy." Pediatric Research, vol. 51, no. 1, Jan. 2002, pp. 40–47, https://doi.org/10.1203/00006450-200201000-00009.

Bradbury, Allison M, et al. "Combination HSCT and Intravenous AAV-Mediated Gene Therapy in a Canine Model Proves Pivotal for Translation of Krabbe Disease Therapy." Molecular Therapy : The Journal of the American Society of Gene Therapy [United States], vol. 32, no. 1, Jan. 2024, PubMed Central, pp. 44–58, https://doi.org/10.1016/j.ymthe.2023.11.014.

Danese, Elisa, and Giuseppe Lippi. "Rare Diseases: The Paradox of an Emerging Challenge." Annals of Translational Medicine, vol. 6, no. 17, Sept. 2018, https://doi.org/10.21037/atm.2018.09.04.

D'Auria, Ludovic, et al. "Psychosine Enhances the Shedding of Membrane Microvesicles: Implications in Demyelination in Krabbe's Disease." PloS One [United States], vol. 12, no. 5, May 2017, PubMed, p. e0178103, https://doi.org/10.1371/journal.pone.0178103.

Del Grosso, Ambra, Gabriele Parlanti, et al. "Chronic Lithium Administration in a Mouse Model for Krabbe Disease." JIMD Reports [United States], vol. 63, no. 1, Nov. 2021, PubMed, pp. 50–65, https://doi.org/10.1002/jmd2.12258.

Del Grosso, Ambra, Lucia Angella, et al. "Dysregulated Autophagy as a New Aspect of the Molecular Pathogenesis of Krabbe Disease." Neurobiology of Disease [United States], vol. 129, Sept. 2019, PubMed, pp. 195–207, https://doi.org/10.1016/j.nbd.2019.05.011.

Del Grosso, Ambra, Marianna Galliani, et al. “Brain-Targeted Enzyme-Loaded Nanoparticles: A Breach Through the Blood-Brain Barrier for Enzyme Replacement Therapy in Krabbe Disease.” Science Advances [United States], vol. 5, no. 11, Nov. 2019, PubMed Central, p. eaax7462, https://doi.org/10.1126/sciadv.aax7462.

Del Grosso, Ambra, Sara Carpi, Laura Colagiorgio, et al. “Investigating the Cellular Effects of GALC Dosing in Enzyme Replacement Therapy for Krabbe Disease Supports the Role of Nanomedicine.” Advanced Biology [Germany], vol. 9, no. 9, Sept. 2025, PubMed Central, p. e00147, https://doi.org/10.1002/adbi.202500147.

Del Grosso, Ambra, Sara Carpi, Miriam De Sarlo, et al. “Chronic Rapamycin Administration via Drinking Water Mitigates the Pathological Phenotype in a Krabbe Disease Mouse Model Through Autophagy Activation.” Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie [France], vol. 173, Apr. 2024, PubMed, p. 116351, https://doi.org/10.1016/j.biopha.2024.116351.

FDA. “Rare Diseases at FDA.” FDA, 21 Nov. 2024, https://www.fda.gov/patients/rare-diseases-fda.

Feltri, M. Laura, et al. “Mechanisms of Demyelination and Neurodegeneration in Globoid Cell Leukodystrophy.” Glia, vol. 69, no. 10, Apr. 2021, pp. 2309–31, https://doi.org/10.1002/glia.24008.

Forge Biologics. “Forge Biologics.” Forge Biologics, 29 Jan. 2024, https://www.forgebiologics.com/forge-biologics-announces-positive-fbx-101-clinical-trial-update-in-patients-with-krabbe-disease-identified-by-newborn-screening-ahead-of-rusp-vote/.

Greco, Melissa R., et al. “Infantile Krabbe Disease (0–12 Months), Progression, and Recommended Endpoints for Clinical Trials.” Annals of Clinical and Translational Neurology, vol. 11, no. 12, Nov. 2024, pp. 3064–80, https://doi.org/10.1002/acn3.52114.

Hawkins-Salsbury, Jacqueline A, et al. “Mechanism-Based Combination Treatment Dramatically Increases Therapeutic Efficacy in Murine Globoid Cell Leukodystrophy.” The Journal of Neuroscience : The Official Journal of the Society for Neuroscience [United States], vol. 35, no. 16, Apr. 2015, PubMed Central, pp. 6495–505, https://doi.org/10.1523/JNEUROSCI.4199-14.2015.

Hill, Chris H, et al. “Azasugar Inhibitors as Pharmacological Chaperones for Krabbe Disease.” Chemical Science [England], vol. 6, no. 5, May 2015, PubMed, pp. 3075–86, https://doi.org/10.1039/c5sc00754b.

Hill, Chris H., et al. “Structural Snapshots Illustrate the Catalytic Cycle of B-Galactocerebrosidase, the Defective Enzyme in Krabbe Disease.” Proceedings of the National Academy of Sciences of the United States of America, vol. 110, no. 51, Dec. 2013, pp. 20479–84, https://doi.org/10.1073/pnas.1311990110.

Jain, Megha, and Orlando De Jesus. “Krabbe Disease.” PubMed, StatPearls Publishing, 23 Aug. 2023, https://www.ncbi.nlm.nih.gov/books/NBK562315/.

Kohlschuttër, Alfried. Lysosomal Leukodystrophies: Krabbe Disease and Metachromatic Leukodystrophy - Handbook of Clinical Neurology. Elsevier, 2013, pp. 1611–18.

Krieg, Sarah Isabel, et al. "Natural History of Krabbe Disease – a Nationwide Study in Germany Using Clinical and MRI Data." Orphanet Journal of Rare Diseases, vol. 15, no. 1, Sept. 2020, https://doi.org/10.1186/s13023-020-01489-3.

Laule, Cornelia, et al. "Hematopoietic Stem Cell Transplantation in Late‐Onset Krabbe Disease: No Evidence of Worsening Demyelination and Axonal Loss 4 Years Post‐Allograft." Journal of Neuroimaging, vol. 28, no. 3, Feb. 2018, pp. 252–55, https://doi.org/10.1111/jon.12502.

Lee, Wing C, et al. "Single-Dose Intracerebroventricular Administration of Galactocerebrosidase Improves Survival in a Mouse Model of Globoid Cell Leukodystrophy." FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology [United States], vol. 21, no. 10, Aug. 2007, PubMed, pp. 2520–27, https://doi.org/10.1096/fj.06-6169com.

LeVine, Steven M, and Sheila Tsau. "Substrate Reduction Therapy for Krabbe Disease: Exploring the Repurposing of the Antibiotic D-Cycloserine." Frontiers in Pediatrics [Switzerland], vol. 9, Jan. 2022, PubMed Central, p. 807973, https://doi.org/10.3389/fped.2021.807973.

Li, Yedda, Christopher A Miller, et al. "Enhanced Efficacy and Increased Long-Term Toxicity of CNS-Directed, AAV-Based Combination Therapy for Krabbe Disease." Molecular Therapy, vol. 29, no. 2, Feb. 2021, pp. 691–701, https://doi.org/10.1016/j.ymthe.2020.12.031.

Li, Yedda, Yue Xu, et al. "Genetic Ablation of Acid Ceramidase in Krabbe Disease Confirms the Psychosine Hypothesis and Identifies a New Therapeutic Target." Proceedings of the National Academy of Sciences, vol. 116, no. 40, Sept. 2019, pp. 20097–103, https://doi.org/10.1073/pnas.1912108116.

Liu, Mengzhu, et al. "Global Detection of DNA Repair Outcomes Induced by CRISPR-Cas9." Nucleic Acids Research [England], vol. 49, no. 15, Sept. 2021, PubMed Central, pp. 8732–42, https://doi.org/10.1093/nar/gkab686.

Matthes, Frank, et al. "Enzyme Replacement Therapy of a Novel Humanized Mouse Model of Globoid Cell Leukodystrophy." Experimental Neurology, vol. 271, Sept. 2015, pp. 36–45, https://doi.org/10.1016/j.expneurol.2015.04.020.

Metovic, Jasna, et al. "Gene Therapy for the Leukodystrophies: From Preclinical Animal Studies to Clinical Trials." Neurotherapeutics, vol. 21, no. 4, July 2024, p. e00443, https://doi.org/10.1016/j.neurot.2024.e00443.

Nam, Bae-Geun, et al. "In Vivo Adenine Base Editing of Mutant Galc Gene Ameliorates Krabbe Disease Progression." Genome Medicine [England], vol. 18, no. 1, Apr. 2026, PubMed, p. 46, https://doi.org/10.1186/s13073-026-01620-2.

Nasir, Gibran, et al. "Krabbe Disease: Prospects of Finding a Cure Using AAV Gene Therapy." Frontiers in Medicine, vol. 8, 2021, https://doi.org/10.3389/fmed.2021.760236.

Nicaise, Alexandra M., et al. “A Microglial Hypothesis of Globoid Cell Leukodystrophy Pathology.” Journal of Neuroscience Research, vol. 94, no. 11, Nov. 2016, pp. 1049–61, https://doi.org/10.1002/jnr.23773.

Orsini, Joseph J, et al. “Krabbe Disease.” Ncbibooks, University of Washington, Seattle, 11 Oct. 2018, https://www.ncbi.nlm.nih.gov/books/NBK1238/.

Papini, Nadia, et al. “Impaired Autophagy in Krabbe Disease: The Role of BCL2 and Beclin-1 Phosphorylation.” International Journal of Molecular Sciences [Switzerland], vol. 24, no. 6, Mar. 2023, PubMed Central, p. 5984, https://doi.org/10.3390/ijms24065984.

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE.

Protein Sci. 2021 Jan;30(1):70-82, https://doi.org/10.1002/pro.3943

Potter, Gregory B., and Magdalena A. Petryniak. “Neuroimmune Mechanisms in Krabbe Disease.” Journal of Neuroscience Research, vol. 94, no. 11, Sept. 2016, pp. 1341–48, https://doi.org/10.1002/jnr.23804.

Qin, Elizabeth Y, et al. “Bone Marrow Transplantation Increases Efficacy of Central Nervous System-Directed Enzyme Replacement Therapy in the Murine Model of Globoid Cell Leukodystrophy.” Molecular Genetics and Metabolism [United States], vol. 107, nos. 1–2, Sept. 2012, PubMed Central, pp. 186–96, https://doi.org/10.1016/j.ymgme.2012.05.021.

Rafi, Mohammad A., et al. “Intravenous Injection of AAVrh10-GALC After the Neonatal Period in Twitcher Mice Results in Significant Expression in the Central and Peripheral Nervous Systems and Improvement of Clinical Features.” Program and Abstracts for the 2015 Meeting of the Society for Inherited Metabolic Disorders, vol. 114, no. 3, Mar. 2015, pp. 459–66, https://doi.org/10.1016/j.ymgme.2014.12.300.

Rodriguez, Sebastian Hernandez, and Toshifumi Yokota. “Building CRISPR-Based Gene-Editing Platforms for Personalized Medicine: The Next Step in Interventional Genetics.” Genes [Switzerland], vol. 17, no. 6, May 2026, PubMed, p. 631, https://doi.org/10.3390/genes17060631.

Saftig, Paul, and Judith Klumperman. “Lysosome Biogenesis and Lysosomal Membrane Proteins: Trafficking Meets Function.” Nih.Gov, 12 Aug. 2009, https://pubmed.ncbi.nlm.nih.gov/19672277/.

Sarkar, Sovan, et al. “Lithium Induces Autophagy by Inhibiting Inositol Monophosphatase.” The Journal of Cell Biology [United States], vol. 170, no. 7, Sept. 2005, PubMed Central, pp. 1101–11, https://doi.org/10.1083/jcb.200504035.

Sharp, Madeleine E, et al. “Stem Cell Transplantation for Adult-Onset Krabbe Disease: Report of a Case.” JIMD Reports [United States], vol. 10, 2013, PubMed Central, pp. 57–59, https://doi.org/10.1007/8904_2012_203.

Slavc, Irene, et al. “Best Practices for the Use of Intracerebroventricular Drug Delivery Devices.” Molecular Genetics and Metabolism [United States], vol. 124, no. 3, July 2018, PubMed, pp. 184–88, https://doi.org/10.1016/j.ymgme.2018.05.003.

Spratley, Samantha J., et al. “Molecular Mechanisms of Disease Pathogenesis Differ in Krabbe Disease Variants.” Traffic, vol. 17, no. 8, May 2016, pp. 908–22, https://doi.org/10.1111/tra.12404.

Thomas, Philip, and Trevor G. Smart. “HEK293 Cell Line: A Vehicle for the Expression of Recombinant Proteins.” Journal of Pharmacological and Toxicological Methods, vol. 51, no. 3, May 2005, pp. 187–200, https://doi.org/10.1016/j.vascn.2004.08.014.

Ullah, Ikram, et al. “A Novel Variant of GALC in a Familial Case of Krabbe Disease: Insights From Structural Bioinformatics and Molecular Dynamics Simulation.” Genes & Diseases, vol. 10, no. 6, 2023, pp. 2263–66, https://doi.org/10.1016/j.gendis.2023.01.018.

Wang, Jiang-Hui, et al. “Adeno-Associated Virus as a Delivery Vector for Gene Therapy of Human Diseases.” Signal Transduction and Targeted Therapy, vol. 9, no. 1, Apr. 2024, pp. 1–33, https://doi.org/10.1038/s41392-024-01780-w.

Won, Je-Seong, et al. “Role of Endogenous Psychosine Accumulation in Oligodendrocytes Differentiation and Survival: Implication for Krabbe Disease.” Brain Research, vol. 1508, May 2013, pp. 44–52, https://doi.org/10.1016/j.brainres.2013.02.024.

Xu, Chengzhe, et al. “Six Novel Mutations Detected in the GALC Gene in 17 Japanese Patients With Krabbe Disease, and New Genotype–Phenotype Correlation.” Journal of Human Genetics, vol. 51, no. 6, Apr. 2006, pp. 548–54, https://doi.org/10.1007/s10038-006-0396-3.

Yoon, Isabel C, et al. “Long-Term Neurodevelopmental Outcomes of Hematopoietic Stem Cell Transplantation for Late-Infantile Krabbe Disease.” Blood, vol. 137, no. 13, Nov. 2020, https://doi.org/10.1182/blood.2020005477.

Zayed, Hatem. “Krabbe Disease in the Arab World.” Journal of Pediatric Genetics, vol. 04, no. 01, July 2015, pp. 001–08, https://doi.org/10.1055/s-0035-1554981.

Zlotogora, J, et al. “Krabbe Disease: Increased Incidence in a Highly Inbred Community.” American Journal of Medical Genetics [United States], vol. 21, no. 4, Aug. 1985, PubMed, pp. 765–70, https://doi.org/10.1002/ajmg.1320210420.

How Does Socioeconomic and Geographic Background Affect Women's Financial Independence in India? By Kyra Aziz

Abstract

This paper investigates whether socioeconomic background determines how financially independent women are in India, and more specifically, whether women from lower socioeconomic backgrounds demonstrate greater practical financial independence than is commonly assumed. Drawing on academic literature, government data, and national survey findings, the paper argues that the dominant perception, that urban, educated, higher-income women are more financially independent, is both widespread and institutionally reinforced, but is not consistently supported by evidence of actual financial behaviour. The paper finds that while higher-income women enjoy greater formal financial access, their practical autonomy is frequently constrained by cultural norms, income dependence, and the delegation of financial tasks to male family members or professional advisors. Lower-income and rural women, by contrast, often demonstrate high levels of practical financial task execution out of structural necessity, even in the absence of formal financial literacy.

Keywords

Financial independence, financial autonomy, women, India, socioeconomic background, financial inclusion, household decision-making

About the Author

Kyra Aziz is an International Baccalaureate Diploma student in India. Her interests lie in the fields of economics, psychology and business, particularly in the Indian context. She founded and leads Pragati ki Udan, a youth-led programme in rural Chhattisgarh, that lead to the development of interest in this particular topic. She has a keen want, to learn about different perspectives, having experienced many from different cultures due to moving for education.

Use of AI

AI was used in this research paper for the purposes of checking grammar and tone. It was also used to find synonyms and appropriate jargon for the topic. Moreover, AI was used to cross check all data used and to make sure that there is no misinterpretation of the works cited. In the initial stages of the paper, AI was also used to get a basic idea of the type of sources that should be used, like how efficient each source type is to analyse the topic.

1. Introduction

There is a common assumption in India that urban, educated women are financially independent, while women from rural or lower-income backgrounds are not. This paper does not take that assumption as fact. It argues instead that the assumption can be deduced from the way

three things are designed: government financial inclusion schemes, financial advertising, and industry reporting on women and money.

One survey shows why the assumption is worth testing. In January 2024, DBS Bank India and CRISIL surveyed more than 800 salaried and self-employed women across ten Indian cities. It reported that 98 per cent took part in long-term family financial decisions, that 47 per cent made financial decisions on their own, and, separately, that 47 per cent sought advice and information from family members (DBS Bank India, "47% of Female Indian Earners"). These are different measures, and the second group is not necessarily the same as the first, but together they make a simple point: taking part in a decision is not the same as having authority over it.

The question behind this paper came from something I observed directly. While working with women in rural Chhattisgarh as part of a financial literacy programme, I found that many of them were already doing things often assumed to be out of their reach: going to the bank on their own, taking and repaying loans, and managing a household budget on a small and irregular income. They were not financially literate in the formal sense, and most could not have explained what a mutual fund was. But they were carrying out financial tasks themselves, largely because there was nobody else to do it. This is a personal observation from one programme, not survey evidence, and it is offered as the reason for the question rather than an answer to it. If financial independence is measured by who actually carries out financial tasks, rather than by who holds a bank account or who is in the room when money is discussed, does the expected ranking between richer urban women and poorer rural women still hold?

This paper asks how socioeconomic and geographic background affect women's financial independence in India. Financial independence is used here to mean two things at once: the authority to decide how money is earned, saved and spent, and the capability to carry out financial tasks personally, such as going to the bank, managing a loan or handling savings, rather than passing them to somebody else.

This definition is adapted from Naila Kabeer's framework for measuring women's empowerment, which separates empowerment into resources, agency and achievements and argues that no one of them can be read alone, since the meaning of any indicator depends on the other two (Kabeer). Kabeer's resources correspond here to financial access; her agency is split into decision authority and task execution. The split is needed because existing studies measure different things and describe them all as financial independence: the DBS and CRISIL survey measures presence in a decision, Banerjee et al. measure responsibility and product holdings, NFHS-5 measures discretion over a resource, and only Chauhan holds account ownership and decision-making apart. Section 3.5 sets out what this paper takes from Kabeer and what it changes.

One limitation of scope should be stated at the outset. This paper compares women from lower socioeconomic backgrounds in rural areas with women from middle and higher socioeconomic backgrounds in urban areas. There are poorer women in cities and wealthier women in villages, and this paper does not examine them. The comparison is used because it is

the one built into the assumption being tested, and because most national data is reported along rural and urban lines.

2. Background

India has expanded women's access to formal financial services considerably over the past decade, and the design of that expansion is itself informative. The Pradhan Mantri Jan Dhan Yojana (PMJDY), launched in 2014, is the largest of these programmes: as of August 2023 it had opened approximately 500.9 million accounts, around 56 per cent held by women and around 67 per cent in rural and semi-urban areas (India, Ministry of Finance). The Deendayal Antyodaya Yojana National Rural Livelihoods Mission (DAY-NRLM) works instead through self-help groups (SHGs), and as of June 2025 reported women from 100.5 million rural households mobilised into more than 9.09 million SHGs across 28 states and six union territories, accessing approximately 11 trillion rupees (if written according to the global number system) in loans since 2013-14 (India, Department of Rural Development). The Mahila Samridhi Yojana and the Mahila Kisan Sashaktikaran Pariyojana also target rural and low-income women. These are figures for programme reach. They record how many women hold an account or belong to a group, not how many decide what happens to the money.

The wording of these schemes matters as much as their scale. The Ministry of Finance describes PMJDY as serving "vulnerable groups such as low-income groups and weaker sections that lack access to basic banking services," and as addressing "financial untouchability by bringing the marginalised sections of the society into the ambit of formal banking" (India, Ministry of Finance). This states who is assumed to lack financial access and, by leaving them out, indicates who is assumed to already have it. Urban and higher-income women are not named as a target group anywhere in the scheme's stated purpose. No policy document says that these women are financially independent, and this paper does not claim that one does. What can be deduced is narrower: these programmes are built on an assumption about where the deficit lies, and it places that deficit outside the urban middle class.

Financial advertising reinforces the same association. HDFC Mutual Fund's "Barni Se Azadi" campaign builds its central image around an all-women Govinda pathak in Mumbai who climb a human pyramid to break a savings jar rather than the traditional clay pot, urging women to move from storing money towards investing it ("HDFC Mutual Fund"). Aditya Birla Sun Life Mutual Fund's #ForHer initiative states its aim as increasing financial inclusion among women, pursued through financial-planning podcasts and literacy articles (Aditya Birla Sun Life Mutual Fund). Both address a woman who already has savings, a phone, and a decision to make about where to put her money. Iyer's analysis of 102 award-winning advertisements, 29 of them Indian, finds that Indian femvertising operates through commodity feminism, treating consumption as the route to empowerment, and largely obscures the structural barriers of caste and class that shape women's financial lives (Iyer). The woman these campaigns address is assumed to hold resources already, and the question put to her is what to do with them; the inclusion schemes ask

a different question, of a different woman. This claim is limited: two campaigns and one content analysis cannot establish what Indian financial advertising does as a whole.

National survey evidence complicates the picture these materials imply, and Section 4 examines it in detail. Two figures indicate the direction. Kumar reports that the share of women with a bank or savings account they themselves use rose from 53 per cent in NFHS-4 to 79 per cent in NFHS-5, while only 22.5 per cent of women who own a mobile phone use it for financial transactions (Kumar). Access has expanded quickly; what women can do with that access has not moved at the same speed.

3. Literature Review

The academic literature clusters around three overlapping questions. The first asks what prevents women from accessing and using financial services: Duvendack et al. on the barriers facing users of banking correspondents, Gupta and Bhadouria on rural women's access and awareness, and Hoorunnisa and Sharma on the barriers women themselves report. The second asks what relationship holds between financial access and financial autonomy: Chauhan on the 2014 account expansion and Chowdhury and Mondal on NFHS-5. The third asks how socioeconomic position shapes both: Banerjee et al. on financial literacy and caste, Chatterjee et al. on household income and women's employment, and Twamley and Sidharth on class, caste and respectability. This paper draws on all three to examine an assumption running through the way their findings are commonly read: that higher socioeconomic status straightforwardly predicts greater financial independence.

3.1 Financial access is not financial autonomy

Access to a financial service does not by itself produce control over it, and Chauhan's 2024 working paper offers the clearest Indian test. She examines the 2014 PMJDY expansion using a difference-in-differences design exploiting regional variation in pre-policy bank branch density, comparing 168 high-impact districts with 161 low-impact ones, drawing on the India Human Development Survey, the Demographic and Health Survey and Reserve Bank of India records, with a district panel of 31,268 households. She finds no significant improvement in women's participation in decisions about large household purchases, or in their spending autonomy, in the districts where account ownership rose most (Chauhan 35).

This is not a finding that accounts did nothing. Chauhan reports that households "became more likely to save in formal instruments, such as fixed deposits and government bonds, and to borrow from formal sources like banks and registered companies" (Chauhan 3). Formal financial behaviour changed; who decided did not. She also reports that the gap the policy meant to close was real, with 48 per cent of unbanked women participating in decisions on large household purchases against 55 per cent of banked women (Chauhan 2). Accounts remain worth having, but the sequence the policy assumes, access first and autonomy following, is not what the evidence shows, so the premise that rural women need accounts before they can become financially independent only captures part of the issue. Two limits apply: this is a working paper rather than

a peer-reviewed article, and a statistically insignificant result is not proof that an effect is absent, only that none was detected on the outcomes measured.

3.2 Financial literacy, earnings and social position

If access alone does not shift financial authority, what does? Banerjee et al. address this using India's national benchmark survey on financial literacy and inclusion, conducted by the National Centre for Financial Education in 2015, covering 59,405 respondents across all states and union territories and examining holdings across 18 financial products. They report that "when women are financially literate, the gender difference is reversed, with women now significantly more responsible for making financial decisions in the household, as compared to financially literate men" (Banerjee et al. 13).

Two qualifications matter. First, the gains are largest when a financially literate woman shares responsibility with her husband rather than holding it alone: such women's holdings in investment products and stocks were roughly 88 per cent higher than under sole responsibility. Deciding alone is therefore not straightforwardly the better outcome, which cautions against treating sole decision-making as the definition of independence. Second, the marginal effects of financial literacy decrease steadily moving down the caste hierarchy, from General Caste to Scheduled Tribe (Banerjee et al. 21), with lower-caste groups showing greater adherence to existing gender norms (Banerjee et al. 6). Financial literacy operates inside a social position rather than overriding it.

National earnings data points the same way. Among married women with earnings, only 18 per cent decide alone how their money is spent, while 67 per cent decide jointly (Kumar). Around 65 per cent of women who earn as much as or more than their husbands have command over their own spending, against 59 per cent of those who earn less (Chowdhury and Mondal).

Hoorunnisa and Sharma surveyed women in the National Capital Region and list the most commonly reported barriers as lack of access to formal financial institutions, social norms and expectations, lack of financial literacy, and gender discrimination (Hoorunnisa and Sharma c363-64). Its reliability is limited: the abstract describes a planned sample of 500 while the tables report 105 respondents, and the percentages do not consistently follow from the counts. Its figures are used here only as an indication of which barriers women name.

That study does not say what its social norms consist of, and this paper does not treat the label as self-explanatory. Duvendack et al. supply the content. Studying the banking correspondent model, they report that women's movements are often restricted by socio-cultural norms and caregiving responsibilities, and that in the studies they review 50 per cent of women had to seek permission from a husband or other family member before leaving the home. Female correspondents face safety problems carrying cash between villages, making families reluctant to let them take the work, and around 10 per cent of correspondents nationally are women. Women also have the lowest access to phones in the region (Duvendack et al.). The norms are therefore concrete: rules about where a woman may travel, whom she may transact with, whether she may carry cash, and whether she holds a phone. Duvendack et al. add one finding that bears on this

paper's definition more sharply than any other: women are frequently proxy users of their own accounts, operated in practice by male family members. Such an account appears in inclusion statistics as her financial access. It is not her financial authority.

3.3 Rural access, awareness and what constrains participation

Evidence on rural women supports separating what a woman knows from what she does. Gupta and Bhadouria surveyed 200 rural women across Uttar Pradesh and Madhya Pradesh using purposive convenience sampling, testing separately whether women's use of financial services influenced household financial decisions and whether their awareness of financial products did. Use of mobile banking, online bill payment and online shopping significantly influenced household decisions about spending, savings and education expenses, while awareness of credit cards, internet banking and loan services showed no significant influence. They conclude that participation in financial decisions remains constrained by sociocultural norms rather than by financial competence alone (Gupta and Bhadouria). This bears on measurement. Financial literacy instruments ask what a woman knows about financial products, and that knowledge is not what moves household decisions; routine use of everyday services is. A woman who transacts regularly but could not describe a mutual fund will score poorly on a literacy measure while doing more of her household's financial work than the score suggests. Two cautions apply: convenience sampling means the study does not support population estimates, and it does not compare rural with urban women, so it cannot establish a gap between them.

3.4 Structural constraints on urban and higher-income women

The literature on urban and middle-class women identifies different constraints. Papanek's concept of family status production describes how a household's social standing is partly generated by work that women do, and by work they visibly do not do, so that keeping a woman out of paid or public work can itself signal that a family has arrived (Papanek). Using the India Human Development Survey, Chatterjee et al. find that other family income explains some, though not all, of the relationship between women's education and their employment, with a shortage of suitable jobs a further explanation (Chatterjee et al.). Rising household income does not straightforwardly raise women's economic participation.

Twamley and Sidharth examine how respectability norms are negotiated by young women in two very different positions: Dalit women aged fifteen to twenty-two in the M Ward slum in Mumbai, working mostly in informal employment, and university-educated middle-class women aged twenty to thirty from middle and upper caste backgrounds in upmarket Baroda (Twamley and Sidharth 5). The two groups differ on class and caste simultaneously; the study does not treat these as alternative names for one division. The comparison runs against expectation. The Baroda women make what the authors call a passive bargain, upholding ideals of respectability "thereby shoring up symbolic capital for a 'good' marriage and class privilege," accepting limits on their movement and on premarital relationships rather than contesting them

(Twamley and Sidharth 21). The Dalit women, despite heavier surveillance, "show more active resistance," pushing at the boundaries of respectable behaviour and using subterfuge to win freedoms (Twamley and Sidharth 1, 5). The women with more education and more household income are the more constrained in their public conduct, and constrained partly by their own strategy.

One limit must be stated plainly. Twamley and Sidharth do not study money: their paper concerns respectability, mobility, relationships and marriage, and both groups are urban. What it establishes is that a mechanism exists by which class position restricts women's public activity, and that it bore more heavily on the wealthier group. Whether it extends to financial tasks is a question this paper raises rather than answers.

3.5 Theoretical framework

This paper accepts Kabeer's argument that no dimension of empowerment can be read on its own, and uses two of her three: resources, treated as financial access, and agency, treated as the capacity to act and to decide. The third, achievements, concerns the outcomes a woman secures, and is not measured here, because the sources reviewed report participation, ownership and decision-making rather than well-being outcomes.

The modification is to agency, which Kabeer treats as a single dimension covering decision-making and the negotiation surrounding it. This paper divides it into decision authority and task execution. The reason is empirical rather than theoretical: in the sources reviewed the two come apart repeatedly. Chauhan finds household saving and borrowing changing while women's decision-making does not (Chauhan 3, 35). Duvendack et al. find women registered as account holders whose accounts are operated by male relatives. Gupta and Bhadouria find that using financial services shifts household decisions while knowing about products does not. A single agency measure would average these together and lose the pattern. This extends Kabeer's argument rather than departing from it: her point is that an indicator's meaning depends on what sits behind it, and splitting agency applies that one level further down. The cost should be acknowledged, since a woman who delegates a task by choice is exercising agency rather than giving it up. This paper therefore does not treat delegation as evidence of dependence in itself. The distinction it draws is between delegation a woman can revoke and delegation she cannot.

4. Analysis

4.1 The assumption is built into institutions

The first step is to establish that the assumption is not merely an impression. It cannot be established directly, since this paper has not surveyed public opinion and no study reviewed measures what Indians believe about which women are financially independent. What can be shown is that the assumption is built into the design of government policy, financial advertising and the reporting of industry surveys. The claim is a deduction from those three, and is limited accordingly.

The policy evidence appears in Section 2, and its logic can be stated sharply. The Rangarajan Committee's definition, adopted as the working definition of financial inclusion in India, describes it as "the process of ensuring access to financial services and timely and adequate credit where needed by vulnerable groups such as weaker sections and low income groups at an affordable cost" (India, Committee on Financial Inclusion). The phrase "where needed" does substantial work: it presumes the need can be located, and locates it in weaker sections and low income groups. PMJDY's account of its own purpose does the same, and neither document says anything about urban or higher-income women. The deduction available is not that policymakers believe those women are financially independent, which no document states, but that these programmes are built around a location of the deficit that falls outside the urban middle class.

Industry reporting supplies the most visible element. The DBS and CRISIL finding that 98 per cent of the women surveyed participate in long-term family financial decisions travelled across the financial press: Sengupta's coverage in ETBFSI led with the independent decision-making finding, Cafemutual reported the family-advice finding, and Business Standard led with the investment allocations (Sengupta; "Many Women Still Seek Advice"; "51% of Women Prefer Low-Risk FDs"). The survey covered 800 salaried and self-employed women across ten cities. In circulation it becomes a statement about urban Indian women, a considerably larger population than the one surveyed. The reporting is accurate about the survey; the step from the survey to the population is where the assumption enters.

Two limits apply. Two campaigns, one content analysis and a handful of press articles are illustrative rather than systematic, and establishing the prevalence of a public belief would require a study of media content or audience perceptions. Whether these institutions created the assumption or inherited one already in circulation is also a question the documents cannot answer.

4.2 Applying the definition to the evidence

On decision authority a socioeconomic gradient exists, and this paper does not dispute it. Analysing NFHS-5 data, Chowdhury and Mondal report that 49 per cent of women aged 15 to 49 lack decision-making power over how to spend their own money, with urban and rural figures of 43 and 51 per cent and 54 per cent among the poorest wealth quintile (Chowdhury and Mondal). The gradient is nonetheless narrower than the assumption implies: eight percentage points between urban and rural, and five between the poorest quintile and the national figure. More than four in ten urban women lack authority over their own money. The assumption treats urban and higher-income women as the group that possesses financial independence. What the data shows is a group in which a large minority does not, separated from the rural majority by a margin rather than a boundary.

A natural objection is that earnings explain this and everything else is secondary, since women without income of their own would not be expected to control money they did not earn. The evidence does not support that simplification. Among married women with earnings, only

18 per cent decide alone how their money is spent (Kumar): these are women who do have an independent income, and fewer than one in five holds sole authority over it. Women earning as much as or more than their husbands exceed those earning less by six percentage points (Chowdhury and Mondal), where an account in which earnings determine authority would predict considerably more. Independent income matters consistently, but it is not sufficient, and the modest size of its effect leaves substantial room for the household and social factors set out in Section 3.

On task execution the assumption fares worse. Around 44 per cent of women hold a bank savings account of their own while not controlling the money in it (Chowdhury and Mondal), and Duvendack et al. describe the mechanism: women are frequently proxy users of accounts registered in their names and operated by male relatives. An account recorded as a woman's financial access may involve no financial act by her at all. A 2025 survey by Haqdarshak and DBS Bank India of 411 women entrepreneurs in rural Madhya Pradesh, Maharashtra and Rajasthan, 402 of them SHG members, found the opposite pattern: 99 per cent held a bank account, 90 per cent saved a portion of their income and 43 per cent funded their business solely through SHG loans, while only 18 per cent made financial decisions independently and 24 per cent reported that their husbands made all decisions (DBS Bank India, "90% of Rural Women Entrepreneurs"). Activity and authority are not tracking each other, within a group as well as between groups. This paper does not claim that rural women execute financial tasks more often than urban women, since no study reviewed compares the two populations on that measure. The claim is narrower: task execution is not captured by the indicators that support the assumption. Account ownership does not measure it when 44 per cent of account holders do not control their accounts, financial literacy does not measure it when awareness of products has no significant influence on household decisions, and participation in long-term financial conversations does not measure it either. The women whose financial activity the assumption discounts are active on a dimension it does not measure.

Evidence on urban women points the same way from the opposite direction. The DBS and CRISIL survey found 47 per cent of metropolitan women seeking advice and information from family members, against 27 per cent consulting financial advisers (DBS Bank India, "47% of Female Indian Earners"). Seeking advice is not the same as surrendering control, and this paper does not treat it as such, but it establishes that reliance on family members is common among precisely the group the assumption treats as independent. Whether that reliance is chosen or required is not settled by the survey, and the literature in Section 3.4 does not settle it either.

4.3 Why the assumption persists

If the evidence does not consistently support the assumption, why does it persist? The answer proposed here follows from Kabeer: what gets counted as financial independence is shaped by who is doing the counting and which population they have in view. The two DBS surveys illustrate this, and are useful precisely because they share a sponsor. Both asked about independent decision-making. Beyond that they measured different things. The metropolitan

survey recorded how respondents allocated investments, reporting 51 per cent in fixed deposits and savings accounts, 15 per cent in mutual funds and 7 per cent in stocks, together with credit card use (DBS Bank India, "47% of Female Indian Earners"). The rural survey recorded whether respondents saved, where their business funding came from, and whether they used digital banking (DBS Bank India, "90% of Rural Women Entrepreneurs"). One instrument asks what a woman owns. The other asks what a woman does. Neither is wrong for its population, but they do not produce comparable pictures, and the first maps more directly onto what the financial services industry recognises as financial engagement.

This produces a circularity. Financial independence is described in terms drawn from the financial lives of higher-income urban women, because those are the terms in which the industry producing the surveys and the advertising operates. Rural and lower-income women are then assessed against that description and found wanting, so the finding is partly an artefact of the description. This is an interpretation rather than a demonstrated finding, and testing it would mean examining a larger set of financial surveys for the assumptions built into their questions. What it offers is an explanation consistent with the evidence: the assumption survives contact with the data because much of the data is generated by instruments that already share it.

One factor cuts across this. Independent income is associated with authority wherever it appears, and belongs to no single socioeconomic group. This paper does not claim that a lower-income woman with her own earnings is more autonomous than a higher-income woman without them, because no source reviewed makes that comparison.

5. Conclusion

This paper asked how socioeconomic and geographic background affect women's financial independence in India. The answer is that the assumption favouring urban and higher-income women is institutionally embedded, and partly but not consistently supported. On decision authority the gradient is real, but narrower than the assumption implies, and more than four in ten women in the favoured group fall on the wrong side of it. On task execution the assumption is not supported, because the indicators that sustain it do not measure task execution at all. Account ownership cannot stand in for financial activity when 44 per cent of women do not control their accounts, and financial literacy cannot stand in for it when awareness of products has no significant influence on household decisions.

Financial independence in India is therefore not one thing, and is not linearly related to socioeconomic status. Independent income is the factor most consistently associated with authority, and belongs to no one socioeconomic group, but it is not sufficient, since only 18 per cent of married women with earnings decide alone how those earnings are spent. That leaves substantial room for household structure, norms about movement and permission, caste position, and the availability of usable financial infrastructure.

The limitations here are specific rather than general. The most consequential is that no source reviewed compares rural and urban women on task execution using the same instrument; the two surveys that come closest share a sponsor but differ in population, year, sampling and

questions. This is why the paper argues that task execution is unmeasured rather than higher in one group, and it is where the argument stops. The sources also measure different sub-populations: NFHS-5 questions about earnings apply only to women with cash earnings, while questions about having money one can decide how to use apply to all women, so women with and without independent income are not consistently separable. Individual sources carry further limits: Chauhan's study is a working paper, Gupta and Bhadouria use convenience sampling with no urban comparison, and Hoorunnisa and Sharma report an abstract sample of 500 against tables of 105. The literature on class and respectability, finally, establishes a mechanism restricting women's public activity but does not examine money.

What would resolve these gaps is a study asking the same questions of rural and urban women across wealth and personal-earnings groups: which financial tasks a woman carries out herself, which she asks others to carry out, and whether she could change that arrangement if she wished. The distinction between delegation a woman can revoke and delegation she cannot is the one these sources could not test, and the one that would settle whether shared financial arrangements reflect cooperation or constraint. The implication for evaluation follows. Where the stated goal is women's control over money, access measures and participation counts are not evidence that the goal has been met. They measure resources and presence. Authority has to be asked about directly.

Works Cited

Aditya Birla Sun Life Mutual Fund. "#ForHer Initiative Campaign Launch and Highlights." Aditya Birla Capital, mutualfund.adityabirlacapital.com/investor-education/herinitiative/video-individual/forher-initiative-campaign-launch-and-highlights. Accessed 7 Sept. 2026.

Banerjee, Anurag N., et al. The Power of a Financially Literate Woman in Intra-Household Financial Decision-Making. NSE-NYU Stern School of Business Initiative for the Study of Indian Capital Markets, Apr. 2021, nsearchives.nseindia.com/s3fs-public/inline-files/THE_PO~1.PDF. Working paper.

Chatterjee, Esha, et al. "Indian Paradox: Rising Education, Declining Women' s Employment." Demographic Research, vol. 38, art. 31, 6 Mar. 2018, pp. 855-78, https://doi.org/10.4054/DemRes.2018.38.31.

Chauhan, Tarana. Accounting for Empowerment? Examining Women's Financial Inclusion in India. 26 Nov. 2024. Center for Philosophy, Politics, and Economics, Brown University. Presented at the Annual Conference on Economic Growth and Development, Indian Statistical Institute, www.isid.ac.in/~acegd/acegd2024/papers/TaranaChauhan.pdf. Working paper.

Chowdhury, Aneek, and Bidisha Mondal. "Women's Control over Their Economic Resources: Evidence from NFHS 5." IWWAGE, 4 Apr. 2024, iwwage.org/womens-control-over-their-economic-resources-evidence-from-nfhs-5/.

DBS Bank India. "47% of Female Indian Earners in Metros Make Independent Financial Decisions: DBS Bank India and CRISIL Survey." DBS, 15 Jan. 2024, www.dbs.com/newsroom/47_of_female_indian_earners_in_metros_make_independent_financial_decisions_dbs_bank_india_and_crisil_survey_in.

DBS Bank India. "90% of Rural Women Entrepreneurs in India Consistently Save a Portion of Their Monthly Income: Survey by Haqdarshak and DBS Bank India." DBS, 6 Mar. 2025, www.dbs.com/newsroom/90_of_rural_women_entrepreneurs_in_india_consistently_save_a_portion_of_their_monthly_income_survey_by_haqdarshak_and_dbs_bank_india.

Duvendack, Maren, et al. "Gender Inclusivity of India's Digital Financial Revolution for Attainment of SDGs: Macro Achievements and the Micro Experiences of Targeted Initiatives." The European Journal of Development Research, vol. 35, no. 6, Dec. 2023, pp. 1369-91, https://doi.org/10.1057/s41287-023-00585-x.

"51% of Women Prefer Low-Risk FDs, Only 7% Invest in Stocks: Survey." Business Standard, 15 Jan. 2024, www.business-standard.com/finance/personal-finance/51-of-women-prefer-low-risk-fds-only-7-invest-in-stocks-survey-124011500105_1.html.

Gupta, Riya, and Shailendra Singh Bhadouria. "Women's Financial Autonomy in Household Financial Decisions: A Study on Rural Women." Journal of Emerging Technologies and Innovative Research, vol. 10, no. 7, July 2023, pp. f395-f408, www.jetir.org/papers/JETIR2307547.pdf.

"HDFC Mutual Fund Turns Dahi Handi into a Symbol of Women's Financial Freedom." Advertising Reporter, [DATE], www.advertisingreporter.com/advertising/hdfc-mutual-fund-turns-dahi-handi-into-a-symbol-of-womens-financial-freedom/.

Hoorunnisa, and Dilip Kumar Sharma. "Factors Determining the Financial Independence of Women in India." International Journal of Creative Research Thoughts, vol. 11, no. 8, Aug. 2023, pp. c360-66, ijcrt.org/papers/IJCRT2308263.pdf.

India, Committee on Financial Inclusion. Report of the Committee on Financial Inclusion. Chaired by C. Rangarajan, Jan. 2008.

India, Department of Rural Development. Cumulative Progress DAY-NRLM as on June, 2025. Press Information Bureau, 2025, static.pib.gov.in/WriteReadData/specificdocs/documents/2025/oct/doc20251023672401.pdf.

India, Ministry of Finance. "Pradhan Mantri Jan Dhan Yojana (PMJDY), National Mission for Financial Inclusion, Completes Nine Years of Successful Implementation." Press Information Bureau, 28 Aug. 2023, www.pib.gov.in/PressReleaseIframePage.aspx?PRID=1952793.

Iyer, Kavitha Venkatasubramany. "Empowering Through Femvertising? Exploring the Post-feminist Discourses in Indian Internationally Award-winning Advertisements Vis-a-Vis the World." Journal of Creative Communications, 2025, https://doi.org/10.1177/09732586251390069.

Kabeer, Naila. "Resources, Agency, Achievements: Reflections on the Measurement of Women's Empowerment." Development and Change, vol. 30, no. 3, July 1999, pp. 435-64, https://doi.org/10.1111/1467-7660.00125.

Kumar, Sunaina. "Financial Inclusion of Women: Current Evidence from India." Observer Research Foundation, 17 Aug. 2023, www.orfonline.org/research/financial-inclusion-of-women-current-evidence-from-india.

"Many Women Still Seek Advice from Family Members to Invest Money: DBS & Crisil Study." Cafemutual, [DATE], cafemutual.com/news/industry/31473-many-women-still-seek-advice-from-family-members-to-invest-money-dbs-crisil-study.

Papanek, Hanna. "Family Status Production: The 'Work' and 'Non-Work' of Women." Signs: Journal of Women in Culture and Society, vol. 4, no. 4, 1979, https://doi.org/10.1086/493663.

Sengupta, Anushka. "47% Women Take Independent Financial Decisions, 51% Prefer Low-Risk Instruments: Report." ETBFSI, 17 Jan. 2024, bfsi.economictimes.indiatimes.com/news/industry/47-women-take-independent-financial-decisions-51-prefer-low-risk-instruments-report/106916509.

Twamley, Katherine, and Juhi Sidharth. "Negotiating Respectability: Comparing the Experiences of Poor and Middle Class Young Urban Women in India." Modern Asian Studies, vol. 53, no. 5, 2019, pp. 1646-74. UCL Discovery, discovery.ucl.ac.uk/id/eprint/1567577/1/Twamley_Sidharth_MAS_Final.pdf.

Value Capture and Strategic Dependence in the AI Stack: Evidence from Fintech, Healthtech, and Agtech in India and Brazil

By Daria Mishina, Supervised by Professor Ganesh Mani

Abstract

The emergence of artificial intelligence confronts emerging economies with a distinctive strategic dilemma: firms can build sophisticated AI applications while remaining dependent on infrastructure beyond their control. This paper examines how such AI-using firms capture economic value under conditions of infrastructural dependence and how that dependence changes as firms internalize more layers of the AI stack. To address this question, the paper develops two concepts. A firm engages in dependent appropriation when it captures economic value through proprietary data, distribution, regulatory access, or sectoral expertise while remaining dependent on external infrastructure such as compute, cloud services, or model platforms. Dependency migration describes a related process: when a firm replaces dependence on an external general-purpose model with a proprietary vertical model, dependence may shift toward cloud and accelerator providers rather than disappear.

These ideas are examined through a systematic matched-sector study of six firms based in India and Brazil: Razorpay and Nubank in fintech, Niramai and Portal Telemedicina in health technology, and Cropin and Agrosmart in agricultural technology. The evidence draws on securities filings, regulatory documents, central-bank and government statistics, technical architecture documentation, and company disclosures. India and Brazil are treated as distinct institutional settings rather than as causal treatment groups. These findings support four main propositions. First, firms may capture economic value from AI applications without owning frontier infrastructure. Second, the national institutions shape the complementary assets and access conditions that firms have but do not deterministically determine the value capture. Third, stronger value capture is associated with certain configurations, which include continuous updating of proprietary data, embedding into distribution networks, and sectoral or regulatory capabilities. Fourth, vertical proprietary AI development does not imply strategic autonomy – firms may reduce dependence on model providers while shifting strategically important dependence toward cloud, compute, and accelerator infrastructure. This study argues against the overly simplistic opposition between successful AI diffusion and infrastructural dependency, which may occur simultaneously in emerging markets.

Introduction

Nubank earned approximately US$2.9 billion in net income in 2025, served 131 million customers by year-end, and developed proprietary AI models using its own financial information. Yet important parts of its technological infrastructure remain connected to foreign cloud and accelerator ecosystems. Razorpay presents a similar paradox in India. In 2026, the company released Vulcan, a proprietary payments foundation model based on roughly four billion transactions, but the model was developed and deployed with NVIDIA accelerated

computing and Amazon Web Services (Nu Holdings 2026a; Razorpay 2026a; Amazon Web Services 2026b). These companies are not technologically independent or passive absorbers of foreign AI technologies. They find themselves in a strategically difficult position because they can appropriate value downstream while remaining dependent upstream. This paper refers to that position as dependent appropriation.

The issue matters because debates around artificial intelligence in emerging economies are often structured around two competing narratives. The first focuses on diffusion. Cloud services, application programming interfaces, open-weight systems, and increasingly advanced AI tools allow companies to access capabilities that previously required larger engineering teams and much greater investment. Experimental evidence supports the proposition that AI can increase productivity without the user controlling the underlying technology. Brynjolfsson, Li, and Raymond (2025), studying 5,172 customer-service agents, find that access to generative-AI assistance increased productivity by around 15 percent on average, with larger effects among less experienced workers. More directly at the firm level, Kim, Kim, and Koning (2026) report a randomized experiment among 515 high-growth startups in which better discovery of AI use cases increased completed tasks, customer acquisition, and revenue without a proportional increase in conventional inputs.

The second narrative stresses concentration and sovereignty. Although access to applications has widened, several essential upstream inputs remain concentrated. Stanford HAI reports that NVIDIA accounts for more than 60 percent of measured global AI compute, that the United States hosts 5,427 recorded AI data centers, and that TSMC fabricates almost all leading AI chips (Stanford HAI 2026). From the first perspective, AI may allow emerging-market firms to overcome some traditional barriers to technological development. From the second, those firms may simply become more efficient tenants of foreign infrastructure. Both interpretations miss part of the problem.

The lack of ownership of semiconductor fabrication or hyperscale cloud infrastructure does not mean that a firm is incapable of developing pricing power, proprietary information, customer lock-in, intellectual property, or regulatory advantage. At the same time, commercial success does not guarantee that the same firm could sustain its AI capabilities in the event of major changes in cloud availability, accelerator access, licensing, or the geopolitical environment. The relevant outcomes are therefore different: who captures the economic value, and who controls the infrastructure required to continue producing it?

The research question is therefore: What strategies do AI-using firms in emerging economies follow in order to generate and capture value while relying on AI infrastructure controlled by third parties, and what happens to this dependence once firms begin internalizing more layers of the AI stack?

The paper makes three contributions. First, it introduces **dependent appropriation** as a strategic position distinct from technological sovereignty and simple platform dependence. A firm may maintain substantial economic leverage further down the supply chain despite lacking ownership over the technological chokepoints behind the product. Second, the paper introduces

dependency migration as the process whereby switching from dependence on an external general-purpose model to dependence on a proprietary domain-specific model changes the nature of dependence: one kind of dependence is reduced while another becomes more important. The result may be a shift from dependence on model providers toward accelerators, cloud infrastructure, storage, orchestration, and managed services. Internalization of one technological layer does not necessarily imply internalization of all layers. Third, the paper connects firm strategy with national institutions. Two relevant examples are India and Brazil, which have large domestic digital markets and sophisticated public digital infrastructure but limited control over leading global accelerator and hyperscale-cloud supply chains. However, governments still have some influence over the conditions under which domestic firms capture value through payment infrastructure, identity systems, data-sharing regulation, and access to computing power.

As far as the empirical analysis goes, the model is based on sector matching rather than on comparing firms operating in different sectors and then attributing differences in their behavior to the institutions in place in each country. Razorpay is matched with Nubank in finance, Niramai with Portal Telemedicina in health technology, and Cropin with Agrosmart in agriculture. In other words, sector matching does not establish a causal India–Brazil effect; however, it substantially reduces the risk that sectoral differences are incorrectly interpreted as institutional effects.

Theory

Diffusion is not production

The first distinction is between AI diffusion and AI production. AI diffusion refers to the adoption and use of existing AI capabilities by firms, individuals, and organizations. AI production, as defined here, refers to control over the capital-intensive capabilities required to develop and operate advanced general-purpose AI systems at scale: accelerators, compute clusters, the supporting infrastructure needed to operate them, frontier-model development, and the financial resources required to sustain these activities.

The development of these capabilities does not necessarily have to go hand in hand. A startup could leverage AI inference services from another company, implement an open-weight model using cloud-based compute, or even customize its own model through a hyperscaler without having the capacity to manufacture semiconductors. This is enough to show that economic value can be created without owning the underlying technology itself. However, it is not enough to claim that technological dependence has been overcome.

Such a difference is especially pertinent for developing nations. High adoption rates might easily be interpreted as proof of having achieved technological independence. On the other hand, an emphasis on only production might overlook the significant amount of value that could yet be created downstream. Neither of these conclusions necessarily follows. Thus, the crucial question in the analysis is not which is more relevant, but rather how their division affects the balance of power.

A layered AI stack

Layer	Principal assets	Main strategic question
Semiconductors	accelerator design, advanced fabrication, memory, packaging	Who controls scarce hardware and its production roadmap?
Compute and cloud	clusters, data centers, storage, orchestration, managed AI services	How difficult is supplier substitution?
General-purpose models	pretrained models, APIs, open weights	Can the model supplier be changed without major loss of capability?
Vertical models and applications	domain models, workflow software, product integration	What assets does the implementing firm actually control?
Customers and data	transaction histories, operational information, distribution, regulated access	Which complementary assets make imitation difficult?

Table 1. Layers of the AI Value Chain

The analysis separates five layers of the AI value chain.Dependence can therefore be reduced at one layer while remaining high at another. A firm relying on an external LLM through a proprietary API faces one form of dependence. A firm training its own transformer on proprietary data using rented NVIDIA GPUs on AWS faces another. Referring to both simply as "foreign AI dependence" overlooks an important strategic difference. The location of dependence matters because the number of available suppliers, switching costs, portability, and political exposure differ substantially across layers.

Value creation is not value capture

A second distinction concerns economic outcomes. Value creation occurs when AI makes an activity more useful, productive, accurate, rapid, or inexpensive. Value capture occurs when the implementing firm appropriates part of that value through revenue, margins, retention, bargaining power, proprietary information, or another mechanism that prevents the gain from accruing entirely to consumers, competitors, or infrastructure providers.

The question of value capture is central to business strategy. Teece (1986) argues that value from innovation depends not only on the innovation itself but also on the complementary assets required for its commercialization. Distribution, manufacturing, specialized capabilities, customer relationships, and appropriability conditions can determine which firm ultimately profits from technological change. Dynamic-capabilities theory extends this logic to firms' ability to integrate and reconfigure resources as technologies and markets change (Teece, Pisano, and Shuen 1997). It would be misleading to consider this research's contribution in the terms of reiteration of the general argument about the connection between complementary assets and value capture because such a relation is a well-established one in the body of strategy literature. On the contrary, the paper analyzes the effect of the possession of economically valuable

complementary assets by the firm that depends on the outside-controlled infrastructure. The evidence regarding the existence of asymmetric interdependence allows concluding that control over concentrated technological and economic networks can provide strategic advantage despite the persistence of business interdependence (Farrell and Newman 2019). Combining the above frameworks, the current research draws a distinction between downstream and upstream and uses the notion of dependency migration to explain how strategic dependence changes depending on the level of technological stack that is internalized by the company.

In the current AI environment, complementary resources take on increased importance because access to advanced AI capabilities is becoming increasingly widespread, while many local resources remain scarce. The companies reviewed in this research possess four important complements: continuously growing proprietary data, broad distribution and customer coverage, institutional positioning, and specialized capabilities and processes. In addition, two proprietary databases can differ substantially in their strategic value. For example, billions of payment transactions generated each day and reused in fraud detection, routing, underwriting, merchant services, and customer management may create substantially more opportunities for appropriation than a smaller database designed to solve a specific diagnostic problem.

Competitive advantage is not strategic autonomy

A firm has a competitive advantage when it can defend an economically valuable position against competitors. A firm has strategic autonomy, as the term is used in this paper, when it has a meaningful ability to sustain core operations or change suppliers in the event of shocks to infrastructure, licensing, prices, or geopolitics. The two criteria can diverge.
A digital bank can be extremely profitable while relying on external cloud services. A government can purchase a large GPU cluster while failing to develop commercially successful AI firms. Strategic autonomy should therefore not be treated as another indicator of corporate success. It is a different outcome, concerned with resilience, substitutability, and control rather than profitability.

Dependent appropriation

The concept of dependent appropriation describes a situation in which a company captures significant economic value from the downstream layers of a technology stack while remaining materially dependent on upstream inputs that cannot easily be replaced in the event of disruption. Thus, the connection becomes clearer when analyzed using a two-dimensional approach compared to a one-dimensional approach.

	Lower strategic autonomy	**Higher strategic autonomy**
Higher value capture	Dependent appropriation	High capture with infrastructural resilience
Lower value capture	Dependent application	Technological capability without effective appropriation

Table 2. Matched-Sector Case Selection

The theoretically interesting position is in the upper-left cell. This approach helps explain why the question, “Does the firm own the underlying AI?” may turn out to be strategically misleading. Ownership of the general-purpose technology is not necessary when the firm is able to retain control over the scarce complementary resources needed to exploit that technology. At the same time, high returns should not be interpreted as evidence that infrastructure dependence is no longer important.

Dependency migration

Dependent appropriation is a condition, while dependency migration is a process. Imagine a company that initially relies on a proprietary external general-purpose model. Over time, it accumulates enough proprietary data and technical capability to train its own domain-specific system internally. Its dependence on the external model provider declines, but the new system still requires compute, storage, networking, deployment infrastructure, and accelerators. If those inputs remain externally controlled, the firm has not become independent. It has relocated its dependence.

I define dependency migration as a shift in the layer at which strategically important dependencies are concentrated, caused by the internalization or substitution of capabilities in another layer of the technology stack. This definition does not imply that all forms of dependency migration are undesirable. For instance, a company may become substantially more resilient by moving from dependence on a closed API to a vertically integrated capability that can operate across multiple clouds, even if it remains dependent on external hardware.

Therefore, at least two factors should be considered when evaluating dependency: concentration, or the number of viable suppliers, and switching costs, or the ease with which a firm can move from one provider or technological configuration to another. The objective is not to eliminate all external dependencies, which would be economically unrealistic for most modern technology companies. The more relevant question is whether those dependencies remain manageable, substitutable, and sufficiently diversified.

National access autonomy and production sovereignty

At the national level, there is also a distinction between production sovereignty and access autonomy. Production sovereignty means domestic control over the technological resources necessary to produce a capability, while access autonomy means the ability of domestic firms and researchers to obtain the necessary resources on reasonable terms. These concepts should not be confused.

A government can subsidize access to imported GPUs and reduce the cost of local experimentation without manufacturing the processors itself. That increases access autonomy while leaving production dependence relatively unaffected. This distinction is particularly useful for India and Brazil because both countries are attempting to expand domestic AI capacity without reproducing every layer of the global semiconductor and cloud stack.

Propositions

Four propositions are examined in the paper. P1 — Diffusion without production ownership: AI can improve application-layer productivity and support sophisticated products without firms owning advanced AI production infrastructure. P2 — Institutional conditioning: domestic institutions influence the complementary assets and access environment of AI-using firms but do not necessarily determine value capture. P3 — Complementary-asset appropriation: among application firms relying on external AI infrastructure, those with stronger combinations of continuously refreshed proprietary information, embedded distribution, and regulatory or domain-specific capabilities are associated with stronger observable value capture. P4 — Dependency migration: when firms internalize domain-specific model capability without internalizing the infrastructure beneath it, strategic dependence shifts toward compute, cloud, or related upstream layers rather than disappearing.

Research design

Structured matched comparison

The paper uses a structured, focused comparison rather than attempting statistical identification from a small sample (George and Bennett 2005). Six firms are matched across three sectors: Razorpay and Nubank in financial technology, Niramai and Portal Telemedicina in health technology, and Cropin and Agrosmart in agriculture technology.

Sector	India	Brazil
Financial technology	Razorpay	Nubank
Health technology	Niramai	Portal Telemedicina
Agriculture technology	Cropin	Agrosmart

Table 3. Value Capture and Strategic Autonomy Framework

However, the matching approach helps to reduce the likelihood of misinterpreting any differences among sectors as the effects of institutions since it contrasts firms belonging to the same sectors in India and Brazil. It is important to bear in mind the fact that there can be no separation between “India effect” and “Brazil effect” in the design of the research. A number of possible explanations persist in the form of firm’s age, industry, financing, founder’s talent, and path dependency. Size of firms represents both an advantage and drawback of the research design since while Nubank is a listed financial company with audited financials, Niramai and Agrosmart represent smaller technology companies.

Evidence hierarchy

Evidence follows a hierarchy of priority. Primary evidence consists of audited financial statements, regulatory records, central bank reports, government data, and scientific literature. Secondary evidence consists of architecture documentation released by cloud service providers and engineering materials published by corporations. Tertiary evidence consists of operational and deployment metrics disclosed by corporations. Media reports are used only when no more direct public evidence is available and when the claim does not concern causal inference.

Claims about technical performance made by corporations and cloud service providers are treated as evidence of value creation rather than evidence of economic value capture. Where the current infrastructure architecture cannot be independently verified, it is classified as unobservable.

Measurement

The analysis is divided into three major categories. First, there are complementary assets, which include proprietary data, distribution and workflow embeddedness, institutional or regulatory access, and domain capabilities. Second, the analysis considers upstream dependence, including dependence on accelerators or compute power, cloud services, and general-model vendors. Third, there is evidence of value capture.

Evidence of value capture is considered stronger when public data demonstrates revenue or gross-profit generation, sustained profitability, significant and long-lasting customer relationships, integration into economically important workflows, and repeated use of proprietary data or distribution assets. The study deliberately avoids calculating a combined "value capture score," because such a measure could create an illusion of precision by aggregating heterogeneous forms of evidence.

Limits of inference

Four limitations are worth noting. The first is survivorship bias: successful firms that adopt AI are more easily observed than unsuccessful firms that adopt AI. The second is asymmetric financial disclosure. Nubank's profitability can be observed through securities filings, whereas the economics of Niramai or Agrosmart are much less transparent. The third is asymmetric architecture disclosure. Even where a particular system is deployed through AWS or Google Cloud, this does not imply that all of the firm's workloads run on those platforms. The fourth concerns AI intensity. The firms considered here are primarily financial, medical, and agricultural companies that use AI, meaning that their competitive advantage is enhanced through AI adoption rather than created solely by AI.

Upstream concentration: why dependence exists

In the modern AI economy, widespread availability of downstream applications coexists with concentrated production of upstream inputs. According to Stanford HAI, global AI compute capacity stood at around 17.1 million H100-equivalents during a period of rapid growth since 2022. More than 60 percent of measured compute comes from NVIDIA, the United States has

5,427 recorded AI data centers, and TSMC fabricates virtually all of the world's leading AI chips (Stanford HAI 2026). This does not mean that economic rents are restricted to upstream activities; Nubank provides evidence against such an assumption. However, these figures show that several technologically difficult-to-substitute inputs are supplied by a limited number of firms and jurisdictions.

The political dimension of this issue becomes clear in the context of U.S. export-control policy. The Framework for Artificial Intelligence Diffusion, introduced in January 2025, sought to regulate international access to advanced computing capacity. In May 2025, the U.S. Department of Commerce announced that it would not enforce the framework and intended to repeal it. In May 2026, the U.S. Government Accountability Office found that formal repeal had not taken place, although Commerce continued to apply a blanket non-enforcement policy. This provides an example of how regulatory changes can influence access to concentrated upstream inputs. For this reason, the country quotas contemplated under the framework are not treated in this paper as a stable restriction imposed on India and Brazil (U.S. GAO 2026).

From a business perspective, this means that lower access costs do not necessarily imply lower strategic dependence. Access to cloud inference may become cheaper while firms remain vulnerable to concentration in upstream inputs. Open-weight models can reduce dependence on model providers without eliminating dependence on computers. Similarly, national GPU programs can improve domestic access without changing the identity of the firms that produce the underlying processors. The application layer and the production layer therefore generate different strategic problems.

India and Brazil as institutional environments

India: public digital rails plus subsidized compute

The key institutional complement that India enjoys is not ownership of the frontier production stack, but an exceptionally broad digital rail network. According to UIDAI, approximately 1.45 billion Aadhaar numbers have been generated, alongside a very large volume of authentications and eKYC transactions. NPCI reports 731 live banks on UPI and 22.7 billion UPI transactions in June 2026 (UIDAI 2026; NPCI 2026). Such institutions can serve as public complements for private companies. A payments company does not need to develop a national real-time settlement system before providing additional financial services, and a digital platform can use existing identity and payment infrastructure without rebuilding it.

The mechanism is supported by broader empirical evidence. Cornelli et al. study 86,163 finance applications across 95 countries and find that the introduction of retail fast-payment systems is associated with higher finance-app adoption, particularly in developing countries and among new entrants such as fintech and big-tech firms. The effect is stronger where there is central-bank participation, real-time settlement, and more open access (Cornelli et al. 2025).

There has also been a notable increase in access to compute in India. As of June 2026, IndiaAI Mission stated that more than 45,000 GPUs were available through shared computing

infrastructure. Subsidized AI computing had helped in the execution of 237 projects, which altogether consumed 9.318 million GPU hours by August 2026 (Government of India 2026). This represents not merely the announcement of a sovereign-computer initiative, but its implementation as an access mechanism. Yet it does not amount to production sovereignty. Greater domestic access to compute through technologies developed and produced within international supply chains changes the cost and availability of capability, but not necessarily ownership of the underlying chokepoints. India therefore provides a useful example of access autonomy without production sovereignty.

Brazil: Pix, Open Finance, data regulation, and planned compute expansion

Brazil has institutional complements to private-sector activity that should be considered, specifically in the sphere of finance. According to Banco Central do Brasil, there are more than 170 million individual users of the service Pix, which equates to about 80 percent of the population and over seven billion transactions as of May 2026 (Banco Central do Brasil 2026a). Open Finance is yet another institutional complement that allows customers to transfer data about their accounts, cards, credits, investments, and other financial matters between the participating organizations according to standardized rules. It is estimated that there are approximately 52 million clients and 3.3 billion data queries per week (Banco Central do Brasil 2026b). These initiatives increase the effectiveness of private financial apps without changing the ownership of AI technology.

Moreover, Brazil introduced the Plano Brasileiro de Inteligência Artificial 2024–2028 that involves investment of R$23 billion in such aspects as infrastructure, AI development, education, public services, innovation, and governance. Yet, by August 2026, LNCC was still in the process of acquiring a new high-performance computing (HPC) system for AI-related needs. Such information is relevant for comparative purposes because by September 2026, India showed stronger signs of having a running subsidized GPU access program, whereas comparable AI-compute capacity in Brazil remained under implementation. Such a difference might affect the costs of experimentation, although this does not establish that Indian firms capture more economic value from AI.

Thus, Brazil shows the same general trend as in the Indian example – public institutions can complement downstream assets and improve access conditions without creating domestic sovereignty over the production stack of AI. Pix and Open Finance make financial data and digital distribution more valuable, and planned investments in computing might reduce some access problems. Neither of these developments, however, represents the notion of semiconductors or hyperscale clouds' sovereignty.

Matched sector evidence

Fintech: Razorpay and Nubank

Razorpay

One example of a proprietary AI capability operating directly above domestic payment infrastructure while relying on foreign compute infrastructure is Razorpay. AWS reports that Razorpay processes approximately US$180 billion in annualized total payment volume and uses AWS-managed services such as Aurora and EMR (Amazon Web Services 2026a). This demonstrates substantial reliance on AWS, but it does not establish that all of Razorpay's systems depend exclusively on AWS.

In August 2026, Razorpay launched Vulcan, its proprietary transformer-based foundation model for payments. According to the company, the model draws on approximately three trillion data points derived from four billion payments and around 3,000 signals per transaction. Although the model architecture and training data are proprietary, its training and deployment rely on NVIDIA GPUs and AWS infrastructure, including Amazon SageMaker (Razorpay 2026a; Amazon Web Services 2026b). This combination can be interpreted as an example of dependency migration: at the model layer, Razorpay is no longer solely dependent on intelligence produced by an external provider, while at the compute and cloud layers, important external dependencies remain.

There is also evidence of deliberate diversification across model providers within Razorpay's AI architecture. In August 2026, Razorpay stated that workloads were increasingly distributed across open and frontier models according to use case rather than concentrated on a single frontier model (Razorpay 2026b). The firm also possesses highly productive complementary assets: large volumes of transactional data, embedded merchant-payment workflows, access to multiple payment methods including UPI, and regulatory authorization from the Reserve Bank of India as an online payment aggregator (Reserve Bank of India 2026).

The commercial evidence is also substantial. During fiscal year 2025, Razorpay's consolidated revenue stood at ₹3,783 crore, whereas its consolidated gross profit was ₹1,277 crore. In addition to that, the company revealed significant consolidated loss, largely due to restructuring charges, ESOP-related charges, and redomiciling costs, but management assured that its online-payments segment was profitable on an EBITDA basis. These figures demonstrate the commercial scale of the organization but do not identify the causal contribution of AI to profitability. The more conservative conclusion is therefore that Razorpay increasingly controls sophisticated intelligence at the vertical-model layer while continuing to rely on externally supplied infrastructure at lower layers of the stack.

Nubank

Nubank provides the clearest example of downstream appropriation in the sample because its commercial performance is directly observable. In 2025, Nu Holdings' revenue was $15.8 billion and the net income was around $2.9 billion, having its customers' base at 131 million by the end of that year (Nu Holdings 2026a). Such indicators provide solid evidence for firm-level value capture, even though they do not prove that such value has been created by AI.

Nubank has also developed significant proprietary AI capabilities. According to its 2026 disclosures, the company operates a proprietary family of financial foundation models called NuFormer, used for credit-card decisioning, unsecured lending, customer service, and growth-related decisions. By the first quarter of 2026, Nubank was already using NuFormer in production for credit cards in Brazil and Mexico and for unsecured lending in Brazil (Nu Holdings 2026b). In other words, the model operates at scale on proprietary financial histories and transaction-related data.

Model ownership, however, does not imply infrastructure independence. AWS documents the migration of Nubank's mission-critical NuPay payment system to Amazon Aurora PostgreSQL, where the referenced database contained more than 31 billion rows. This demonstrates substantial use of AWS infrastructure for a mission-critical system, but it does not imply that the entire corporation is hosted exclusively on AWS (Amazon Web Services 2026c). Overall, Nubank combines profitability, large-scale and continuously refreshed proprietary financial data, embedded consumer distribution, regulatory access, proprietary AI capability, and continued exposure to external infrastructure. It therefore represents the clearest example of dependent appropriation in the sample. Nubank's commercial success despite this dependence shows that control over financial intelligence does not necessarily imply control over all of the upstream inputs required to sustain it.

Fintech comparison

The distinguishing factor in the case of the two fintech companies is not that a dependency relationship exists in one economy but not in the other. Both companies remain connected to foreign infrastructure, but each also possesses a highly productive set of complementary assets. This example suggests that the economic productivity of such complementary assets may be more important for value capture than control over the entire production stack. This does not mean that upstream control is unimportant; rather, it means that commercial appropriation and strategic autonomy are distinct outcomes.

Health technology: Niramai and Portal Telemedicina

Niramai

Niramai has a distinctive comparative advantage. Its product Thermalytix combines machine-learning algorithms with thermal imaging to conduct breast-health screening. Google Cloud documents Niramai's use of TensorFlow, Google Kubernetes Engine, Firebase, and containerized machine-learning models (Google Cloud n.d.). The system therefore represents a domain-specific AI application rather than simply an interface built on top of a generic language model.

The company's main complementary strengths are scientific and regulatory. In March 2022, the FDA made a 510(k) substantial-equivalence determination for Niramai's SMILE-100 telethermographic system. This determination applies specifically to SMILE-100 and should not

be interpreted as a general FDA approval of Niramai's products (U.S. FDA 2022). The company reports having screened more than 500,000 women across over 20 countries and holding approximately 40 patents. These figures are self-reported operational indicators rather than evidence of financial performance.

Thermal images generated through screening are proprietary and clinically specific; however, they are unlikely to be refreshed at the same frequency as payments or banking data. Distribution also depends heavily on cooperation with hospitals, clinics, and screening programs. Publicly available information on profitability is too limited to compare Niramai's economic performance directly with that of Nubank or Razorpay. Niramai therefore provides an example of strong domain-specific technological and regulatory capabilities, but with substantially weaker observable evidence of economic appropriation.

Portal Telemedicina

Portal Telemedicina provides the Brazilian healthcare comparison. The company operates a remote-diagnosis and healthcare-management platform integrated with medical devices and clinical workflows. According to company materials, its products have reached more than 33 million people, and its platform is integrated with a substantial share of medical devices in its target market. These figures are company-reported and have not been independently verified.

Portal's AI is trained on proprietary medical data rather than only on publicly available information. Google documents the company's use of TensorFlow together with a historical dataset of more than 30 million examinations and health records for machine-learning applications. Portal's main complementary assets therefore include not only its AI capabilities but also medical data, integration with clinical devices, interoperability with hospital information systems, access to specialized medical workflows, and deployment across healthcare facilities.

At present, publicly available information does not provide a sufficiently complete picture of the company's cloud architecture to estimate switching costs reliably. This dimension should therefore remain classified as unobservable rather than inferred. As with Niramai, publicly available information on profitability is too limited to support a direct financial comparison with Nubank.

Healthtech comparison

The healthcare cases present an important boundary condition for the complementary-assets argument. Although both companies possess proprietary data and distinctive AI capabilities, the economic characteristics of those data differ substantially from those observed in fintech. Medical datasets may be difficult to replicate and strategically valuable, but the deployment of medical solutions typically requires validation, procurement, integration, and less frequent end-user interaction.

The broader conclusion is that proprietary data are a heterogeneous strategic asset whose value depends on several factors, including refresh rate, potential for repeated reuse, proximity to

distribution, and the firm’s ability to monetize the predictions or decisions generated from those data.

Agriculture technology: Cropin and AgrosmCropin

Cropin offers another example of the combination of proprietary domain intelligence and rented infrastructure. According to Amazon Web Services (AWS), Cropin uses proprietary artificial-intelligence and machine-learning technology that are built in partnership with Amazon SageMaker along with AWS storage, database and messaging service for agriculture use cases. Similarly, Cropin describes its Intelligent Agriculture Cloud as an AWS solution (Amazon Web Services 2024; Cropin 2024).

In July 2026, Cropin and Google Cloud announced the creation of a platform named OrbitAI which is a Google Cloud based platform using Google Cloud enterprise AI and data solutions. As per Cropin, the OrbitAI platform uses proprietary agricultural intelligence which spans across 103 countries, over 400 crops and over one billion acres (Cropin 2026).

However, it will be wrong to give too much importance to the above development. It has not been reported anywhere that Cropin has moved all its services from AWS to Google Cloud. Instead, the evidence suggests that proprietary agricultural intelligence can operate across different hyperscaler ecosystems, products, and time periods. A firm may therefore remain dependent on the cloud layer while still retaining some degree of substitutability between providers. In this sense, Cropin is not only an example of dependency migration, but also of dependency reconfiguration. The strategic asset lies in the agricultural intelligence itself: accumulated crop knowledge, proprietary data, and domain-specific expertise. The hyperscaler remains important, but it does not necessarily control these agronomic complements.

Agrosmart

Agrosmart's architecture is more closely connected to physical agricultural operations. The company records activities in over nine countries, involving more than 100,000 farmers, over 90 crops, and approximately 48 million hectares of land. These numbers form the company adoption statistics (Agrosmart 2026).

A Globalstar case study documents an Agrosmart solution combining weather stations, soil-moisture sensors, digital rain gauges, proprietary algorithms, and satellite data transmission (Globalstar n.d.). This illustrates another form of infrastructure dependence. The external input is not simply an LLM API or cloud database; satellite communication and sensor infrastructure connect the application to technological assets outside the pure software layer.

Agrosmart's strongest complements are therefore field-level operational data, agronomic expertise, hardware and workflow integration, and relationships with agricultural producers. The company's current cloud architecture is not sufficiently transparent in reliable public sources and should therefore remain unclassified. Reliable public profitability data are also unavailable.

Agtech comparison

However, a broader principle emerges from these agricultural cases. Domain-independent artificial intelligence does not make domain-specific knowledge obsolete. General-purpose models do not contain proprietary soil analysis, farming histories, crop signatures, data feeds, or farmer relationships. At the same time, ownership of these complementary assets does not imply ownership of the underlying technological infrastructure. Cropin and Agrosmart therefore illustrate specialized applications built on top of broader technological systems.

Cross-case findings

Finding 1: diffusion without production ownership

Firms 1–6 are consistent with P1. In all six cases, the companies exhibited the ability to build sophisticated AI-enabled products or services without the need for leading-edge capabilities in semiconductor fabrication or hyperscale cloud computing. This observation is consistent with the empirical evidence showing that users can benefit from productivity improvement through AI without owning the AI infrastructure itself.

However, this statement should not be overstated. These cases do not show that AI made the formation of these companies possible, since most of them were established before the recent wave of generative-AI growth. Rather, they show that application-layer specialization can develop independently of ownership of upstream production capabilities.

Finding 2: institutions condition opportunities but do not determine winners

The findings are consistent with P2 but do not support a deterministic claim about the effect of national institutions. India's UPI system and Brazil's Pix system both function as national payment rails. Evidence from the BIS study independently suggests that retail fast-payment systems are associated with greater adoption of private financial applications, particularly in developing economies and among fintech and big-tech entrants. This provides a plausible institutional mechanism through which public digital infrastructure can support subsequent private innovation.

India's current shared-compute program also appears to provide a more immediate subsidized access mechanism than Brazil's still-developing advanced AI-compute capacity. However, the comparison does not produce a consistent finding that one country outperforms the other in terms of firm-level performance. Among the firms analyzed, Nubank provides the strongest evidence of commercial appropriation. The broader conclusion is therefore that national institutions shape the conditions under which firms build and exploit complementary assets, but they do not deterministically determine value capture.

Finding 3: complementary assets differ in economic productivity

P3 receives the strongest support from the cases. While many companies in the sample use AI systems, possession of an AI system alone is not sufficient to distinguish stronger commercial applications from weaker ones. Razorpay and Nubank stand out because they

combine proprietary intelligence with three other assets: high-frequency first-party data, embedded distribution, and regulated access.

What matters is how these assets interact. A transaction creates information that can be reused for other tasks such as authorization, fraud prevention, routing, lending, relationship management, and product development. The creation of a business or consumer relationship generates additional data, while that data can improve the product, help maintain the relationship, and in turn generate further data. This reinforcing dynamic is less evident in settings where data generation is less frequent, as is often the case in health and agriculture.

The conclusion is therefore not simply that "more data is better." Rather, the cases suggest that complementary assets support stronger appropriation when they can be repeatedly reused and reinforce one another over time.

Finding 4: dependency migrates rather than disappears

P4 provides the paper's most distinctive theoretical interpretation of the case evidence. Razorpay internalizes payment intelligence through Vulcan without giving up its dependence on AWS and NVIDIA. Nubank internalizes the financial-model intelligence with NuFormer, while some critical infrastructure continues to be sourced externally. Niramai internalizes specific machine learning capabilities used in medicine with known deployments using Google Cloud technology. Finally, Cropin uses its own agricultural intelligence in combination with hyperscaler infrastructure.

Taken together, the cases are consistent with a pattern in which internalization of vertical intelligence can reduce dependence at the model layer while leaving strategically important dependence at compute, cloud, or other upstream layers.

This study does not prove that dependency migration increases dependence in general. On the contrary, the multi-model design of Razorpay and the use by Cropin of different hyperscalers' infrastructures show that some dependencies may be more substitutable than others. As a result, dependency migration may have at least two strategic implications: (1) dependency migration with lock-in, when the new upstream dependency becomes non-substitutable, and (2) dependency migration with diversification, when the firm internalizes the specific capability, but preserves competition among infrastructure providers.

Discussion

The application layer is not necessarily a low-value layer

It is often inferred that economic power will tend to concentrate where technological bottlenecks exist. This inference is partly correct. Upstream infrastructure providers may enjoy substantial bargaining power because they control resources that are scarce and difficult to replicate, but downstream firms can also build sustainable competitive advantages through complementary assets that remain outside the control of those infrastructure providers.

Nubank illustrates this dynamic particularly well. Its dependence on external firms for parts of its infrastructure has not prevented it from capturing substantial economic value. Technological dependence, therefore, does not imply an inability to appropriate significant economic returns. This is particularly relevant for emerging-market strategy because it challenges the assumption that firms must replicate the entire AI stack in order to capture meaningful economic value.

Commercial success is not sovereignty

The reverse mistake is equally important. Strong application firms do not imply national technological autonomy. Razorpay's proprietary model still relies on NVIDIA and AWS technology, while India's shared-compute program substantially increases access but continues to operate within a global accelerator ecosystem characterized by concentrated design and fabrication.

National policy should therefore distinguish between increasing entrepreneurship, increasing domestic value capture, reducing the effective price of compute, reducing exposure to particular foreign suppliers, and establishing sovereign production capacity. These objectives are related, but they are not interchangeable.

The relevant strategic unit is the dependency portfolio

Dependency migration, however, calls for a reassessment of the classic build-versus-buy decision. A firm that chooses to develop its own model does not simply move from a state of dependence to one of independence; it reconfigures the structure of its dependencies. Internal development may weaken one form of dependence while creating or intensifying another. The use of open-weight models may reduce dependence on a model vendor while increasing the firm’s engineering and compute requirements. Managed cloud services may reduce internal capital requirements while increasing switching costs. Distributing workloads across multiple vendors may improve resilience while also increasing architectural complexity.

The goal should therefore not be that of maximizing ownership, but rather reaching a balance of efficiency between scarcity, substitutability, robustness, and costs.

Alternative explanations

However, several other possible theories remain plausible. For example, sector differences might matter in the sense that finance is naturally characterized by more frequent monetization of opportunities in contrast to health or agriculture. Scale effects matter as well since both proprietary datasets and proprietary distribution capabilities may stem from the firm’s success as well as help in building up the success. The large Nubank dataset is one such example where the dataset itself is partly a consequence of the company’s existing success as a firm.

The research also suffers from survivorship bias because the sample only includes firms that survived in the market. Firms which failed to succeed in the market may themselves have had proprietary datasets or AI technology. Moreover, disclosure bias is important as publicly

listed firms offer far more evidence compared to private firms. Finally, reverse causality and joint determination issues remain an important concern. The success of companies might generate more proprietary data and capital to develop proprietary models, which in their turn might contribute to company performance.

In aggregate, all these problems rule out any causal inference like "proprietary data cause profitability". However, they do not rule out the strategic description which lies at the heart of this study – firms with commercial appropriation can still be dependent upstream.

Implications

For founders

However, the key strategic consideration for founders goes beyond whether or not to build their own AI capabilities. The better question is which capabilities will remain scarce as access to general-purpose models becomes increasingly widespread. In many application-oriented companies, the answer is unlikely to be the generic model itself. Instead, it may be operational data, regulatory licenses, customer distribution, integration into existing workflows, reputation, domain expertise, or specialized hardware infrastructure.

While a proprietary model may strengthen some of these capabilities, founders also need to understand the dependencies that lie beneath the model. A practical analysis of resilience means considering what would happen if the company’s key model provider, cloud provider, or accelerator supplier disappeared tomorrow, and which functions could continue operating for 30, 90, or 180 days. This provides a more practical test of resilience than simply asking whether the company “owns its AI.”

For investors

The apparent discounting of application-layer businesses because they depend on leased infrastructure as well as the tendency to categorize such companies based on their proprietary artificial intelligence systems is something that ought to be avoided in investment analysis. The two variables must be separated and an investor should realize that a business may be strong on one and weak on the other.

It is therefore important for the investor to check whether proprietary information is regularly updated through routine activities, whether distribution is built-in as opposed to purchased, whether regulatory access presents an entry barrier, whether there is a single complementary asset that supports various money-making products, whether the model providers have the capacity to duplicate applications by going downstream and whether cloud and model dependencies are easily replaceable. The most appealing applications will probably be those that combine local complementary assets with foreign substitutes.

For governments

These implications also challenge conventional approaches to technological sovereignty policy. While duplicating every element of the global AI value chain is unlikely to be economically viable for most countries, the alternative does not have to be passive dependence on foreign providers. Government intervention need not be universal; it can instead be selective.

India provides one example, where the state subsidizes access to compute capacity, reducing the cost of experimentation without requiring semiconductor sovereignty. India and Brazil also show how governments can build public digital infrastructure that supports private innovation. UPI and Pix reduce transaction frictions, while Open Finance facilitates the portability of financial data under standardized regulatory conditions.

Hence, the main policy question that follows is: what dependencies are economically viable, what are too risky strategically, and what are manageable through competition, diversification, infrastructure, capability building, and alliances. This approach is closer to managed interdependence than to technological autarky.

Conclusion

In sum, AI is expanding access to sophisticated technological capabilities, while access to the infrastructure underlying those capabilities remains highly uneven. This combination does not produce a simple story of either technological leapfrogging or subordination. Instead, the cases examined in this paper point to a phenomenon of dependent appropriation.

The examples from India and Brazil show how firms can accumulate proprietary data, develop domain-specific AI capabilities, integrate into customer workflows, satisfy regulatory requirements, and—in the most pronounced cases of Nubank and Razorpay—capture substantial economic value without owning the entire upstream AI infrastructure stack. The strongest examples of appropriation are those in which complementary assets reinforce one another: transactions generate data, distribution generates additional transactions, regulatory access embeds the firm more deeply into important workflows, and proprietary intelligence increases the value of the customer relationship.

Commercial success, however, does not make firms independent of infrastructure. As firms develop more sophisticated proprietary AI capabilities, their dependencies may migrate toward other layers of the stack. Vulcan is proprietary, but the compute infrastructure on which it relies is not. Nubank possesses proprietary financial intelligence, but it does not own the semiconductor infrastructure underlying that intelligence. Niramai's specialized medical machine learning coexists with documented use of Google Cloud infrastructure, while Cropin combines proprietary agricultural intelligence with global hyperscaler infrastructure.

This is why the binary question of whether a firm "owns its AI" is becoming less informative. More relevant questions are which layers it controls, which layers generate economic rents, which dependencies are substitutable, and which dependencies could interrupt the business if access were removed. The same logic applies at the national level. India is rapidly expanding shared-compute access, increasing access autonomy without eliminating dependence

on foreign semiconductor supply chains. Brazil has developed institutional complements through Pix and Open Finance without thereby becoming an upstream AI production economy.

The core argument of this paper is that value capture and strategic autonomy are distinct optimization problems. A firm in a developing economy may optimize the first without fully solving the second, while a government may improve the second without automatically generating the first. Research on AI dependence therefore should not be limited to asking who "has AI." From a strategic perspective, greater attention should be paid to which layers of the stack are controlled, which are substitutable, and where economic rents are ultimately captured.

Future research should examine infrastructure changes longitudinally, incorporate failed firms, study switching between suppliers, and draw on interviews with founders and technical managers. Larger datasets on firms' technological architectures could test whether proprietary vertical-model development systematically shifts dependence upward in the stack and whether diversified architectures reduce vulnerability to pricing, supply, and regulatory shocks. Such evidence could move research on AI dependence beyond static descriptions of technological concentration toward a more dynamic theory of dependency adaptation and value appropriation.

Work cited

Aghion, P., Jones, B. F., and Jones, C. I. (2017). “Artificial Intelligence and Economic Growth.” NBER Working Paper No. 23928. doi:10.3386/w23928.

Agrosmart. (2026). Corporate and technology materials.

Amazon Web Services. (2024). “Cropin and the Asian Disaster Preparedness Center Increase Crop Yields through Climate Resilience Initiative.”

Amazon Web Services. (2026a). “Scaling Digital Payments to $180 Billion While Reducing Infrastructure Costs by 76% with AWS Enterprise Support.”

Amazon Web Services. (2026b). “Razorpay Launches Vulcan, India’s First AI Payments Foundation Model, Fueled by NVIDIA and AWS.”

Amazon Web Services. (2026c). “Migrating Mission-Critical Payments at Nubank to Amazon Aurora PostgreSQL.”

Banco Central do Brasil. (2026a). “Pix em números.”

Banco Central do Brasil. (2026b). “Open Finance: documentation and statistics.”

Brynjolfsson, E., Li, D., and Raymond, L. R. (2025). “Generative AI at Work.” Quarterly Journal of Economics, 140(2), 889–942. doi:10.1093/qje/qjae044.

Cornelli, G., Frost, J., Warren, J., Yang, C., and Velásquez, C. (2025). “Retail Fast Payment Systems as a Catalyst for Digital Finance.” BIS Working Papers No. 1228, revised June 2025.

Cropin. (2024). “Accelerating Progress in the Agriculture Ecosystem with the Intelligent Agriculture Cloud.”

Cropin. (2026). “Cropin and Google Cloud Launch OrbitAI; the World’s First Agentic AI Platform for Food and Agriculture.”

Farrell, H., and Newman, A. L. (2019). “Weaponized Interdependence: How Global Economic Networks Shape State Coercion.” International Security, 44(1), 42–79.

George, A. L., and Bennett, A. (2005). Case Studies and Theory Development in the Social Sciences. MIT Press.

Globalstar. (n.d.). “Solutions for Remote Monitoring and Agroeconomic Data Collection.”

Google. (2019). “AI Is Helping Rural Patients Get Crucial Medical Care.”

Google Cloud. (n.d.). “NIRAMAI Health Analytix: Making Breast Cancer Diagnosis Easier with Google Cloud.”

Ministério da Ciência, Tecnologia e Inovação. (2024). Plano Brasileiro de Inteligência Artificial 2024–2028. Government of Brazil.

Government of India, Press Information Bureau. (2026). IndiaAI Mission shared-compute implementation materials.

Kim, H., Kim, D., and Koning, R. (2026). “Mapping AI into Production: A Field Experiment on Firm Performance.” INSEAD Working Paper No. 2026/20/STR. doi:10.2139/ssrn.6513481.

National Payments Corporation of India. (2026). “Unified Payments Interface (UPI) Product Statistics.”

NIRAMAI Health Analytix. (2026). Corporate and technology materials.
Nu Holdings Ltd. (2026a). Annual Report on Form 20-F for the Year Ended December 31, 2025.
Nu Holdings Ltd. (2026b). Q1 and Q2 2026 Forms 6-K and earnings materials.
Portal Telemedicina. (2026). Corporate and technology materials.
Razorpay. (2026a). “One Foundation Model, Built for India’s Payments Ecosystem.”
Razorpay. (2026b). “The Winner Doesn’t Take It All.”
Reserve Bank of India. (2026). “Authorised Payment System Operators.”
Stanford Institute for Human-Centered Artificial Intelligence. (2026). The 2026 AI Index Report. Stanford University.
Teece, D. J. (1986). “Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy.” Research Policy, 15(6), 285–305. doi:10.1016/0048-7333(86)90027-2.
Teece, D. J., Pisano, G., and Shuen, A. (1997). “Dynamic Capabilities and Strategic Management.” Strategic Management Journal, 18(7), 509–533. doi:10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z.
U.S. Food and Drug Administration. (2022). “510(k) Premarket Notification K212965: SMILE-100 System.”
U.S. Government Accountability Office. (2026). Applicability of the Congressional Review Act to the Rescission of the Artificial Intelligence Diffusion Rule, B-337935.
Unique Identification Authority of India. (2026). “Aadhaar Dashboard.”

Drone Swarms for Faster Wildfire Detection: From Simulation to Concurrent Flight
By Ilayda Koker

Keywords

Engineering, Robotics, Swarm Robotics, Unmanned Aerial Vehicles, Wildfire Detection.

Abstract

Wildfires are increasing in frequency and severity, making rapid detection and early action critical for limiting damage. This study tests scaling of an autonomous aerial detection system from one drone to multiple drones improves wildfire detection speed, coverage, and accuracy. It analyzes whether sequential single-drone emulation predicts simultaneous multi-drone performance. A Python HSV-threshold detector program was deployed on quadcopters across a 100 ft × 100 ft field. The experiments were conducted at two altitudes and different visibility conditions. In a 96-trial factorial experiment, scaling from one to many sequentially emulated zones reduced detection time by 63% and reduced the smoke-condition miss rate from 38% to 12% (Welch's t-test, $p < 0.05$ across all four conditions). In a separate 32-trial experiment with two physically simultaneous drones, concurrent flight detected hotspots 28–41% faster than a single drone. That confirmed the initial hypothesis. Concurrent operation added 3.24 seconds to full-field coverage time ($p < 0.0001$), while detection-time overhead was only 1.24 seconds ($p = 0.67$). The analysis showed that zone count and altitude level had an impact on detection time ($R^2 = 0.372$). Smoke, on the other hand, didn't show up as a factor. Field notes suggest this might be due to survivorship bias—something worth keeping in mind. A follow-up with many concurrently flying drones would help validate what's really going on.

1. Introduction

Wildfires are among the most destructive natural disasters in the western United States. The 2025 Los Angeles fires caused an estimated $250–275 billion in damage and over 31 deaths ("2025 LA Wildfires"). Post-fire analyses consistently show that detection speed is the most critical factor in limiting damage—a fire caught within minutes is far easier to contain than one that has already spread.

This project asks a narrow, testable version of that broader problem: does adding coordinated aerial search units to a detection system improve performance in a way that is statistically significant, predictable, and — critically — real under true operating conditions rather than only under a convenient experimental proxy? The original experiment tested the first two questions using a single physical drone that sequentially flew each unit's assigned zone to emulate a multi-drone swarm. This manuscript directly tests the third question: whether that emulation actually predicts what happens when multiple drones fly at the same time.

The purpose of this research is personal as much as technical (This research was motivated by the possibility of using drone technology to protect communities and the environment). Drone technology has been widely adopted for destructive purposes in modern conflict; this project

began with the question of whether the same underlying technological advancement can be used toward protecting the environment and life instead. Living in California's Central Valley, wildfire smoke and evacuation are a routine part of fire season, which made the detection-speed gap feel like a problem worth addressing. This is what distinguishes our approach from the UAV wildfire-detection literature in Section 2. Most of those papers assume the technology is a given and just ask how to improve it. We started with a more basic question: was this even worth building?.

What this paper really offers is a direct experimental comparison between simulated swarm scaling and real simultaneous multi-drone flight. We quantify both the upside—faster detection—and the downside: the coordination overhead that comes with flying multiple drones at the same time. It situates the work in existing literature (Section 2), develops the statistical theory used to validate results (Section 3), and presents methods for both the original factorial experiment and the new concurrency validation (Section 4). Section 5 reports the 96-trial factorial results. Section 6—new in this manuscript—reports the 32-trial true-concurrent validation and directly compares it against the sequential-emulation assumption used throughout Section 5. Section 7 discusses what this comparison means; Sections 8–12 cover limitations, scope of work, conclusions, future work, acknowledgments, references, and author information.

2. Related Work

This project analyzes UAV-based wildfire detection and multi-agent swarm coordination theory. Ambrosia et al. stated that optical cameras on UAVs can detect fire signatures from altitude. Anderson and Smith identified a gap in the literature: most published UAV wildfire-detection systems evaluate single-drone performance, with little controlled testing of coordinated multi-unit performance — the direct motivation for this project.

Phan and Liu looked at coordinated multi-UAV path planning for coverage, and their work shaped the zone-splitting approach we adopted. Dorri, Kanhere, and Jurdak make the point that coordination can boost coverage efficiency by cutting out redundant searching and idle time. But as Section 6 shows, there's a catch—real coordination introduces overhead that theoretical or emulated studies often miss. As far as we know, this is one of the few student-level UAV projects that actually measures that overhead against a sequential-emulation baseline, instead of just assuming the theory will hold up.

NASA’s Fire Information for Resource Management System (FIRMS) provided real-world incident and detection-latency benchmarks used to contextualize this system's detection times relative to satellite-based fire monitoring.

3. Statistical and Regression Theory Background

3.1 The Two-Sample t-Test

We use a two-sample t-test to see if the two groups have different means, while taking each group's internal variability into account. The standard pooled-variance version assumes

both groups come from populations with the same variance. But that's often not the case, so we use Welch's t-test instead—it doesn't make that assumption and adjusts the degrees of freedom accordingly. Given the spread we see in our data, the Welch approach is the safer choice.

3.2 Simple and Multiple Linear Regression

The basic starting setup is $y = \beta_0 + \beta_1 x + \varepsilon$. When we start using more than one predictor, that same equation expands to $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$. The idea is very similar; we are just using more terms. Each coefficient tells you the effect of that one variable, while keeping all the others fixed. That's what makes it the right tool for Section 6.2—it lets us separate out the effect of true concurrency on flight time from the noise of altitude and smoke. Raw averages would just confuse things; this way we can actually see what's going on.

To check how well the model fits, we use $R^2 = 1 - (SSE/SST)$. But here's a property worth keeping in mind throughout the paper: if you fit a regression to a handful of group means, you'll almost always get a higher R^2 than if you fit the same relationship to the raw trial-level data. That's because averaging smooths out the within-group noise before the model ever sees it. We demonstrate this directly in Section 5.

3.3 Multicollinearity, Assumptions, and Diagnostics

As predictors themselves are correlated, individual coefficient estimates become unstable, which means they are multicollinearity. The Variance Inflation Factor, $VIF = 1/(1-R^2)$ from an auxiliary regression of each predictor, supports this finding. This is where VIF above 5 is commonly treated as a concern threshold.

OLS regression comes with a couple of standard assumptions—residuals should be independent and roughly normal. Rather than just taking those on faith, we test them directly. We use the Shapiro-Wilk test for normality and check for heteroscedasticity by looking at the correlation between residual magnitude and fitted values. Both are covered in Section 5.3.

4. Experimental Design and Methods

4.1 Engineering Goal and Hypothesis

The engineering question we are trying to answer is straightforward: if we send more drones to search autonomously, how much faster do we find targets? How will the coverage and detection accuracy be impacted? And on top of that—if we simulate multiple drones by running them one after another, does that actually tell us anything useful about what happens when they all fly at the same time?

We expect that adding more coordinated search units will improve detection speed and localization accuracy, roughly in proportion to the number of units added. As for concurrent versus sequential flight, we expect the same general pattern, but with some overhead in the concurrent case.

4.2 System Build, Fire Simulation, and the Original Factorial Design

Our detection system has been composed of three main components: a DJI Mini drone, Mission Planner for autonomous waypoint navigation, and a custom Python algorithm that uses HSV-based color thresholding to detect objects from camera input.
It does not require any human intervention once deployed. A red-colored mat at a known GPS coordinate simulated a wildfire hotspot; a fog machine simulated smoke.
The starting 96-trial experiment tested zone count (1, 2, or 3), altitude (10ft or 20ft), and visibility (clear or smoke) in a full factorial design, 8 trials per configuration. For configurations of more than one zone, a single physical drone flew each zone's pattern in sequence, with the field divided into proportional zones (10% overlap) to simulate what simultaneous multi-drone flight would do.

4.3 New: True Concurrent 2-Drone Validation

This manuscript adds 32 new trials (8 per altitude/smoke condition) in which two physical drones flew simultaneously over the original, full 100ft × 100ft field — not a field divided and shrunk for each unit. This setup directly tests the assumption we made back in Section 4.2. Rather than having one drone cover half the field by itself, here we have two drones actually sharing the full airspace and splitting the detection task in real time. That brings along the usual coordination headaches—keeping them spaced out, avoiding overlapping coverage, and logging GPS and timestamps independently. The same 1-drone baseline trials from the original experiment serve as the single-drone comparison point; the previous zone-split 2-drone trials serve as the “emulated” comparison point; the new trials are labeled “actual” throughout Section 6.

4.4 Field Notes on Non-Detection Trials

We made a couple of data-collection tweaks this time around. Non-detection trials are now explicitly recorded as "did not detect within the flight window" rather than just "missed"—and we added a field-note column for extra context. Across both the original and new datasets, every non-detection trial comes with the same experimenter assessment: smoke was the primary reason the drone missed the target. We made that judgment independently of any statistical analysis and came back to it alongside the numbers in Section 7.2.

4.5 Software Implementation and Program Architecture

The detection-to-hover pipeline described conceptually above was implemented as a single Python program, drone_red_mat_finder.py (approximately 800 lines), built on OpenCV, NumPy, and Pygame. The program is organized into four cooperating components — a color detector, a video-stream handler, a drone controller, and a mission display — coordinated by a top-level mission loop, described below in the order data flows through the system.
Detection. The “RedMatDetector” class converts each incoming BGR camera frame to HSV color space and applies two hue-range masks (0–10° and 170–180°) combined with a bitwise

OR, since red wraps around both ends of the HSV hue wheel. A 5×5 morphological opening removes small noise pixels and a subsequent closing fills small gaps within the mat's silhouette. The count of surviving red pixels is compared against a configurable threshold (400 pixels by default) to flag a candidate detection, and the image moments of the mask locate the pixel-space centroid of the detected region. To reduce false positives from transient glare or lighting flicker, a detection is only reported as "stable" once it has appeared in at least 3 of the last 5 analyzed frames — a simple temporal-smoothing filter layered on top of the per-frame color threshold. Navigation. Search waypoints are generated by generate_lawnmower_waypoints(), which lays out a boustrophedon (back-and-forth) pattern across the configured search rectangle, alternating direction on each pass and spacing passes a fixed distance apart (2.5 m by default). Waypoint positions are computed in GPS coordinates using an equirectangular offset approximation (offset_gps()), and the distance and compass bearing to each waypoint during flight are computed with the haversine formula, the same great-circle distance method commonly used for short-range GPS navigation.

Flight control and video. The DroneController class issues Virtual Stick commands (pitch, roll, yaw, throttle) as HTTP requests to a DJI Bridge instance running on the operator's phone, exposing takeoff, fly-to-waypoint, hover, and return-to-home/land actions. Camera frames are pulled from the drone's RTMP live stream by the VideoStream class on a dedicated background thread, so frame capture never blocks the flight-control loop. The program also includes a simulator mode that substitutes a synthetic green-field frame with a red rectangle injected after a fixed delay, allowing the detection and navigation logic to be exercised and timed end-to-end without a physical flight — the mode used for the majority of pre-field software validation and debugging described in this paper.

The run_mission() function in the program manages the whole process. It creates the waypoint path needed, initiates the video capture thread, and finally displays a checklist to be confirmed by the user. The operator must confirm readiness before takeoff, after which the drone proceeds waypoint by waypoint. The process scans frames for roughly 1.5 seconds at each stop. Upon a stable detection, the drone hovers over the mat and logs its approximate GPS location. That concludes the mission; otherwise, it completes the full search pattern and returns home. A live heads-up display (MissionDisplay) shows the annotated camera feed, waypoint progress, and detection status throughout, and the operator retains keyboard override at all times — abort-and-return-home, emergency land, and pause — consistent with the FAA Part 107 requirement that a remote pilot in command be able to intervene in an autonomous flight at any point. This manual oversight layer sits outside the automatic detection decision itself: the color-threshold trigger, centroid calculation, and GPS/timestamp logging described in Section 4.2 all execute without operator input, and it is this decision step — not the presence of a supervising pilot — that this paper characterizes as autonomous.

Configuration and reuse. All tunable parameters — flight altitude, pass spacing, detection-pixel threshold, and the search-rectangle dimensions — are exposed as constants at the

top of the script rather than hard-coded into the logic, so the same program could be reconfigured between trials without editing the detection or navigation code itself.

5. Results — Factorial Experiment (96 Trials)

As part of the results, we're showing the same factorial values from the earlier manuscript. We double-checked the data file after adding the field-note clarifications, and everything lines up—not a single mismatch across all 96 trials. Table 1 summarizes the statistics for all 12 configurations, and Figures 1 and 2 show detection-time scaling and miss-rate trends.

5.1 Summary Statistics

Table 1. Summary statistics for the original 96-trial factorial experiment across all 12 zone × altitude × smoke configurations. Listed Det. avg(s) gives mean time to first detection. Cov. avg(s) gives mean time to full-field coverage. Loc. err. defines mean localization error, which is the distance from the detection zone.; Suppression defines the percentage of trials with localization error under 3 ft, which is considered a successful detection. Miss is simply the percentage of non-detection trials; values after ± are standard deviations.

Config.	Alt.	Smoke	Det. avg (s)	Cov. avg (s)	Loc. err. (ft)	Suppression	Miss
1 zone	10ft	No	44.0 ±28.6	95.5 ±1.3	3.04	62%	0%
1 zone	10ft	Yes	44.2 ±22.5	94.4 ±2.2	2.48	50%	25%
1 zone	20ft	No	25.3 ±15.2	54.6 ±2.3	2.91	62%	0%
1 zone	20ft	Yes	28.0 ±12.3	54.4 ±2.5	2.84	38%	38%
2 zones	10ft	No	26.5 ±16.8	52.8 ±1.7	2.14	88%	0%
2 zones	10ft	Yes	27.3 ±12.4	53.8 ±2.1	2.23	75%	12%
2 zones	20ft	No	16.0 ±8.2	32.1 ±2.1	2.92	62%	0%
2 zones	20ft	Yes	17.3 ±4.8	33.5 ±1.9	2.40	62%	25%
3 zones	10ft	No	16.3 ±9.4	35.0 ±1.9	1.84	100%	0%
3 zones	10ft	Yes	17.6 ±7.9	33.6 ±2.2	1.70	88%	12%
3 zones	20ft	No	10.3 ±4.7	19.6 ±1.4	1.94	88%	0%
3 zones	20ft	Yes	11.1 ±3.5	20.3 ±1.8	2.17	88%	12%

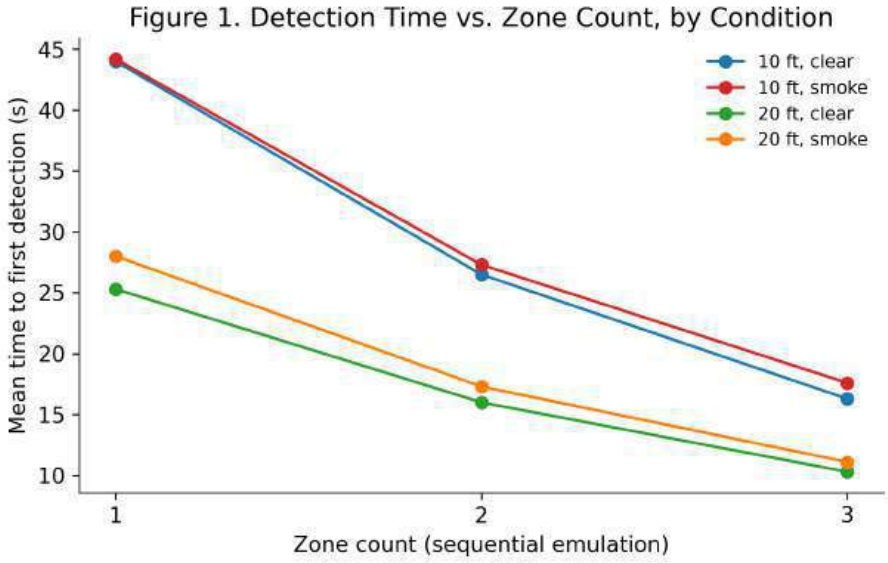


Figure 1. Mean time to first detection as a function of zone count, shown separately for each altitude/smoke condition. All four conditions show a similar, near-linear decline (Section 5.3 examines how much of this apparent linearity survives at the trial level). This figure was created using Python (matplotlib).

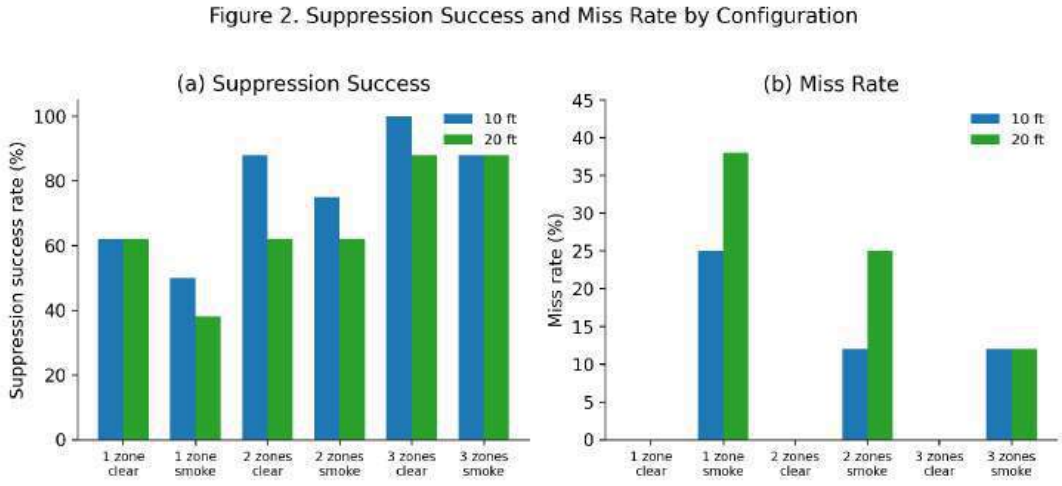


Figure 2. Suppression success rate (a) and miss rate (b) by zone count, altitude, and smoke condition. The miss-rate reduction under smoke (38% → 12%, 1 zone → 3 zones) is the clearest single visual evidence for the paper's headline claim. This figure was created using Python (matplotlib).

5.2 Hypothesis Testing (Welch-Corrected)

Table 2. shows how the single-drone and three-drone configurations compared across all four altitude and smoke conditions, using a Welch-corrected t-test.

Condition	t-statistic	Original p (pooled)	Corrected p (Welch)	Significant?
10ft, no smoke	2.605	0.0208	0.0299	Yes, $p < 0.05$
20ft, no smoke	2.672	0.0182	0.0272	Yes, $p < 0.05$
10ft, with smoke	2.752	0.0135	0.0329	Yes, $p < 0.05$
20ft, with smoke	2.967	0.0059	0.0359	Yes, $p < 0.05$

5.3 Configuration-Mean vs. Trial-Level Regression

If we regress mean detection time against zone count using configuration-level means—following the original analysis—we get an R^2 between 0.976 and 0.982. But when we fit a multiple regression directly to the 86 trial-level records that had a valid detection time (using zone count, altitude, smoke, and a zone×smoke interaction as predictors), that R^2 drops to 0.372 (adjusted $R^2 = 0.341$). Zone count and altitude come out highly significant ($p < 0.001$), while smoke and its interaction don't reach significance in this specification. Full coefficients are reported in Table 3, and Figure 3 compares this trial-level fit against the configuration-mean fit, with residual diagnostics shown in Figure 4. We come back to why that is in Section 7.2.

Table 3. Multiple regression coefficients predicting time to first detection from zone count, altitude, smoke, and their interaction, fit to the 86 trials with a valid detection.

Term	Coefficient	p-value
Zones	−10.688	< 0.0001
Altitude	−1.115	0.0004
Smoke	1.556	0.8507
Zones × Smoke	−0.162	0.9656

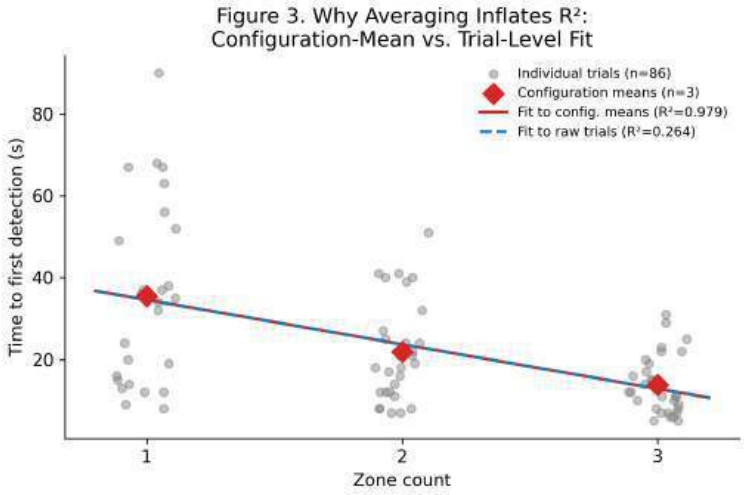


Figure 3. The same zone-count effect fit to configuration means (red, n = 3) versus fit directly to the 86 individual trials (blue dashed, n = 86). The two fitted lines are nearly identical, but R^2 falls from 0.98 to 0.26 once trial-level noise is no longer averaged away — the mechanism behind this paper's central methodological caution. This figure was created using Python (matplotlib).

We ran some diagnostics on the model. Residual mean is effectively zero— that leaves not much to go after. The Shapiro-Wilk test gives us W = 0.969 with p = 0.039, so we're seeing a bit of non-normality. This is not severe either. There is evidence of heteroscedasticity though—the correlation between residual magnitude and fitted values is r = 0.664 ($p < 0.0001$). On the overfitting front, we did an 80/20 split and got a test $R^2 = 0.389$ versus training $R^2 = 0.366$, so that looks fine. Multicollinearity? Zone count and altitude are clean (VIF ~1–2), but smoke and its interaction are a bit high at 7.4 and 8.3

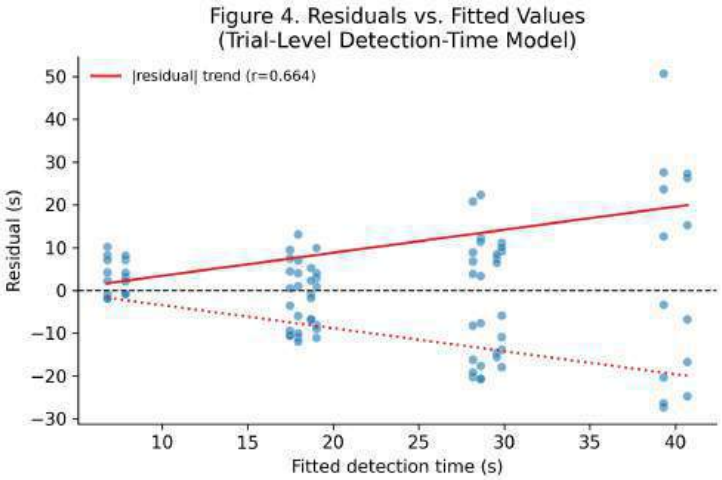


Figure 4. Residuals plotted against fitted values for the trial-level detection-time model. The widening spread at higher fitted values is the heteroscedasticity reported in-text (r = 0.664, $p < 0.0001$). This figure was created using Python (matplotlib).

6. Results — True Concurrent 2-Drone Validation (32 New Trials)

This section directly tests the assumption, carried through all of Section 5, that a single drone sequentially covering a shrunk zone accurately predicts what multiple drones flying simultaneously over the full field would achieve.

6.1 Emulated vs. Actual: Condition-by-Condition Comparison

Table 4. Emulated vs. actual two-drone detection and coverage times by condition, with overhead percentage and Welch's t-test p-values.

Condition	Metric	Emulated (sequential)	Actual (concurrent)	Overhead	p-value
10ft, no smoke	Detection (s)	26.5	27.9	+5.2%	0.869
10ft, no smoke	Coverage (s)	52.8	57.0	+8.1%	0.0001
10ft, smoke	Detection (s)	27.3	26.1	−4.2%	0.859
10ft, smoke	Coverage (s)	53.8	57.2	+6.5%	0.071
20ft, no smoke	Detection (s)	16.0	18.0	+12.5%	0.656
20ft, no smoke	Coverage (s)	32.1	34.1	+6.2%	0.053
20ft, smoke	Detection (s)	17.3	20.2	+16.3%	0.256
20ft, smoke	Coverage (s)	33.5	37.1	+10.8%	0.0149

Figure 5. Emulated (Sequential) vs. Actual (Concurrent) 2-Drone Flight

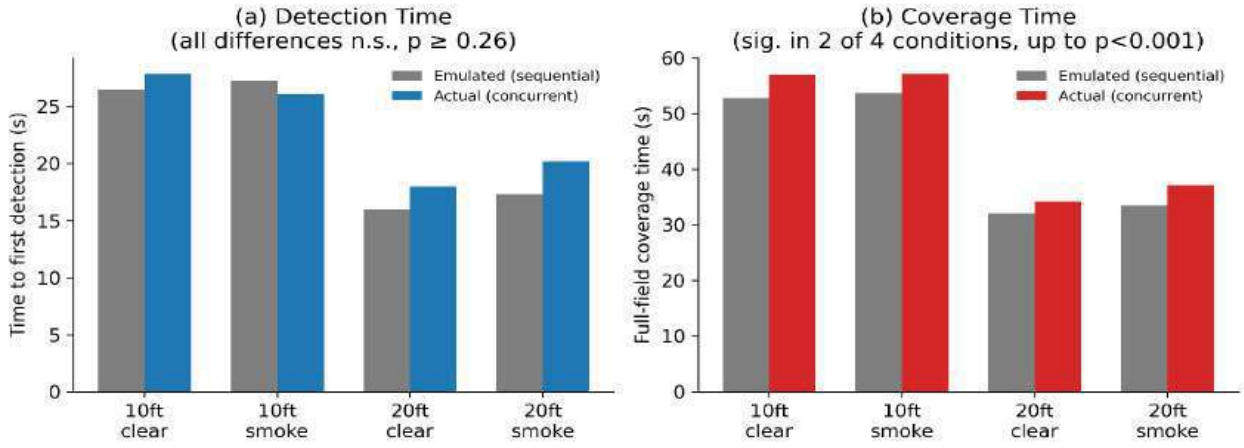


Figure 5.Emulated (sequential single-drone) versus actual (concurrent two-drone) detection time (a) and full-field coverage time (b) across all four tested conditions. Coverage-time overhead is consistent and reaches significance in two of the four conditions; detection-time differences show no consistent direction and are not significant in any condition. Generated using Python (matplotlib).

Coverage-time overhead is directionally consistent (positive) in all four conditions and reaches statistical significance in two of four (10ft/no smoke and 20ft/smoke) (Table 4). Detection-time overhead is directionally inconsistent (positive in three conditions, negative in one) and not statistically significant in any single condition — consistent with detection time simply being noisier than coverage time throughout this dataset (Section 5.3). Figure 5 plots these emulated-versus-actual comparisons directly.

6.2 Concurrency Impact & Isolating

Raw condition-by-condition experiment and comparisons mix together the effects of altitude, smoke, and concurrency. In order to isolate concurrency's effect specifically, a multiple regression was fit to the combined 64-trial emulated + actual 2-drone dataset, with altitude, smoke, and a concurrent-flight indicator (0 = emulated, 1 = actual) as predictors:

Detection Time Model

Table 5. Multiple regression predicting detection time from altitude, smoke, and a concurrent-flight indicator (0 = emulated, 1 = actual; n = 58).

Term	Coefficient	p-value
Altitude	−0.919	0.0028
Smoke	0.591	0.8419
Concurrent (actual = 1)	+1.241	0.6735

n = 58, R^2 = 0.158. Controlling for altitude and smoke, true concurrent flight added an estimated 1.24 seconds to detection time on average. This is very much not statistically distinguishable from zero (p = 0.67). (Table 5)

Coverage Time Model

Table 6. Multiple regression predicting full-field coverage time from altitude, smoke, and a concurrent-flight indicator (0 = emulated, 1 = actual; n = 58).

Term	Coefficient	p-value
Altitude	−2.120	< 0.0001
Smoke	1.569	0.0149
Concurrent (actual = 1)	+3.241	< 0.0001

This gives us n = 58 and R^2 = 0.957. With altitude and smoke held fixed, actual concurrent multiple drone flight adds an estimated 3.24 seconds to full-field coverage time on average. The impact is highly significant (p < 0.0001) and it is the cleanest, most defensible measure of coordination overhead we've been able to extract from this project. Figure 6 plots both models' concurrency-overhead coefficients with 95% confidence intervals side by side. (Table 6)

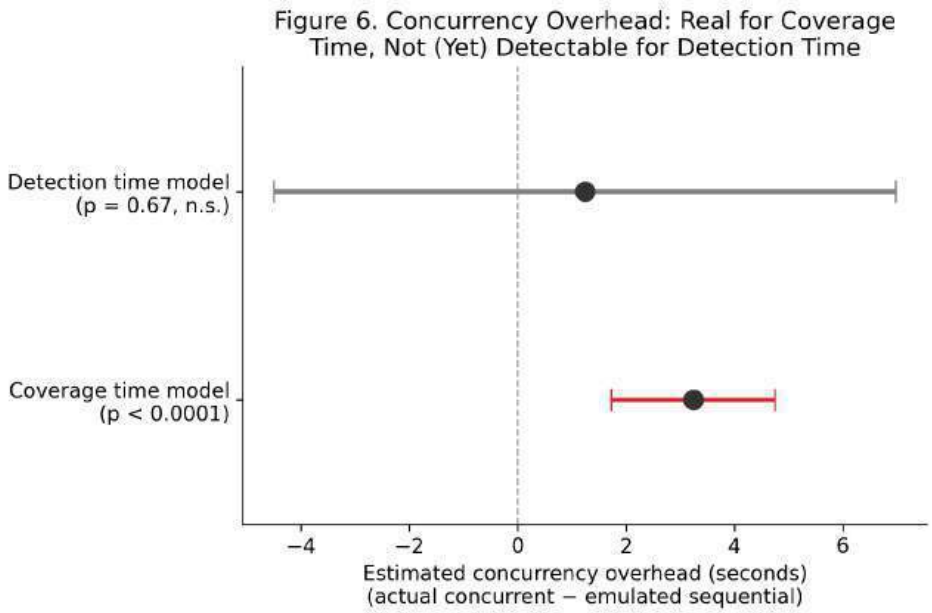


Figure 6. Estimated concurrency-overhead coefficient (actual minus emulated) with 95% confidence intervals for the detection-time model and the coverage-time model. The coverage-time interval excludes zero; the detection-time interval does not, illustrating why one overhead is statistically confirmed and the other is not. This figure was created using Python (matplotlib).

6.3 Does True Concurrent Flight Still Beat a Single Drone?

Table 7. Time to first detection for a single drone versus true concurrent two-drone flight, by condition, with percent improvement and independent-samples t-test p-values.

Condition	1-Drone (s)	Actual 2-Drone (s)	Improvement	p-value
10ft, no smoke	44.0	27.9	36.6%	0.192
10ft, smoke	44.2	26.1	40.8%	0.117
20ft, no smoke	25.2	18.0	28.7%	0.272
20ft, smoke	28.0	20.2	28.0%	0.232

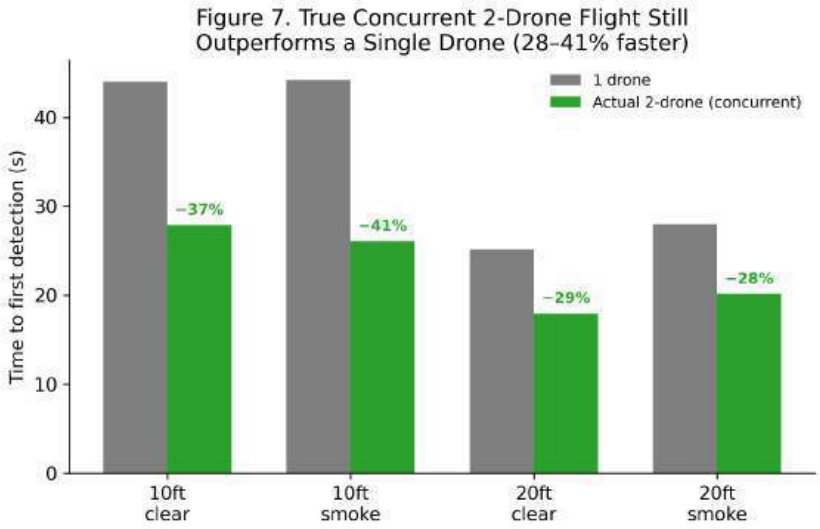


Figure 7 compares time to first detection for a single drone versus two concurrent drones across conditions. Even with the coverage-time overhead from Figures 5 and 6, the concurrency advantage holds steady at 28–41% across every condition. Created with Python (matplotlib)

Despite the confirmed coordination overhead in coverage time, true concurrent 2-drone flight remained 28–41% faster at detection than a single drone across every tested condition. (Table 7) None of these individual comparisons reach statistical significance at this sample size (n=8 vs n=8 per condition), consistent with the pattern already seen for 1-vs-2-zone comparisons in the original factorial experiment (Section 5.2) — the 1-vs-3 comparison was the one that reliably reached significance there as well, suggesting a larger jump in unit count, not a larger per-condition sample, is what it will take to confirm this two-drone advantage statistically. Figure 7 plots this per-condition comparison directly.

7. Discussion

7.1 The Sequential-Emulation Assumption Was Reasonable, but Not Free

The main question motivating the additional experiments was whether the original experiment's central methodological shortcut—emulating multi-drone search with a single sequentially flown drone— failed to reflect reality. To be clear, this doesn't upend the main takeaway. True concurrent 2-drone flight still handily beats a single drone—by about the same margin (28–41%) that the emulation predicted. But buried underneath that result was a real, measurable cost: coordinating two physical drones over the same airspace adds about 3.2 seconds to full-field coverage time on average. The most likely culprits are spacing and avoidance behavior, plus the practical difficulty of keeping two independent flight computers synchronized—not any shortcoming in the detection algorithm itself. Detection time — the more operationally important metric — showed a smaller, statistically inconclusive overhead (about 1.2 seconds), suggesting the coordination cost shows up more in finishing the full search than in how quickly the fire is first found.

In real-world wildfire operations, running a coordinated two-drone system successfully cuts down detection time. However, field conditions will eat into some of the efficiency gains seen in our emulation, especially during full-coverage missions. Operational teams should build this real-world overhead into their planning rather than expecting the raw emulation figures out in the field.

7.2 Smoke, Non-Detections, and a Convergence of Evidence

The previous manuscript's regression found smoke was not a statistically significant predictor of detection time (Section 5.3) and attributed this to survivorship bias: censored (non-detection) trials, disproportionately caused by smoke, were excluded from the time regression entirely, removing exactly the cases where smoke's effect would appear most strongly. This manuscript's clarified field notes provide independent, qualitative confirmation of that inference — every single non-detection trial in the dataset, across both the original 96-trial experiment and the new 32-trial concurrency validation, was independently assessed by the experimenter as smoke-driven. That the statistical analysis and the field observations converge on the same explanation, arrived at separately, is stronger evidence than either alone.

8. Limitations and Scope of Work

This section discusses the limitations and scope of the present work.

8.1 Limitations

- **Concurrency validated for two drones only:** The sequential-emulation assumption has now been empirically tested and partially confirmed for the 2-drone case, but the 3-drone configuration — which produced this project's strongest results — remains unvalidated by true concurrent flight. Coordination overhead may not scale linearly from two drones to many many.
- **Small sample for the concurrency comparison:** 8 trials per condition (32 total) limits the precision of the detection-time overhead estimate in particular, which was not statistically significant but also cannot be confidently ruled out as meaningful given the wide confidence interval.
- **Confirmed heteroscedasticity and mild non-normality** in the trial-level detection-time model (Section 5.3), meaning standard-error estimates should be interpreted with some caution.
- **Elevated multicollinearity** between smoke and its interaction term in the factorial-experiment regression (VIF 7.4–8.3).
- **Uncontrolled wind and single-field testing:** 6–8 mph winds were consistent across a set of experiments. All trials were conducted in one field with one geography and lighting profile.

8.2 Scope of Work

In this study, the author completed the following:

- Developed the research idea and question, along with the hypothesis for expected results. Designed the full experiment and required protocols. The zone-splitting strategy was unique to emulate a single-drone-to-multi-drone scenario.
- Wrote the Python HSV-threshold detection algorithm with basic library use.
- Programmed all drone waypoint missions using Mission Planner. The same flow is used, including synchronized simultaneous-flight coordination for the new 2-drone concurrent trials
- Executed all 96 original field trials and all 32 true-concurrent trials. Collected standardized data recording and field-note assessment of non-detection causes
- Performed statistical analysis: descriptive statistics, corrected hypothesis tests to correlate results. Including trial-level and configuration-level regression, VIF, residual diagnostics, train/test validation, and the new concurrency-overhead regression models

Resources used beyond what is freely available to any student: two personally owned DJI Mini (used) drones and batteries. Off-the-shelf free/open-source software (Mission Planner, QGroundControl, Python); a commercially licensed waypoint app (Litchi, ~$25); and access to a private outdoor field. No institutional, laboratory, or paid research support was used.

9. Conclusions and Future Work

9.1 Conclusions

This project tested whether increasing coordinated autonomous search units improves wildfire detection performance, first through a sequential single-drone emulation (96 trials) and, newly in this article, through true concurrent two-drone flight (32 additional trials). Our factorial experiment proved that scaling from one to many wildfire-emulated zones improved detection speed and lowered miss rates (Welch-corrected $p < 0.05$). Actual physical conditions played a significant role in experimentation results. That led to our trial-level multiple regression to account for roughly 37% of the overall variance between individual flights. The concurrency validation found that true simultaneous two-drone flight preserves the core benefit — 28–41% faster detection than a single drone — while introducing a small, precisely quantified, statistically significant coordination overhead in full-coverage time (+3.24 seconds, $p < 0.0001$) that the original sequential emulation could not detect. Together, these results support the underlying hypothesis while giving a substantially more complete, and more honest, picture of what deploying this system for real would actually look like. A key novelty of this project is that it does not stop at simulated swarm scaling: it experimentally tests that prediction using two drones flying simultaneously. This direct comparison turns an experimental assumption into a measurable result and provides a practical foundation for future multi-drone wildfire-detection research.

9.2 Applications and Future Work

- Run the same concurrent validation with three or more physical drones to see how coordination overhead scales and see the nonlinear behavior of it on scaling.
- Model detection as a censored time-to-event outcome so the smoke effect on non-detections—now supported by field notes—gets incorporated into the analysis rather than dropped.
- Isolating the wind factor will be critical. Collecting per-experiment wind speed and making part of the test data to factor into the results will help to explain real-world scenarios of wildfires.
- Instrument the two-drone concurrent trials to log inter-drone spacing directly, to identify the specific mechanical source of the coverage-time overhead
- To deal with optical sensor limitations, integrate thermal infrared sensing and improve the isolation of the fire impact. The thermal conditions can be emulated via heat source at the target area.
- Replace the HSV color-threshold algorithm with a CNN trained on real wildfire imagery.

10. Acknowledgments

The author gratefully acknowledges Dr. Mir Mehedi Faruk for his mentorship and guidance on this research project and manuscript.

11. Authors

Ilayda Koker is a high school student and independent researcher from California's Central Valley. Her work looks at how autonomous drone swarms can help detect wildfires—pairing real flight tests with statistical analysis. She's driven by the fires that have hit her community and wants to use drone tech to protect, not just observe.

Works Cited

Anderson, J., and R. Smith. "Autonomous UAV Systems for Wildfire Detection: A Review." Journal of Field Robotics, 2023. Providing an overview of UAVs combating wildfires.

Ambrosia, Vincent G., et al. "UAV Remote Sensing in Support of Wildland Fire Management." International Journal of Wildland Fire, 2011.

Associated Press, and Earth.com. "2025 LA Wildfires: Damage Estimates." Earth.com, Mar. 2025.The 2025 Los Angeles wildfires were catastrophic, ranking among the costliest natural disasters in U.S. history. The two largest, the Palisades and Eaton fires, destroyed approximately 16,000 structures and burned over 55,000 acres. The human toll was also immense, with at least 31 direct fatalities and over 150,000 residents forced to evacuate their homes

CAL FIRE. "2018 Camp Fire & 2017 Tubbs Fire Incident Summaries." California Department of Forestry and Fire Protection, 2018. The extreme behavior of both fires highlights the importance of speed. The Camp Fire's extreme spread, with most destruction occurring in the first four hours. Tubbs Fire's rapid 12-mile advance in three hours, leads to a key point: a fire detected and targeted within minutes is fundamentally easier to contain than one that has already grown beyond a small perimeter

Dorri, Ali, et al. "Multi-Agent Systems: A Survey." IEEE Access, 2018. Survey" published in IEEE Access is a comprehensive overview of the field of multi-agent systems (MAS). It was written by Ali Dorri, Salil S. Kanhere, and Raja Jurdak, who have a strong publication record together in distributed systems, IoT security, and blockchain

Federal Aviation Administration. "Part 107—Small Unmanned Aircraft Systems." FAA, faa.gov/uas. Part 107 is the primary set of Federal Aviation Administration (FAA) rules governing the operation of small unmanned aircraft systems (UAS), commonly known as drones. It covers everything from commercial photography to agricultural monitoring and research operations like the UAV wildfire detection tests

NASA Earthdata. "FIRMS—Fire Information for Resource Management System." FIRMS is a NASA system that distributes near real-time (NRT) active fire data from satellite observation. It was originally developed by the University of Maryland with funding from NASA's Applied Sciences Program and the UN Food and Agriculture Organization, and was transitioned to NASA's Land, Atmosphere Near real-time Capability for Earth observation (LANCE)

Phan, C., and H. Liu. "Coordinated Multi-UAV Path Planning for Coverage." Robotics and Autonomous Systems, 2022. Addresses the Multi-UAV Coverage Path Planning (mCPP) problem.

Shapiro, S. S., and M. B. Wilk. "An Analysis of Variance Test for Normality." Biometrika, 1965.

Welch, B. L. "The Generalization of Student's Problem When Several Different Population Variances Are Involved." Biometrika, 1947.

Reassessing the Neuropsychological Burden of Pediatric Hemophilia Across Treatment Eras By Rahim Noor

Abstract

Hemophilia is a congenital X-linked bleeding disorder that results from deficiencies in clotting factors VIII (hemophilia A) and IX (hemophilia B). Despite significant advances in morbidity and mortality outcomes, the disorder's effect on the developing brain is less understood, as older literature is confounded by Human Immunodeficiency Virus (HIV) acquired through contaminated factor concentrate. This review examines what the published literature reports about attention-deficit hyperactivity disorder (ADHD), anxiety, and depression in children and adolescents with hemophilia across treatment eras and evaluates how much of the reported effect is linked to hemophilia itself rather than HIV. A structured review was conducted using PubMed to identify studies reporting ADHD, anxiety, or depression in pediatric and adolescent hemophilia populations. Studies were included if they described at least one target outcome and were available in English. Prevalence statistics were arranged by assessment instrument and treatment era. Across studies more than 20 years apart, ADHD was reported at two to three times the general population rate. Anxiety and depression rates were elevated, but prevalence estimates varied with assessment instruments. The broad cognitive deficits that were reported during the HIV era were not observed under modern treatment. Hemophilia is associated with a moderate, consistent neuropsychological burden, distinct from the effects of HIV. The consistent prevalence of attention difficulties underscores the need for routine screening in pediatric hemophilia care.

Introduction

Hemophilia is an inherited bleeding disorder caused by a deficiency of one of the proteins responsible for blood clotting (Srivastava et al.). Hemophilia A results from a deficiency in clotting factor VIII, and hemophilia B results from a deficiency in clotting factor IX. Both types of the disorder are X-linked recessive and predominantly affect males (Iorio et al.; Srivastava et al.). Globally, hemophilia A occurs in approximately 17.1 cases per 100,000 males and hemophilia B in 3.8 cases per 100,000 males, making hemophilia A roughly four times more common (Iorio et al.). Severity of hemophilia is classified by clotting factor activity, with severe below one percent, moderate from one to five percent, and mild above five percent (Iorio et al.). Unfortunately, hemophilia also creates a substantial financial burden on affected individuals due to the high cost of the clotting factor. Hemophilia is a lifelong condition with no current cure, so patients require treatment throughout their entire lives at different frequencies. In the United States, average annual treatment for each patient exceeds $195,000, of which approximately eighty percent of the cost comes purely from factor replacement (Zhou et al.; Rodriguez-Merchan; Chen et al.). This review examines how treatment history has shaped the mental health of children and adolescents with hemophilia and argues that the neuropsychological burden once predominantly associated with hemophilia is, in fact, largely

driven by HIV. However, this does not eliminate hemophilia's moderate yet consistent neuropsychological effects, which are primarily manifested as attention difficulties.

Patients with severe hemophilia bleed spontaneously, as the bleeds are most often in joints and muscles. The repeated bleeding into joints and muscles that people with hemophilia face damages the cartilage and synovium over time, often leading to chronic pain and progressive degeneration defined as hemophilic arthropathy (Burke et al.). Bleeding can also occur in the central nervous system, as intracranial hemorrhage is the most serious complication of hemophilia. Most notably, intracranial hemorrhage can cause death and long-term disability, and it also demonstrates how a coagulation disorder can directly cause neurological injury (Srivastava et al.).

Treatment for people with hemophilia has changed considerably over the past four decades *(Fig 1)*. Specifically, patients were historically treated on demand, with clotting factor administered only after bleeding had already begun. This previous method controlled acute bleeding but did not prevent repeated bleeding from causing progressive joint damage. On the other hand, the modern standard of care has relied on prophylaxis, in which factor is infused on a regular schedule to prevent bleeds beforehand (Srivastava et al.). The safety of hemophilia treatment has also improved, as before viral inactivation was introduced in the mid-1980s, factor concentrates were manufactured by pooling plasma from thousands of donors, so a single contaminated donation could expose many patients, and these products often transmitted HIV and hepatitis C during treatment (Kessler). The introduction of viral inactivation and recombinant products has successfully eliminated transfusion-derived infections in high-income countries (Kessler). Thus, children and adolescents with hemophilia who are treated today often have much fewer bleeding episodes, less accumulated joint damage, and a lower risk of HIV and hepatitis C infection.

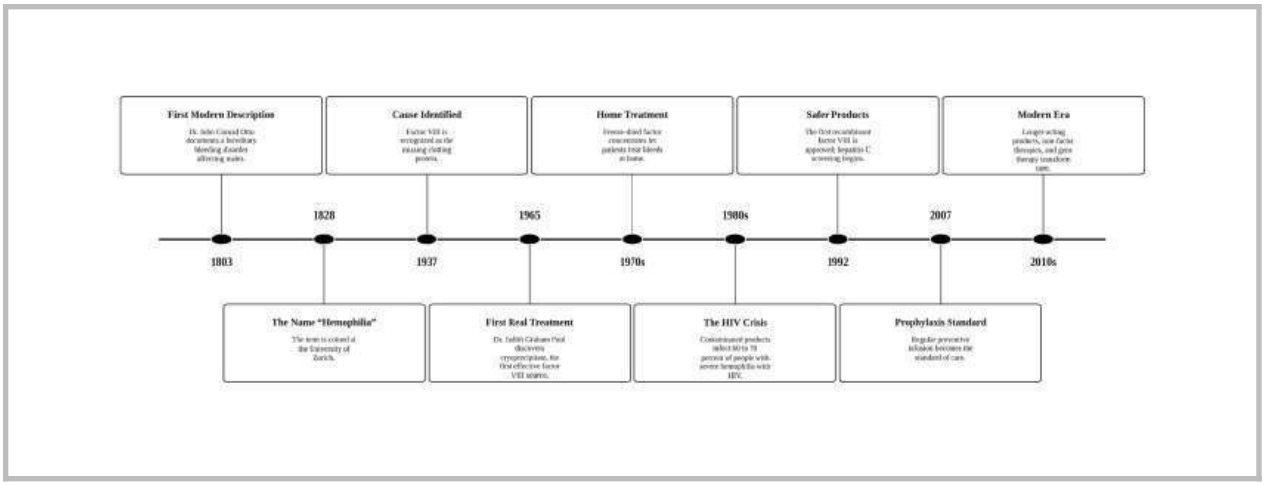


Fig 1: History of hemophilia treatment. Timeline showing major developments in the understanding and treatment of hemophilia, from its first modern description in 1803 to current therapies. Major changes include the introduction of factor replacement, the HIV crisis caused by contaminated blood products, the development of safer recombinant products, and the transition toward prophylaxis and modern treatment.

Hemophilia is considered a hematologic disorder, and its effects on cognition and mental health have been studied for several decades (Al-Huniti et al.; Buranahirun et al.). Over this time, three different mechanisms have been suggested to link hemophilia to neuropsychological outcomes: direct neurological injury from intracranial hemorrhage; the burden of chronic illness, which includes frequent venous access, activity restrictions, chronic pain, and school absence; and comorbid HIV infection contracted through contaminated factor concentrate throughout the 1980s and 1990s (Buranahirun et al.). Since the examined literature covered two distinct treatment eras, one distinguished by contaminated blood products and the other by modern prophylaxis, whether the findings reported in the older studies still apply to children treated today remains unresolved.

This review examines whether hemophilia itself increases the risk of ADHD, anxiety, and depression in children and adolescents, or whether the findings reported in earlier studies were mostly influenced by HIV infection contracted through earlier treatment. It is hypothesized that hemophilia is associated with a moderate but consistent neuropsychological burden. However, the majority of the burden associated with hemophilia in the earlier studies is driven by HIV rather than hemophilia itself, and attention difficulties remain a consistent result among children and adolescents with hemophilia today. The following sections review the available evidence in three stages: reported prevalence of ADHD, anxiety, and depression; findings from the HIV era; and outcomes reported under modern treatment conditions.

Methods

Search Strategy

A structured literature review was performed using PubMed (https://pubmed.ncbi.nlm.nih.gov/) to identify studies that reported anxiety, depression, and ADHD in children and adolescents with hemophilia. The final PubMed search was conducted in May 2026. The search combined the terms “hemophilia” OR “haemophilia” with outcome-related terms including “ADHD,” “attention,” “anxiety,” “depression,” and “neuropsychological.” The exact Boolean search string was: (hemophilia OR haemophilia) AND (ADHD OR attention OR anxiety OR depression OR neuropsychological). There was no lower limit placed on the publication date, as studies that were conducted before and after the introduction of viral inactivation and routine prophylaxis were both required in order to compare treatment eras.

Inclusion Criteria

Studies were included if they examined a pediatric or adolescent population with hemophilia, reported at least one of the three target outcomes, and were available in English. Priority was given to meta-analyses, review articles, and longitudinal cohort studies. Additionally, studies of adult populations were included only when they provided prevalence baselines for comparison or included participants of all ages. Adult and mixed-age studies were

used only to provide contextual comparisons and were not treated as primary evidence for pediatric or adolescent neuropsychological outcomes.

Screening and Data Extraction

A total of 243 records were reviewed by title and abstract, and 34 studies met the inclusion criteria and were preserved for full-text review *(Fig 2)*. For each study that met the inclusion criteria, data were extracted on the author and year, country, age group, sample size, treatment era, assessment instrument, reported outcome, and prevalence if applicable. Assessment instruments varied across the selected studies and included the Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS), the Child and Adolescent Psychiatric Assessment (CAPA), the Screen for Child Anxiety Related Emotional Disorders (SCARED), the Children's Depression Inventory (CDI), the Beck Depression Inventory (BDI), and the ADHD-IV rating scale.

Data Synthesis and Quality Assessment

Since the included studies differed in design, sample size, and assessment methods, their findings were summarized by topic instead of being combined into a formal meta-analysis. Prevalence statistics were organized by instrument class, which separated structured diagnostic interviews from symptom inventories and rating scales because both methods produce estimates that cannot be compared directly. Furthermore, internalizing problems, such as anxiety and depression, were analyzed separately from attentional difficulties. Studies were also grouped by treatment era to compare findings before and after the introduction of viral inactivation and prophylaxis. Specifically for this review, 1992 was used as an operational cutoff between historical and modern treatment periods because recombinant factor VIII was first approved in that year. Therefore, studies primarily reflecting treatment before 1992 were classified as HIV-era studies, while studies reflecting treatment after 1992 were classified as modern-era studies. This cutoff is a simplification, as viral inactivation, recombinant products, prophylaxis, and other improvements in comprehensive care were adopted gradually and at different times across countries. Lastly, summary statistics from previously published meta-analyses were used as supporting evidence and are clearly identified throughout the review. Since this review summarized the findings rather than combining them statistically, a formal risk-of-bias score was not implemented. Instead, study quality was evaluated on sample size, study design, and the use of standardized, validated measures, which were considered when interpreting the findings.

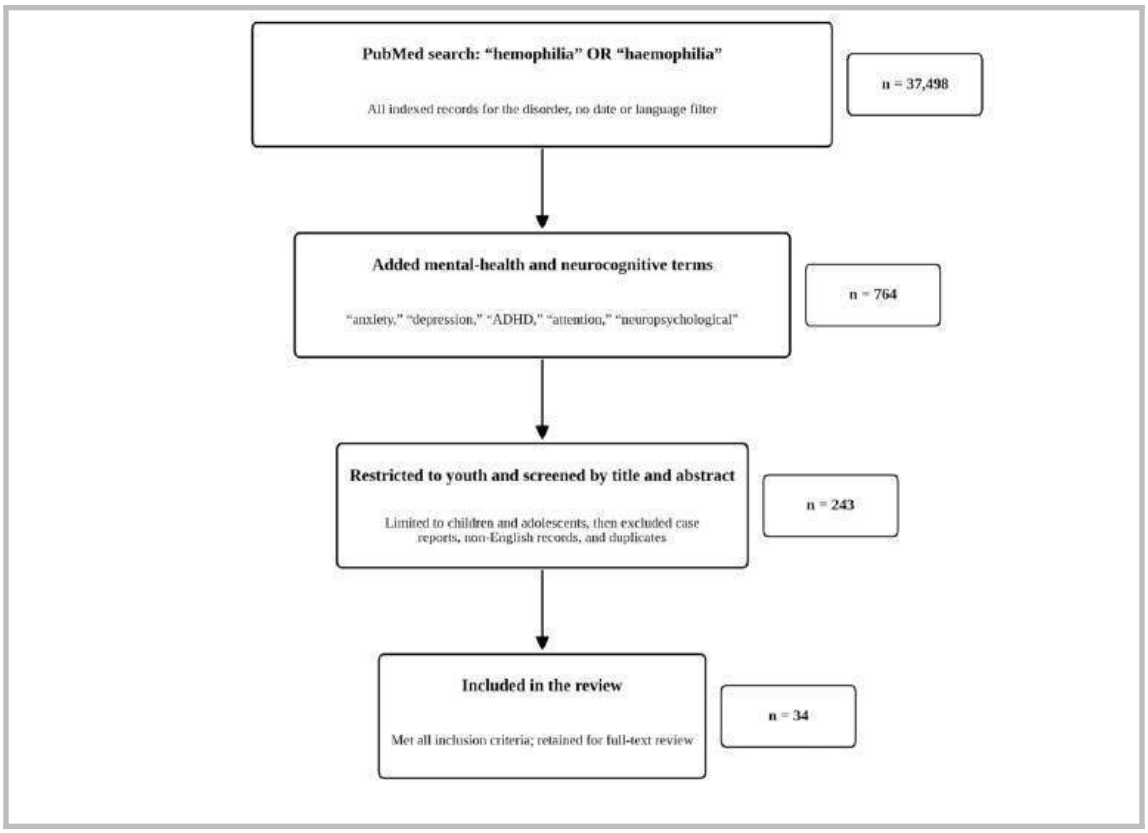


Fig 2: Study selection process. Flow diagram showing the literature search and screening process used in this review. The initial PubMed search identified 37,498 records, which was narrowed to 764 after adding mental health and neuropsychological search terms. After restricting the results to children and adolescents and screening titles and abstracts, 243 records remained, with 34 studies ultimately included in the review.

Results

Prevalence of ADHD, Anxiety, and Depression

ADHD is the most consistently reported neuropsychological finding in the pediatric and adolescent hemophilia literature reviewed (Al-Huniti et al.; Mrakotsky et al.; Wodrich et al.) *(Table 1)*. The first study to test the association directly was performed in 2003 at a hemophilia treatment center in Phoenix, Arizona. Of 78 eligible boys aged from 5 to 14, 34 participated, and all of the boys were confirmed HIV negative by both serological and PCR-based assays (Wodrich et al.). In terms of parent rating scales, 26 percent exceeded the clinical cut-off for inattentive ADHD, 18 percent for hyperactive-impulsive ADHD, and 18 percent for both inattentive and hyperactive-impulsive ADHD. On the other hand, teacher-completed rating scales demonstrated lower rates, as only 4 percent of participants exceeded the cut-off for inattentive ADHD and none for hyperactive-impulsive ADHD. Many participants received ADHD medication during school hours, which explains the discrepancy between the different rater groups, and a retrospective chart demonstrated that 29 percent of the participants already had a prior ADHD diagnosis (Wodrich et al.).

Table 1 | Characteristics of studies included in the review: Summary of the studies included in this review, organized by publication year. The table presents study year, location, participant sample size, design, era, primary focus area, assessment method, and key findings.

Study	Year	Country	n	HIV status	Design	Era	Focus / measure	Diagnostic vs. screening	Key result
Pediatric and adolescent neuropsychological studies									
Wodrich & Recht	2003	US	34	HIV−	Cross-sectional	Modern	ADHD (rating scale)	Screening	26% exceeded ADHD cut-off; 29% already diagnosed
Spencer et al.	2009	US	19	Not reported	Case-control	Modern	ADHD (ADHD-IV)	Screening	More hyperactivity and inattention than controls
Abali et al.	2014	Turkey	42	Not reported	Cross-sectional	Modern	Anxiety (SCARED)	Screening	38% anxiety; tied to overprotective parenting
Ghanizadeh & Baligh-Jahromi	2009	Iran	83	Not reported	Cross-sectional	Modern	Depression/anxiety (K-SADS)	Diagnostic	Depression 6%, anxiety 5%; some suicidal thoughts
Hassan et al.	2011	Egypt	50	Not reported	Cross-sectional	Modern	Depression/mood (BDI)	Screening	68% depression; worse joints linked to lower mood
Trzepacz et al.	2003	US	40	Not reported	Case-control	Modern	Social/emotional (multi-rater)	Screening	More sadness and lower self-image than controls
Canclini et al.	2003	Italy	NR	Not reported	Case-control	Modern	Coping/anxiety (STAI)	Screening	Different coping and more emotional strain than controls
Kvist et al.	1990	Finland	NR	Not reported	Cohort	HIV era	School/emotional (records)	Screening	Mild/moderate cases had more emotional disturbance
Miles et al.	2012	Canada	16	Not reported	Case-control	Modern	Neurocognitive after ICH	Diagnostic	Weaker language and math; most scored average
Loveland et al. (HGDS baseline)	1994	US	333	Mixed HIV+/−	Longitudinal cohort	HIV era	Neurocognitive battery	Diagnostic	~25% scored low; achievement below IQ
Sirois et al. (HGDS)	1998	US	298	Mixed HIV+/−	Longitudinal cohort	HIV era	Neurocognitive + MRI	Diagnostic	Brain scans and neuro signs linked to lower scores
Usner et al. (HGDS)	1998	US	333	Mixed HIV+/−	Longitudinal cohort	HIV era	Neurocognitive battery	Diagnostic	Coordination and severity linked to lower scores

Study	Year	Country	n	HIV status	Design	Era	Focus / measure	Diagnostic vs. screening	Key result
Loveland et al. (HGDS)	2000	US	333	Mixed HIV+/−	Longitudinal cohort	HIV era	Neurocognitive battery	Diagnostic	Immune decline linked to weaker cognition
Cheung et al.	2023	China	NR	Not reported	Cross-sectional	Modern	Neurocognitive + imaging	Diagnostic	Brain and behaviour factors linked to weaker cognition
eTHINK (Mrakotsky et al.)	2024	US	551	HIV−	Cross-sectional	Modern	Cognition/behaviour (battery)	Diagnostic	Mostly normal development; 21% ADHD in teens
Shapiro et al.	2001	US	131	Not reported	Cohort	HIV era	Academic achievement	Diagnostic	More special-ed; bleeding linked to weaker achievement
Boggs et al.	2023	US	NR	Not reported	Quality improvement	Modern	ADHD screening in clinic	Screening	Screening can be built into routine clinic visits
Contextual sources (epidemiology, cost, guidelines, adult, adherence, reviews)									
Iorio et al.	2019	Int.	Registries	—	Registry meta-analysis	—	Epidemiology	—	Prevalence 17.1 (A) and 3.8 (B) per 100,000 males
Al-Huniti et al.	2020	Int.	28 studies	—	Meta-analysis	—	Mental health	—	Higher depression, anxiety, ADHD than general public
Ayano et al.	2023	Int.	588 studies	—	Umbrella review	—	General-population ADHD	—	ADHD ~8% overall; ~10% among boys
Srivastava et al. (WFH)	2020	Int.	Guideline	—	Clinical guideline	—	Treatment standards	—	Prophylaxis is standard of care; ICH can cause death
Zhou et al.	2015	US	222	—	Cost analysis	—	Economic	—	Hemophilia A ~$195,000/yr; ~80% from factor
Rodriguez-Merchan	2020	Int.	Review	—	Review	—	Economic	—	Minimizing treatment cost without reducing quality
Chen et al. (HUGS Vb)	2017	US	NR	—	Cost analysis	—	Economic	—	Severity and regimen linked to higher costs
Burke et al. (CHESS US)	2021	US	101	—	Survey (adults)	—	Economic / PRO	—	85% chronic pain; mean annual bleed rate 1.73

Study	Year	Country	n	HIV status	Design	Era	Focus / measure	Diagnostic vs. screening	Key result
Kessler	2006	Int.	Review	—	Review	—	Liver disease	—	Viral inactivation reduced transmitted infection
Buranahirun et al.	2020	US	Review	—	Review	HIV era	Neuropsychological (HGDS)	—	HGDS found elevated attention problems/ADHD in HIV era
Steen Carlsson et al. (MIND)	2023	Nordic	3,246	—	Prescription registry (all ages)	—	Internalizing	—	Higher depression/anxiety medication use than controls
Koseoglu et al.	2024	Turkey	90	—	Cross-sectional (adults)	Modern	Internalizing	—	~24% depression, ~74% anxiety in adults
Hoefnagels et al.	2020	Int.	Qualitative	—	Qualitative study	—	Adherence	—	Self-management transition makes adherence harder
Witkop et al.	2016	US	NR	—	Survey	—	Adherence	—	Busy schedules and forgetting predict non-adherence
Pai & McGrady	2014	Int.	Meta-analysis	—	Meta-analysis	—	Adherence (chronic illness)	—	Interventions improved adherence in chronic illness
Konigs et al.	2024	Int.	Review	—	Review	—	Adolescent care	—	Teens may see treatment as a burden and skip it
Palareti et al. (Cochrane)	2020	Int.	7 trials, 362	—	Review of trials	—	Interventions	—	Support programs helped youth; evidence limited

A follow-up study, conducted in 2009, from the same center also compared boys with hemophilia to a control group without hemophilia and withdrew participants from stimulant medication for three days before the rating scales were completed (Spencer et al.). In this study, 19 boys with hemophilia aged from 6 to 14 were assessed, and teachers rated the hemophilia group higher on both hyperactivity-impulsivity and inattention. However, parents rated the group higher on hyperactivity-impulsivity only. Reading and mathematics scores showed trends toward lower performance without reaching any statistical significance, and special education participation was higher in the hemophilia group (Spencer et al.). The eTHINK study, whose data were collected between 2018 and 2019, enrolled 551 males aged from 1 to 21 with hemophilia A (n = 433) or hemophilia B (n = 101) and documented ADHD in approximately 21 percent of adolescent participants (Mrakotsky et al.). In terms of the general population of

children and adolescents, meta-analytic estimates place ADHD prevalence at 8.0 percent and approximately 10 percent among boys only (Ayano et al.). In regard to the hemophilia literature, reported rates across the studies are two to three times the expected population rate (Al-Huniti et al.; Mrakotsky et al.; Wodrich et al.).

Anxiety and depression are reported at elevated rates; however, the prevalence estimates for anxiety and depression vary more than for ADHD. A study of 42 boys aged from 7 to 16 in Turkey found anxiety in 38 percent of participants, and state anxiety scores exceeded trait anxiety scores (Abali et al.). One possible explanation for this is the high levels of maternal overprotection and strict discipline that are also reported in the study (Abali et al.). Additionally, a study of 83 children and adolescents aged from 5 to 19 conducted in Iran used the K-SADS and identified depression in 6 percent and anxiety in only 5 percent of participants, and some participants even described suicidal thoughts or preference for death (Ghanizadeh and Baligh-Jahromi). Depression estimates in other studies are even higher, as in a study conducted in Egypt of 50 adolescents with hemophilia A reported depression in 68 percent of participants using the Beck Depression Inventory and found that worse joint outcomes were associated with a worsening mood (Hassan et al.). A clinic-based study of 90 patients reported depression in approximately 24 percent and anxiety in approximately 74 percent (Koseoglu et al.).

The prevalence estimates vary considerably across the studies; however, two controlled studies offer insightful explanations for these differences. Specifically, a comparison of 40 boys with hemophilia and another 40 boys without hemophilia found comparable social functioning but more described sadness and lower self-esteem in the hemophilia group (Trzepacz et al.). These findings indicate that children and adolescents with hemophilia may struggle emotionally even when they appear to be functioning well socially (Trzepacz et al.). Secondly, a separate case-control study found that patients with hemophilia often used different coping strategies and described greater emotional strain than those in the control group (Canclini et al.). Thus, these studies suggest that the wide range in reported anxiety and depression reflects a genuine emotional burden that may be masked by normal social functioning in children and adolescents with hemophilia.

Reported rates of anxiety, depression, and ADHD varied according to which assessment method was used and the treatment era studied *(Fig 3).* The variation among the prevalence estimates of anxiety, depression, and ADHD is mostly due to differing instrumentation methods. For example, structured diagnostic interviews often lead to lower prevalence estimates than symptom inventories because the diagnostic interviews require participants to meet a certain diagnostic criterion, while symptom inventories classify individuals based on symptom scores that exceed a preestablished threshold. Among the depression studies included in the review, every estimate created from a structured interview was at most 12 percent, whereas every estimate derived from a symptom inventory was at least 17 percent *(Fig 3A).* Thus, direct comparison across instrument classes is not applicable; however, all methods demonstrate a general trend, in which children and adolescents with hemophilia consistently indicate higher rates of anxiety and depression than the general population. Since direct comparison is not

appropriate, individual prevalence estimates must be interpreted cautiously, as the same population can appear more or less affected depending on the assessment method used. A meta-analysis of 28 studies also found higher rates of depression, anxiety, and ADHD in people with hemophilia than in the general population (Al-Huniti et al.). The MIND study examined dispensed prescriptions for 3,246 people with hemophilia from four Nordic countries over 11 years and found higher use of antidepressant and anxiolytic medications than in matched control groups. The use of antidepressant and anxiolytic medication was most elevated in the group with moderate to high factor consumption (Steen Carlsson et al.).

Lastly, intracranial hemorrhage affects the brain directly rather than through the burden of living with hemophilia. Two studies examined children with a documented history of intracranial bleeding. The first study compared six boys aged from 4 to 12, who had a history of intracranial hemorrhage, to 11 controls and found deficits in motor and visual function, with four of the six boys also requiring additional learning support at school (Kvist et al.). The second study compared boys aged from 3 to 17 with and without a history of intracranial bleeding, matched for hemophilia severity, and found weaker performance in language and mathematics among the boys with a history of intracranial bleeding (Miles et al.). The majority of the participants in both studies scored within the average range on the neurocognitive and academic tests they took, and both studies contained a small sample size. Despite their small sample sizes, both studies suggest that the cognitive deficits observed after intracranial hemorrhage may be associated with direct brain injury, whereas anxiety and depression may be more related to the emotional burden of living with hemophilia.

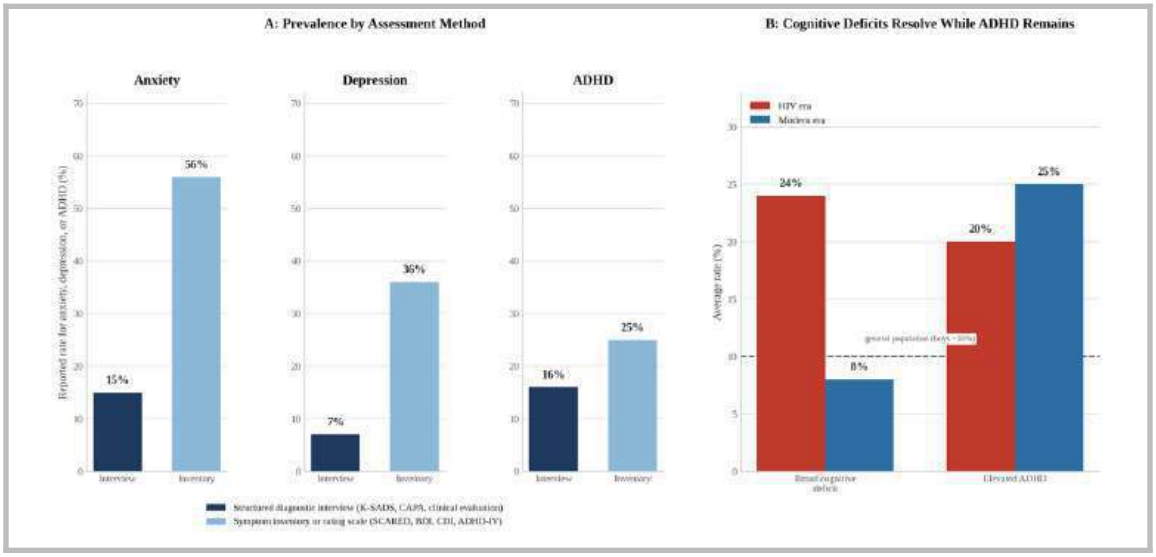


Fig 3: Neuropsychological findings across assessment methods and treatment periods. Bars represent unweighted averages calculated from the studies included in each category and are intended to show general trends rather than pooled statistical estimates. (A) Anxiety, depression, and ADHD prevalence varied by assessment method, with symptom inventories and rating scales producing higher estimates than structured diagnostic interviews. (B) Broad cognitive deficits were less common under modern treatment than during the HIV era, while elevated ADHD remained present.

Neuropsychological Outcomes During the HIV Era

The largest study regarding cognition associated with hemophilia is the Hemophilia Growth and Development Study (HGDS), a multi-site longitudinal study that included 333 boys with moderate to severe hemophilia beginning in 1988. The cohort included 207 HIV positive and 126 HIV negative boys aged from 6 to 19. A baseline assessment determined that the two groups of boys performed relatively similarly on neuropsychological testing when the study first began, and approximately 25 percent of the full cohort scored below average in at least three domains of testing (Loveland, J. Stehbens, et al.). Academic achievement and adaptive skills fell below the levels predicted by measured intelligence as well (Loveland, J. Stehbens, et al.). Beyond these general cognitive findings, analyses of the HGDS cohort also reported elevated rates of attention problems and ADHD relative to the general population, indicating that attention difficulties were already present in children with hemophilia during the HIV era (Buranahirun et al.).

A companion analysis examined the relationship between neuropsychological performance, neurological examination, and magnetic resonance imaging in 298 participants aged from 7 to 19. In this analysis, lower neuropsychological scores were linked to a history of academic difficulty, coordination and gait abnormalities, and parental education level (Sirois et al.). Another separate analysis of the entire cohort also found that participants without coordination or gait abnormalities scored higher in reading and mathematics, and that lower academic achievement was associated with greater hemophilia severity (Usner et al.).

Because many participants within the study had both hemophilia and HIV, cross-sectional data could not separate the effects of each condition; however, longitudinal analysis made this crucial distinction possible. Specifically, growth curve modeling of neuropsychological performance across the HGDS cohort demonstrated that declines in test performance corresponded with declines in immune function. Further, participants whose immune status deteriorated showed the largest decrease in cognitive performance, and those participants whose immune status remained stable did not show a similar decline. Thus, this finding suggests that the majority of the observed neuropsychological deterioration is due to HIV progression rather than to hemophilia (Loveland, J. A. Stehbens, et al.).

Since HIV positive and HIV negative participants performed similarly at baseline, the deficits observed at the beginning of the study were not caused by HIV itself. Academic achievement fell below intelligence-predicted levels across the full cohort rather than only in the HIV positive group. Therefore, this evidence from the HGDS data indicates that HIV accounted for the decline over time, whereas hemophilia itself only accounted for a smaller baseline difference that was already present before such decline began. This distinction is extremely important because it separates the effects of HIV from the effects of hemophilia. If HIV had caused the neuropsychological deficits alone, the HIV positive group would have performed much worse from the beginning of the study. However, both groups began the study at similar levels, and the differences only appeared as immune function declined, further supporting that HIV was responsible for the majority of the cognitive decline (Loveland, J. A. Stehbens, et al.).

Neuropsychological Outcomes Under Modern Treatment

The eTHINK study examined 551 males aged from 1 to 21 throughout United States hemophilia treatment centers and used standardized measures of cognition, emotional and behavioral adjustment, and adaptive functioning to perform the study. Most participants performed similarly to children and boys their own age in the general population. This result differs from the HGDS findings, in which approximately a quarter of the participants scored below average across multiple domains (Mrakotsky et al.).

Despite the largely typical cognitive performance, several differences were found in attention. Participants in the 4 to 5 and 10 to 21 age ranges demonstrated worse attention performance. Specifically, ADHD was present in approximately 21 percent of adolescent participants, and 64 participants with comorbid ADHD reported more adaptive and executive function problems in daily life. Because attention remained an area of concern, the authors recommended screening for attention and learning difficulties (Mrakotsky et al.).

Under modern conditions, the cognitive impairments that were reported in the HGDS are no longer seen *(Fig 3B)*. Specifically, general intellectual functioning, the proportion of participants scoring below average across multiple domains, and the overall pattern of academic underachievement are absent in the modern cohort, and the removal of HIV and the reduction in bleeding burden help explain this improvement (Loveland, J. A. Stehbens, et al.; Mrakotsky et al.). ADHD rates were elevated in an HIV negative sample examined in 2003 (Wodrich et al.), in a controlled comparison conducted in 2009 (Spencer et al.), and in the modern eTHINK cohort in 2024 (Mrakotsky et al.). The consistency of this finding across studies in the modern treatment era suggests that the attention difficulties in this population are unlikely to be explained by factor replacement alone. Anxiety and depression also remain elevated, as registry data covering 3,246 people found higher psychotropic prescription rates across age groups and severity levels (Steen Carlsson et al.). Despite prophylaxis reducing the frequency of intracranial hemorrhages, they still continue to occur and, unfortunately, lead to neuropsychological damage. For example, a recent study, using neuroimaging alongside behavioral measures, identified both brain-based and behavioral factors linked to weaker cognitive performance in a modern cohort, which confirms that measurable neuropsychological risk has not disappeared under the modern standard of care (Cheung et al.).

During adolescence, adherence to prophylaxis often decreases as patients begin managing their own treatment. Studies of adolescents and young adults with hemophilia have found that this transition from parental care to self-management can make adherence more difficult (Hoefnagels et al.; Witkop et al.). Other barriers also include busy schedules, stress, forgetting treatment, and the demands of regular infusions (Hoefnagels et al.). However, these challenges are not unique to hemophilia. Studies of adolescents with other chronic illnesses have also found that peer relationships, increasing independence, and the responsibility of managing treatment can interfere with adherence (Pai and McGrady). Interventions designed specifically for adolescents with chronic illnesses have improved treatment adherence, suggesting that similar approaches may also benefit adolescents with hemophilia (Pai and McGrady). Poor adherence is

especially important because missing prophylaxis increases the risk of bleeding and joint damage (Königs et al.). Studies of adolescents with hemophilia A have also found that worse joint outcomes are associated with lower mood (Hassan et al.). Therefore, psychological well-being affects treatment adherence, and the physical consequences of poorer adherence further affect psychological health.

Discussion

Hemophilia is associated with elevated rates of ADHD, anxiety, and depression in children and adolescents; however, the magnitude of the association depends on the treatment era. For example, studies from the 1990s reported low neuropsychological scores, weak academic achievement, and adaptive skills below intelligence-predicted levels. Yet studies under modern treatment conditions, such as the eTHINK study, report largely normal development, and HIV is largely responsible for this difference (Mrakotsky et al.). Specifically, longitudinal analysis of the HGDS cohort found that declines in neuropsychological performance tracked declines in immune function, and once viral inactivation and recombinant products removed HIV from the treated population, the majority of the decline resolved (Loveland, J. A. Stehbens, et al.). However, ADHD remains an exception, as elevated rates were reported during the HIV era in the HGDS cohort and continue to appear in modern HIV-negative cohorts (Buranahirun et al.). After accounting for the confounding effects of HIV, attention difficulties remain one of the most consistent neuropsychological findings in pediatric hemophilia (Al-Huniti et al.).

Implications and Significance

The success of hemophilia treatment is typically measured by physical outcomes such as annual bleeding rates and joint health. However, the evidence reviewed in this paper suggests that these measures do not recognize the full burden of the disease in children and adolescents. The meta-analysis of 28 studies found that ADHD, anxiety, and depression are frequently underdiagnosed and undertreated despite being emphasized repeatedly throughout the literature (Al-Huniti et al.). Among these conditions, ADHD is the most impactful because it directly affects attention, learning, and academic performance. Multiple studies also identified higher rates of special education participation among children with hemophilia, suggesting that attention difficulties have educational consequences if untreated (Shapiro et al.). Unlike recurrent bleeding, these neuropsychological difficulties are not routinely assessed during hemophilia care.

Connection to Objectives

This review set out to determine what the published literature reports about ADHD, anxiety, and depression in children and adolescents with hemophilia, to evaluate how much of the older evidence can be attributed to hemophilia itself rather than to HIV, and to assess what risk remains under modern care. In this review, the first two objectives were achieved since the literature reports elevated rates of all three conditions, and the longitudinal analyses from the HGDS helped distinguish the effects of HIV from those of hemophilia. The third objective was

met only partially, as the eTHINK study provides strong evidence regarding cognitive and behavioral outcomes under modern treatment, but it is a single study, and it is cross-sectional. No study has followed a cohort of children treated with prophylaxis from early childhood through adolescence while measuring psychological outcomes. Residual risk under modern care can therefore be described but not precisely quantified. The consistency of the ADHD association among the treatment eras examined is the strongest finding of this review.

Recommendations

Routine mental health and attention screening should be incorporated into pediatric hemophilia care. A quality improvement study demonstrated that formal screening for inattention and hyperactivity can be integrated into existing hemophilia clinic visits (Boggs et al.). A Cochrane review of psychological interventions in hemophilia demonstrated that support programs improved self-management and confidence in children and adolescents; however, the authors rated the quality of the underlying evidence as low (Palareti et al.). Future research should prioritize longitudinal designs in modern cohorts. A similar study conducted in an HIV-negative population receiving prophylaxis could begin to provide evidence needed to address the questions that remain unresolved in this review. Future research should also standardize instrumentation, as variation in instrumentation methods made direct comparison of prevalence impossible.

Limitations

The studies included in this review differed in study design, sample size, and assessment methods, which made statistical pooling unfeasible. Additionally, since structured diagnostic interviews and symptom inventories identify psychological disorders differently, prevalence estimates from the two methods could not be compared directly. Rather than combining all of the prevalence estimates, the studies were grouped according to the assessment method they used. Sample sizes ranged from six to 551 participants, and the smaller studies provided less precise estimates than the larger ones. Studies from multiple treatment eras were intentionally included to compare outcomes before and after the introduction of viral inactivation and prophylaxis. Consequently, treatment practices, HIV exposure, and the diagnostic criteria used to identify ADHD and depression differed across studies. Changes to the diagnostic criteria for ADHD over time may also affect comparisons between older and more recent research. Nearly all participants were male because hemophilia is an X-linked disorder. Thus, these findings may not fully apply to symptomatic female carriers. Several studies were conducted at a single hemophilia treatment center, and some prevalence estimates were obtained from a previously published meta-analysis rather than the original studies. Finally, only English-language studies were included, and most modern evidence came from high-income countries. Because most modern studies come from high-income countries, these findings may not reflect outcomes in places where access to prophylaxis and comprehensive hemophilia care is limited.

Closing Thought

Much of the early literature on the neuropsychological effects of hemophilia was influenced by HIV contracted through contaminated treatment. As treatment improved and HIV transmission through clotting factor concentrates was eliminated, much of the apparent neuropsychological burden declined. However, children and adolescents with hemophilia continue to show elevated rates of attention difficulties compared with the general population. An ADHD prevalence of approximately 21 percent may not appear as alarming as the widespread cognitive impairments reported during the HIV era, but it still represents about one in five adolescents whose attention difficulties can affect academic performance if left unidentified. Like other chronic illnesses, hemophilia can affect more than physical health. Therefore, routine screening for attention and mental health difficulties should remain part of hemophilia care, even when the bleeding disorder itself is well controlled.

Works Cited

Abali, Osman, et al. "An Examination of the Symptoms of Anxiety and Parental Attitude in Children with Hemophilia." Turkish Journal of Medical Sciences, vol. 44, no. 6, 2014, pp. 1087–90. PubMed, https://doi.org/10.3906/sag-1310-82.

Al-Huniti, Ahmad, et al. "Mental Health Disorders in Haemophilia: Systematic Literature Review and Meta-Analysis." Haemophilia: The Official Journal of the World Federation of Hemophilia, vol. 26, no. 3, May 2020, pp. 431–42. PubMed, https://doi.org/10.1111/hae.13960.

Ayano, Getinet, et al. "The Global Prevalence of Attention Deficit Hyperactivity Disorder in Children and Adolescents: An Umbrella Review of Meta-Analyses." Journal of Affective Disorders, vol. 339, Oct. 2023, pp. 860–66. PubMed, https://doi.org/10.1016/j.jad.2023.07.071.

Boggs, Jacklyn E., et al. "Screening for Inattention, Hyperactivity and Impulsivity in Children with Haemophilia: A Quality Improvement Intervention." Haemophilia: The Official Journal of the World Federation of Hemophilia, vol. 29, no. 1, Jan. 2023, pp. 210–18. PubMed, https://doi.org/10.1111/hae.14656.

Buranahirun, Cathy, et al. "Neuropsychological Function in Children with Hemophilia: A Review of the Hemophilia Growth and Development Study and Introduction of the Current eTHINK Study." Pediatric Blood & Cancer, vol. 67, no. 1, Jan. 2020, p. e28004. PubMed, https://doi.org/10.1002/pbc.28004.

Burke, Tom, et al. "Clinical, Humanistic, and Economic Burden of Severe Hemophilia B in the United States: Results from the CHESS US and CHESS US+ Population Surveys." Orphanet Journal of Rare Diseases, vol. 16, no. 1, Mar. 2021, p. 143. PubMed, https://doi.org/10.1186/s13023-021-01774-9.

Canclini, M., et al. "Psychological Aspects and Coping in Haemophilic Patients: A Case-Control Study." Haemophilia: The Official Journal of the World Federation of Hemophilia, vol. 9, no. 5, Sept. 2003, pp. 619–24. PubMed, https://doi.org/10.1046/j.1365-2516.2003.00807.x.

Chen, Christina X., et al. "Economic Burden of Illness among Persons with Hemophilia B from HUGS Vb: Examining the Association of Severity and Treatment Regimens with Costs and Annual Bleed Rates." Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research, vol. 20, no. 8, Sept. 2017, pp. 1074–82. PubMed, https://doi.org/10.1016/j.jval.2017.04.017.

Cheung, Yin Ting, et al. "Clinical, Neuroimaging, and Behavioural Risk Factors for Neurocognitive Impairment in Chinese Patients with Haemophilia: A Multicentre Study." Haemophilia: The Official Journal of the World Federation of Hemophilia, vol. 29, no. 4, July 2023, pp. 1074–86. PubMed, https://doi.org/10.1111/hae.14816.

Ghanizadeh, A., and P. Baligh-Jahromi. "Depression, Anxiety and Suicidal Behaviour in Children and Adolescents with Haemophilia." Haemophilia: The Official Journal of the

World Federation of Hemophilia, vol. 15, no. 2, Mar. 2009, pp. 528–32. PubMed, https://doi.org/10.1111/j.1365-2516.2008.01971.x.

Hassan, T. H., et al. “Assessment of Musculoskeletal Function and Mood in Haemophilia A Adolescents: A Cross-Sectional Study.” Haemophilia: The Official Journal of the World Federation of Hemophilia, vol. 17, no. 4, July 2011, pp. 683–88. PubMed, https://doi.org/10.1111/j.1365-2516.2010.02475.x.

Hoefnagels, J. W., et al. “The Perspectives of Adolescents and Young Adults on Adherence to Prophylaxis in Hemophilia: A Qualitative Study.” Patient Preference and Adherence, vol. 14, 2020, pp. 163–71. PubMed, https://doi.org/10.2147/PPA.S232393.

Iorio, Alfonso, et al. “Establishing the Prevalence and Prevalence at Birth of Hemophilia in Males: A Meta-Analytic Approach Using National Registries.” Annals of Internal Medicine, vol. 171, no. 8, Oct. 2019, pp. 540–46. PubMed, https://doi.org/10.7326/M19-1208.

Kessler, Craig M. “Update on Liver Disease in Hemophilia Patients.” Seminars in Hematology, vol. 43, no. 1 Suppl 1, Jan. 2006, pp. S13-17. PubMed, https://doi.org/10.1053/j.seminhematol.2005.11.021.

Königs, Christoph, et al. “Teenagers and Adolescents with Hemophilia-Need for a Specific Approach.” Journal of Clinical Medicine, vol. 13, no. 17, Aug. 2024, p. 5121. PubMed, https://doi.org/10.3390/jcm13175121.

Koseoglu, Fatos D., et al. “Depression and Anxiety in Patients with Hemophilia A and B.” International Journal of Psychiatry in Medicine, vol. 59, no. 1, Jan. 2024, pp. 20–33. PubMed, https://doi.org/10.1177/00912174231196343.

Kvist, B., et al. “School Absences, School Achievements and Personality Traits of the Haemophilic Child.” Scandinavian Journal of Social Medicine, vol. 18, no. 2, June 1990, pp. 125–32. PubMed, https://doi.org/10.1177/140349489001800206.

Loveland, K. A., J. A. Stehbens, et al. “Declining Immune Function in Children and Adolescents with Hemophilia and HIV Infection: Effects on Neuropsychological Performance. Hemophilia Growth and Development Study.” Journal of Pediatric Psychology, vol. 25, no. 5, 2000, pp. 309–22. PubMed, https://doi.org/10.1093/jpepsy/25.5.309.

Loveland, K. A., J. Stehbens, et al. “Hemophilia Growth and Development Study: Baseline Neurodevelopmental Findings.” Journal of Pediatric Psychology, vol. 19, no. 2, Apr. 1994, pp. 223–39. PubMed, https://doi.org/10.1093/jpepsy/19.2.223.

Miles, B. S., et al. “Effect of Intracranial Bleeds on the Neurocognitive, Academic, Behavioural and Adaptive Functioning of Boys with Haemophilia.” Haemophilia: The Official Journal of the World Federation of Hemophilia, vol. 18, no. 2, Mar. 2012, pp. 229–34. PubMed, https://doi.org/10.1111/j.1365-2516.2011.02632.x.

Mrakotsky, Christine, et al. “The eTHINK Study: Cognitive and Behavioral Outcomes in Children with Hemophilia.” The Journal of Pediatrics, vol. 272, Sept. 2024, p. 114089. PubMed, https://doi.org/10.1016/j.jpeds.2024.114089.

Pai, Ahna L. H., and Meghan McGrady. “Systematic Review and Meta-Analysis of Psychological Interventions to Promote Treatment Adherence in Children, Adolescents, and Young Adults with Chronic Illness.” Journal of Pediatric Psychology, vol. 39, no. 8, Sept. 2014, pp. 918–31. PubMed, https://doi.org/10.1093/jpepsy/jsu038.

Palareti, Laura, et al. “Psychological Interventions for People with Hemophilia.” The Cochrane Database of Systematic Reviews, vol. 3, no. 3, Mar. 2020, p. CD010215. PubMed, https://doi.org/10.1002/14651858.CD010215.pub2.

Rodriguez-Merchan, E. Carlos. “The Cost of Hemophilia Treatment: The Importance of Minimizing It without Detriment to Its Quality.” Expert Review of Hematology, vol. 13, no. 3, Mar. 2020, pp. 269–74. PubMed, https://doi.org/10.1080/17474086.2020.1716726.

Shapiro, A. D., et al. “Defining the Impact of Hemophilia: The Academic Achievement in Children with Hemophilia Study.” Pediatrics, vol. 108, no. 6, Dec. 2001, p. E105. PubMed, https://doi.org/10.1542/peds.108.6.e105.

Sirois, P. A., et al. “Hemophilia Growth and Development Study: Relationships between Neuropsychological, Neurological, and MRI Findings at Baseline.” Journal of Pediatric Psychology, vol. 23, no. 1, Feb. 1998, pp. 45–56. PubMed, https://doi.org/10.1093/jpepsy/23.1.45.

Spencer, M. L. S., et al. “Inattention, Hyperactivity-Impulsivity, Academic Skills and Psychopathology in Boys with and without Haemophilia.” Haemophilia: The Official Journal of the World Federation of Hemophilia, vol. 15, no. 3, May 2009, pp. 701–06. PubMed, https://doi.org/10.1111/j.1365-2516.2009.01993.x.

Srivastava, Alok, et al. “WFH Guidelines for the Management of Hemophilia, 3rd Edition.” Haemophilia: The Official Journal of the World Federation of Hemophilia, 26 Suppl 6, Aug. 2020, pp. 1–158. PubMed, https://doi.org/10.1111/hae.14046.

Steen Carlsson, Katarina, et al. “High Use of Pain, Depression, and Anxiety Drugs in Hemophilia: More than 3000 People with Hemophilia in an 11-Year Nordic Registry Study.” Research and Practice in Thrombosis and Haemostasis, vol. 7, no. 2, Feb. 2023, p. 100061. PubMed, https://doi.org/10.1016/j.rpth.2023.100061.

Trzepacz, Angie M., et al. “Social, Emotional, and Behavioral Functioning of Children with Hemophilia.” Journal of Developmental and Behavioral Pediatrics: JDBP, vol. 24, no. 4, Aug. 2003, pp. 225–32. PubMed, https://doi.org/10.1097/00004703-200308000-00002.

Usner, D. W., et al. “Hemophilia Morbidity, Cognitive Functioning, and Academic Achievement.” The Journal of Pediatrics, vol. 133, no. 6, Dec. 1998, pp. 782–87. PubMed, https://doi.org/10.1016/s0022-3476(98)70151-9.

Witkop, M. L., et al. “Predictors of Non-Adherence to Prescribed Prophylactic Clotting-Factor Treatment Regimens among Adolescent and Young Adults with a Bleeding Disorder.” Haemophilia: The Official Journal of the World Federation of Hemophilia, vol. 22, no. 4, July 2016, pp. e245-250. PubMed, https://doi.org/10.1111/hae.12951.

Wodrich, D. L., et al. “Is Attention Deficit Hyperactivity Disorder Over-Represented among HIV-Seronegative Boys with Haemophilia? Preliminary Results from Our Centre.”

Haemophilia: The Official Journal of the World Federation of Hemophilia, vol. 9, no. 5, Sept. 2003, pp. 593–97. PubMed, https://doi.org/10.1046/j.1365-2516.2003.00790.x.

Zhou, Zheng-Yi, et al. “Burden of Illness: Direct and Indirect Costs among Persons with Hemophilia A in the United States.” Journal of Medical Economics, vol. 18, no. 6, June 2015, pp. 457–65. PubMed, https://doi.org/10.3111/13696998.2015.1016228.

Assessing Flood Mitigation Strategies for Urban Resilience: A Comparative Review of New York City and HoustonBy Eva Pirolli

Abstract

As climate change intensifies extreme precipitation and urbanization continues to increase, flooding has become one of the most pressing challenges facing cities across the United States. By removing soil and vegetation, expanding populations, and decreasing urban space, cities are becoming exacerbated by an anticipated increase in the number of intense rainfall events. This paper conducts a comparative review of the flood mitigation strategies in New York City and Houston, two cities with distinctly different geographic and climatic characteristics. Drawing on peer-reviewed research, government reports, and engineering studies, the review evaluates traditional gray infrastructure, green infrastructure, and emerging flood adaptation technologies based on measures including runoff reduction, peak flow reduction, implementation costs, environmental impacts, and long-term resilience. New York City's findings demonstrate that distributed green infrastructure and neighborhood-scale stormwater systems are particularly effective in reducing localized flood risk within a dense urban environment. Houston's findings demonstrate that large-scale detention, reservoirs, and conveyance systems can manage substantially greater volumes of water but remain vulnerable when extreme rainfall exceeds their design capacity. The comparison therefore shows that neither centralized gray infrastructure nor distributed green infrastructure is independently sufficient. Flood mitigation is most effective when infrastructure is matched to local land availability and urban development patterns. For both the research field and society, the results suggest that future strategies must determine where water can be safely managed and where preserving floodplains or limiting development offers a more sustainable solution.

Introduction

Over the past 50 years, flooding has become increasingly prominent in many urban cities across the United States. Urbanization has caused many challenges for communities to control and mitigate flood risks. As population density increases and land becomes more scarce, urban populations become particularly exposed, creating zones of vulnerability for flooding. Additionally, climate change is also contributing to these threats. As the Earth warms due to an intensified greenhouse effect, precipitation is becoming more extreme and frequent. In the United States, nine of the top ten years for extreme one-day precipitation have occurred since 1995 (CAP 2025). Economic risks are also affected. Recent major floods have impacted communities, turning recovery from these events into regular financial shocks. According to the National Oceanic and Atmospheric Administration (NOAA), an average flooding event costs approximately $4.5 billion (National Centers for Environmental Information 2026). These losses are not staggering, instead they are becoming more and more costly as these repairs accumulate.

Studies and models reinforce that cities are currently using many different strategies to manage and reduce floodwater. Cities in the East coast are actively utilizing more green

infrastructure and sewer systems, while the South and West coast have the advantage of more open land, allowing for different strategies including levees and basins to block in the water accumulated. Furthermore, programs like the National Flood Insurance Program (NFIP) have historically encouraged rebuilding in flood-prone areas following flooding events, which actually increases long-term vulnerability of the land. However, newer policies are now approaching these threats with other methods- including wetland restoration and restrictions on development in high-risk zones. Despite this progress, there is still an important gap in understanding the right combination of collection systems, redirection systems, and green infrastructure in contrasting cities like New York City and Houston.

To better examine how flood mitigation should be approached and adapted to different urban environments, this paper analyzes New York City and Houston. New York City experiences increasingly frequent intense "flash floods" with torrential downpours. These storms can easily overwhelm crowded cities like NYC through overflowing streets and flooding transit systems across boroughs. With about 70% of the surface now paved with impervious surfaces, heavy rainstorms become detrimental threats to the city's infrastructure and citizens (New York City Department of Environmental Protection 2026). For Houston, data shows that rainfall intensity is on the rise as it continually increased by 22% since 1970 (Climate Central 2026). A city like Houston, experiences factors like urban sprawl, clay soil, and topography of the region has led to greater risk in flooding. In addition, Houston is highly vulnerable to hurricanes due to its geographic location on the Texas Gulf Coast, which exposes many zones up to a full year's worth of water in just a few days (translating to 50+ inches). These cities were selected because they represent distinct geographic, climatic, and economic conditions in the United States. Accordingly, this study addresses the following research question: How should flood mitigation strategies be adapted to the geographic, climatic, and urban characteristics of New York City and Houston to maximize long-term effectiveness and sustainability?

Specifically, this paper evaluates the current mitigation practices, proposed future approaches, and associated economic, political, and environmental trade-offs within each city. Through this comparative analysis, the paper seeks to determine how flood mitigations strategies can be adapted to local conditions to improve long-term urban resilience. This paper begins by examining the different types of flood mitigation strategies and the factors that influence their implementation. It then reviews the current mitigation efforts, proposed future approaches, and limitations in the cities of New York City and Houston. Finally, the analysis compares the effectiveness, costs, land requirements, and environmental impacts of these strategies to identify how flood mitigation can be most effectively adapted to local conditions.

Background

Flood mitigation refers to the proactive actions taken to reduce long-term risk of damage to human life, infrastructure, and environment. Rather than responding to flood threats, it is a preventive approach that is designed to lessen flood risks. This is done through reducing runoff, increasing water storage, or limiting exposure in vulnerable zones. Flood mitigation strategies

fall into 3 main categories based on how they manage stormwater: collection systems, redirection systems, and green infrastructure. Collection strategies are designed to capture and relay runoff before it accumulates in threatened areas. Redirection systems follow the process of moving water away from flood-prone zones. In contrast, green infrastructure takes a different approach, it works by restoring hydrologic processes through features that increase infiltration, reduce runoff and peak flow, and water quality improvement.

Collection systems are a highly effective method at storing and conveying water to prevent damage to cities. In fact, it can save up to 92% of flood damage costs (Sohn et al., 2020). A few common examples of collection systems include detention basins, retention basins, and underground storage tanks. Detention basins, also known as dry ponds, are "vegetated depressions designed to hold runoff from impermeable surfaces and allow the settling of sediments and associated pollutants"(The European Commission, 2022). These basins are more successful at reducing runoff volume, peak, time to peak, and hydrograph slope in lesser storms (Hoss et al., 2016). In addition, low-lying areas can serve as natural sources of detention basins to store water during extreme flood events (Jayawardena & van Roon, 2017). The limitation for this technology is that it does not significantly reduce the total runoff volume. Studies show that detention ponds slow down the peak flow of storm runoff by an average of only 0.3% (Emerson et al., 2025). For instance, retention basins also known as stormwater ponds perform the opposite function of holding water permanently long-term instead of short term hold before release like the detention basin functionality. Ponds have been regarded as the best stormwater management BMP alternatives for reducing flood peaks (Ayalew et al., 2015). However, the limitation of retention basins is that they require a massive footprint of approximately over 5 acres to maintain a permanent body of water while accommodating more stormwater. Maintenance for retention basins is also extensive compared to other strategies since they require routine sediment dredging every 5-15 years. Underground storage tanks store stormwater below ground in urban areas, a beneficial strategy for areas that have limited space, often paired with pumping systems. The goal for this system is to maximize storage in dense cities; however, it does have the limitation of high construction and maintenance costs associated.

Redirection systems move water away efficiently and reduce flooding risks in vulnerable areas. Notable examples of this strategy include storm drains and sewer networks, flood bypass channels, and culverts/canals. Sewer networks utilize pipes and inlets to collect runoff from streets, carrying water to treatment plants or discharge points. Sewer networks have been a common methodology to combat flooding, however, in recent years many cities are pulling away from this technique as it proves inefficient for major flood mitigation. Flood bypass channels are artificial channels designed to divert excess floodwater. This system has proved to be controversial in its effectiveness as many geographical factors ultimately dilute its performance. Culverts and canals carry water under roads or railways and canals redirecting the flow of water across regions. This system is often associated with highly effective and cost-efficient buried structures preventing road erosion during extreme flood events. Furthermore, its hydraulic design including circular pipes provide maximum hydraulic efficiency for handling smaller water flows

while larger structures such as Arch Culverts manage the larger water volumes allowing for a flexible approach to flood mitigation.

Green infrastructure can mitigate flood risks by reducing stormwater runoff and protecting floodplains in a sustainable approach. Unlike conventional gray infrastructure, which primarily focuses on rapidly conveying water away from existing infrastructure, nature-based strategies can reduce runoff closer to where precipitation occurs. Salient examples include bioswales, rain gardens, and permeable pavement. Bioswales are landscape channels filled with vegetation and serve its primary purpose in slowing down runoff and filtering pollutants. Bioswales are regarded as highly efficient with its design mimicking the one of a wetland environment, trapping heavy metals and filtering out toxic chemicals through engineered soil and deep-rooted plants. This initiative is especially rewarding in urban areas that have many impervious surfaces and high traffic. Rain Gardens act as shallow planted depressions that absorb stormwater, designed for residential or urban areas. Because they utilize specialized porous soils and native plants, this system absorbs approximately 30% more water than an equivalent area of conventional lawn (Penn State Extension, 2024; Pacific Northwest National Laboratory). Green Roofs are vegetated layers located on roofs of buildings. A global synthesis of 75 peer-reviewed studies found that green roofs retained an average of 62% of rainfall, while individual experimental studies have reported retention rates nearing 80% depending on factors such as substrate depth and vegetation (VanWoert et al., 2005). This strategy not only consists of waterproofing but it also takes on the job of nourishing plant growth, showcasing high efficiency that surpasses its initial high cost of implementation.

Review of Flood Mitigation in New York City

New York City has increasingly shifted from traditional reliance on gray infrastructure toward a more integrated approach, combining engineered flood protection with green and nature-based infrastructure (New York City Department of Environmental Protection, 2024). This transition is especially important because approximately 70% of New York City's surface is covered by impervious materials, limiting the amount of rainfall that can infiltrate into the ground (NYC Department of Environmental Protection, 2023). As a result, this increases pressure on the city's flood exposure: nearly 440,000 New Yorkers, 14,500 businesses, and approximately $250 billion in property value are located within the city's 100-year floodplain (City of New York, 2026). Consequently, New York City's mitigation strategy must address both the largest volume of stormwater produced by intense rainfall and the extensive physical and economic exposure created by its dense coastal development. Rather than relying on one system, the city has developed large-scale coastal protection, stormwater storage, natural drainage, and decentralized runoff-reduction programs.

Current Flood Mitigation Efforts

One of New York City's largest current flood-mitigation efforts is the East Side Coastal Resiliency (ESCR) Project, which was developed to combat the devastating impacts made by

Hurricane Sandy in 2012. It is designed to protect Lower Manhattan from future coastal flooding. The project creates a 2.4-mile flood-protection system along Manhattan's East Side using raised parkland, floodwalls, berms, 18 swinging or sliding floodgates, and upgraded underground drainage infrastructure (New York City Mayor's Office, 2024). At a total cost of $1.45 billion, the ESCR Project represents an investment of approximately $604 million per mile of protected coastline, based on the project's total cost and length. The system is designed to protect more than 110,000 residents and billions of dollars of infrastructure and property from future flood events (New York City Mayor's Office, 2024). The first completed section cost $163 million and was finished $10 million under its original projected budget, demonstrating that large-scale flood infrastructure can achieve financial efficiencies even at substantial overall investment (New York City Mayor's Office, 2024). The project's physical scale is especially significant because it creates a line of continuity in its protection across such a dense urban zone. Unlike stormwater-storage strategies, this ESCR project primarily prevents coastal floodwater from entering vulnerable neighborhoods through physical barriers that could otherwise become trapped behind those barriers. This distinction is imperative because New York's coastal flood risk requires addressing both storm surge and urban drainage. Research on coastal flood-risk management in New York has similarly emphasized that effective adaptation requires considering multiple strategies and their trade-offs rather than relying on a single intervention (Ceres et al., 2020). Consequently, ESCR establishes why densely developed coastal areas may require expensive structural protection when large-scale natural buffers are constrained by existing infrastructure.

A second major effort is the New York City Cloudburst Program, which targets short-duration, high-intensity rainfall that can overwhelm conventional sewer systems. Traditional gray stormwater infrastructure was primarily designed to convey runoff through centralized networks of pipes, drains, channels, and pumps; however, many existing systems were constructed for smaller populations, more permeable urban landscapes, and historical climate conditions (Price et al., 2021). As urban development has grown exponentially, it has increased impervious surfaces resulting in a need for a more sufficient long-term solution. Cloudburst infrastructure provides a complimentary approach by combining surface and underground systems to manage excess stormwater during flood events. Research on Copenhagen's Cloudburst Management Plan demonstrates how designated surface flow pathways can supplement conventional underground drainage systems, allowing cities to create additional capacity for extreme storms without relying only on sewer networks (Lerer et al., 2017). New York City implements this approach with a variety of scales depending on the characteristics of the drainage problem. For instance, the city's first Cloudburst project at the NYCHA South Jamaica Houses cost approximately $6 million and can capture 330,000 gallons during a rain event. In addition, it can manage up to 3.5 million gallons annually (New York City Department of Environmental Protection, 2026). Furthermore, the Homecrest Cloudburst project is a notable example of a planned effort costing $95 million. It is expected to manage approximately 30 million gallons annually and designed for downpours of more than two inches

of rain in a single hour (New York City Department of Environmental Protection, 2026). The Homecrest project has an expected annual management capacity of about 8.6 times greater than the South Jamaica project while costing approximately 15.8 times as much. This variation demonstrates that Cloudburst infrastructure can be scaled according to the size and intensity of the drainage problem. It also highlights a distinction between Cloudburst infrastructure and the East Side Coastal Resiliency Project. While the ESCR project is primarily evaluated by the population and area protected from flooding, the Cloudburst projects can be evaluated by their capacity to capture and manage stormwater during intense rainfall events.

Another notable effort New York City has invested heavily in is the Staten Island Bluebelt Program. This project integrates wetlands, streams, ponds, and other natural drainage systems to collect and convey stormwater (New York City Department of Environmental Protection, 2026). Research on nature-based infrastructure finds that these systems can provide multiple benefits beyond flood mitigation. This includes water-quality improvements, habitat creation, and the preservation of natural processes, making them particularly valuable in areas where sufficient land is available (Esraz-Ul-Zannat et al., 2024). The Bluebelt approach is particularly valuable because it can use large areas of land to manage stormwater whilst also providing ecological benefits. For instance, the expanded New Creek Bluebelt cost approximately $110 million, covers 94 acres, and drains a much larger 2,249 acre watershed (New York City Department of Environmental Protection, 2026). This means that the Bluebelt's drainage area is approximately 24 times larger than the physical area occupied by the Bluebelt. Rather than requiring the entire drainage area to be converted into flood infrastructure, the strategy uses strategically located natural drainage pathways like wetlands to receive and slow stormwater. A separate $14.3 million Bluebelt in Midland Beach covers approximately 3 acres while receiving stormwater from 111.5-acre drainage acres (New York City Department of Environmental Protection, 2026). This means that the drainage area is approximately 37 times larger than the Bluebelt itself. Peer-reviewed research specifically examining the Staten Island Bluebelt has found that the program combines stormwater infrastructure with the preservation of natural wetlands and drainage corridors, providing both flood-management and ecological benefits (Pickett et al., 2024). However, the effectiveness of this approach depends on the availability of suitable land available. The contrast between Staten Island and Manhattan reinforces the importance of geographic conditions when selecting flood mitigation strategies.

In addition to these large-scale projects, New York City has implemented its Green Infrastructure Program to reduce the amount of stormwater entering the combined sewer system. Unlike ESCR and Bluebelts, which protect larger areas through centralized or landscape-scale infrastructure, green infrastructure attempts to manage rainfall at or near its source (New York City Department of Environmental Protection 2026). The physical scale of individual projects can be relatively small, but their cumulative effects are significant. For example, a peer-reviewed study of bioswales retained approximately 40% of the stormwater it received from a drainage area 231 times its size while also reducing overall nutrient pollution (from 7.7 to 6kg nitrogen per year) by decreasing combined sewer overflow volumes (Shetty et al., 2019). Similarly, a

$16.6 million porous-pavement project on Beach 108th Street in Rockaway is expected to allow nearly 1.3 million gallons of stormwater to infiltrate into the ground each year (New York City Department of Environmental Protection, 2023). More recently, NYC committed $32.6 million to install seven miles of porous pavement in Kensington and Borough Park, with six of the seven miles completed by December 2025 (New York City Department of Environmental Protection, 2025). These projects demonstrate that green infrastructure can manage stormwater without requiring large underground storage structures or expanded drainage channels. However, the effectiveness of green infrastructure cannot be represented by one universal percentage. NYC's 2025 Green Infrastructure Annual Report calculates stormwater capture using practice-specific equivalency rates and then converts captured stormwater into estimated combined sewer overflow reductions using drainage-area-specific rates. This means that the same type of green infrastructure can produce different estimated reductions depending on its location, size, surrounding drainage area, and connection to the sewer system. Green infrastructure is therefore most effective as a distributed network of smaller interventions rather than as a replacement for large-scale coastal or watershed infrastructure.

Plans in Place and Upcoming

New York City's future flood mitigation plans increasingly focus on expanding the scale and capacity of the strategies already being implemented. In 2026, the city announced a $95 million Cloudburst project in Homecrest that will use porous pavement and underground storage to manage approximately 30 million gallons of stormwater annually (New York City Department of Environmental Protection, 2026). The city is also continuing to expand Bluebelt infrastructure, with more than 25 Bluebelt projects currently under way and nearly $55 million allocated for a new Kissena Park Bluebelt. DEP modeling found that the proposed Kissena Park Bluebelt could cost approximately $40 million, mitigating flooding across a 90-acre area and providing relief to an estimated 3,000 residents (New York City Department of Environmental Protection, 2024). These projects indicate that NYC's future strategy is moving toward a more distributed system in which multiple neighborhood-scale interventions supplement major coastal defenses. Research similarly recommends increasing the flood-management capacity of green infrastructure. These modifications would include deeper surface depressions, more hydraulically efficient inlets, and connections to subsurface storage. In areas with high water tables, low-permeability soils, or shallow bedrock, these practices would also likely need to be lined and connected via underdrains to local catchbasins. To overcome the perception that enhanced Green Infrastructure facilities are overdesigned, and to justify the additional costs associated with their construction and maintenance, the hybrid role intended would need to be recognized formally and encoded in new interagency agreements (Rosenzweig et al., 2024). The overall trend suggests that New York City is not replacing gray infrastructure. Instead, it is increasing the number of scales at which stormwater is managed.

Current Issues and Limitations

Despite significant investment, New York City's current flood mitigation system continues to face important limitations because the city experiences multiple forms of flooding that cannot always be addressed with the same type of infrastructure. A comprehensive peer-reviewed assessment identifies several distinct flood hazards affecting New York City, including rainfall-driven flooding, coastal flooding, groundwater flooding, and compound flooding, demonstrating that flood risk cannot be understood as a single, uniform problem (Rosenzweig et al., 2024). Groundwater flooding presents a particularly difficult challenge because it occurs when elevated groundwater levels cause water to enter buildings through foundations, basements, or floor drains; consequently, increasing the capacity of surface drains or sewer systems may not eliminate flooding caused by water rising from below ground (Rosenzweig et al., 2024). This concern has become increasingly significant as sea-level rise and changing precipitation patterns can raise groundwater levels in coastal areas. Recent research by Glas et al. found that shallow groundwater conditions can occur alongside heavy precipitation and storm surge in coastal New York, meaning that groundwater can contribute to flood hazards simultaneously with other drivers rather than operating as an entirely separate risk (Glas et al., 2026). In response to these unresolved risks, New York City has begun developing additional groundwater-risk planning measures, including proposals for groundwater flood-risk mapping. Together, this research suggests that expanding surface drainage infrastructure alone cannot address all of New York City's future flood risks.

Another major limitation is compound flooding, in which multiple hazards occur simultaneously and can produce greater damage than any individual flood mechanism alone. Research found that the joint occurrence of storm surge and heavy precipitation has increased in many major U.S. coastal cities and specifically attributed New York City's increase in compound events to changing storm-surge weather patterns that also favor heavy precipitation (Wahl et al., 2015). More recent NYC-specific research similarly found that sea-level rise substantially increases the hazard associated with compound rainfall and storm-surge flooding and projected that the probability of destructive Sandy-like compound flooding could increase by as much as five times by the end of the century (Sarhadi et al., 2024). These findings demonstrate an important limitation of infrastructure designed for only one flood mechanism: a coastal barrier may reduce storm-surge exposure while intense rainfall simultaneously overwhelms the sewer system, creating flooding from another direction. Research examining NYC's drainage system similarly concludes that high sea-level conditions can compound rainfall-driven urban flooding, increasing the vulnerability of coastal areas to pluvial flooding (Madaus et al., 2024). Consequently, future flood infrastructure must increasingly account for interactions between multiple hazards rather than treating rainfall, coastal surge, and groundwater as independent problems.

Finally, research on flood resilience increasingly demonstrates that technical infrastructure alone cannot determine whether adaptation strategies are effective because flood risks and vulnerabilities vary substantially between neighborhoods. New York City's Jewel Streets neighborhood illustrates this problem: the area experiences chronic flooding related to its

low elevation, high water table, and inadequate stormwater and sanitary sewer infrastructure, conditions that differ significantly from those affecting densely developed coastal neighborhoods (New York City Department of Housing Preservation and Development, 2025). The city's planning process therefore incorporated residents and community organizations into a yearslong planning effort, reflecting the need to identify local conditions that may not be apparent through citywide infrastructure planning alone (New York City Department of Housing Preservation and Development, 2025). This approach is consistent with broader flood-resilience research emphasizing that adaptation strategies must account for the social and geographic characteristics that shape local vulnerability rather than applying identical solutions across an entire city. The creation of the Southeast Queens Flooding Adaptation Task Force similarly reflects the recognition that chronic groundwater and stormwater flooding require coordination and planning tailored to neighborhood-specific conditions. Ultimately, New York City's remaining challenge is not simply to construct more flood-control infrastructure, but to determine the appropriate combination of systems, scale, location, and community involvement for each type of flood risk.

Review of Flood Mitigation in Houston

Houston's flood mitigation strategy differs substantially from New York City's because the region has greater land availability and exposure to both extreme rainfall and hurricanes. Houston is characterized by flat terrain, low-infiltration soils, extensive impervious surfaces, and a subtropical climate that exposes the region to severe rainfall events, making it particularly vulnerable to flooding. As a result, Houston has historically emphasized large-scale collection and redirection systems. This includes detention basins, reservoirs, channel improvements, and drainage networks that temporarily store or convey stormwater away from developed areas. Research examining Houston’s flood-control infrastructure found that the region has become highly engineered around channelization and detention, although these systems have been challenged by increasingly severe rainfall events (Peeples et al., 2023). The scale of Houston’s investment reflects the magnitude of its flood risk: in 2018, Harris County voters approved a $2.5 billion flood bond to finance flood damage reduction projects (Harris Country Flood Control District, 2018).

Current Flood Mitigation Projects

One of Houston's largest completed flood mitigation efforts is Project Brays, a more than 30-year partnership between the Harris County Flood Control District and the U.S. Army Corps of Engineers. Completed in 2022 at a total cost of approximately $480 million, the project combines channel conveyance improvements and stormwater detention to reduce flood risk along Brays Bayou(Harris County Flood Control District, 2022). The completed system includes 21 miles of channel improvements extending from the Houston Ship Channel to Fondren Road and four stormwater detention basins. Together, the basins cover more than 800 acres and can store approximately 3.5 billion gallons of stormwater (Harris County Flood Control District, 2022). Project Brays therefore demonstrates the benefits of combining redirection and collection

rather than relying on a single flood-control method. Channel improvements increase the amount of water that can be conveyed through the watershed. In contrast, the detention basins act as temporary storage of excess stormwater and reduce peak flows. Peer-reviewed research examining Project Brays found that the completed improvements can substantially reduce flood levels, showing water-surface elevations decreasing by approximately 1.2 meters in the upstream half of Brays Bayou during an Atlas 14 100-year storm (Peeples et al., 2023). The economic benefits of Houston's flood-control infrastructure is also substantial. During the May-June 2015 flooding, the U.S. Army Corps of Engineers estimated that the combined Houston Flood Channel improvements along Brays and Sims Bayous prevented approximately $4.3 billion in flood damages (U.S. Army Corps of Engineers, 2015). However, these benefits come with significant construction and land costs, and channel expansion alone cannot eliminate flood risk when rainfall exceeds the capacity of the system. Therefore, Project Brays illustrates both the value and limitations of large-scale engineered conveyance.

Addicks and Barker Reservoirs

A second major effort is the Addicks and Barker Reservoirs. Constructed by the U.S. Army Corps of Engineers in response to catastrophic flooding during 1929 and 1935, the reservoirs collect rainfall west of the city and release stored water in a controlled manner into Buffalo Bayou. Together, the reservoirs cover watersheds of approximately 136 and 130 square miles, respectively, and were specifically designed to reduce downstream flood damage (U.S. Army Corps of Engineers). Their original construction cost was approximately $4 million, although maintaining and modernizing the aging infrastructure has required substantial additional investment (U.S. Army Corps of Engineers). Both reservoirs are normally kept largely empty to preserve their capacity to capture water during major rainfall events before gradually releasing it once downstream conditions allow (U.S. Army Corps of Engineers, 2020).

The reservoirs demonstrate the long-term economic value of stormwater storage. The Army Corps has estimated that Addicks and Barker have prevented more than $6.75 billion in flood damage since their completion. During the May-June 2015 flood event alone, the reservoirs were estimated to prevent approximately $2.1 billion in damages (U.S. Army Corps of Engineers). However, Hurricane Harvey demonstrated that even major reservoirs have limits when rainfall exceeds their design and operating capacity. Modernization has therefore become an important component of Houston's strategy; a $75 million contract upgraded the outlet control structures at Addicks and Barker to improve the resilience and operation of the dams (U.S. Corps of Engineers, 2020). The reservoirs therefore remain an important part of Houston's flood protection system, but their limitations demonstrate why the city must combine regional storage with additional detention, conveyance, and nonstructural approaches.

Plans in Place and Upcoming

Houston's future flood-mitigation plans increasingly focus on expanding the regional system established through the 2018 Flood Bond while continuing to develop additional capacity

for flood-damage reduction. Planned and ongoing projects include channel-conveyance improvements, new and expanded stormwater detention facilities, drainage improvements, and modernization of existing flood-control infrastructure (Harris County Flood Control District, 2026). The Flood Control District's capital-improvement strategy specifically combines increased conveyance capacity in bayous and drainage channels with the excavation and expansion of detention basins, which temporarily store excess stormwater and gradually release it after the immediate flood threat has passed (Harris County Flood Control District, 2026). This approach manages floodwater at multiple locations within a watershed rather than relying exclusively on downstream channels or reservoirs. For example, current and planned bond-funded projects include combinations of regional detention and upstream channel improvements, reflecting a strategy of increasing storage and conveyance capacity throughout the larger drainage system (Harris County Flood Control District, 2026).

Houston is also beginning to complement its traditional gray infrastructure with green and nature-based approaches that manage runoff closer to its source. Green infrastructure can reduce the volume of rainfall entering the drainage network and therefore decrease the amount of water that must ultimately be conveyed or stored by larger detention and channel systems. A peer-reviewed modeling study of proposed green infrastructure in Houston's Sunnyside neighborhood found that the interventions reduced average annual runoff by approximately 3%, decreasing modeled runoff volume from 15,169,032 cubic feet to 14,661,951 cubic feet (Newman et al., 2022). During a modeled 100-year storm, however, the same green-infrastructure practices reduced runoff by approximately 37%, from 401,341 cubic feet to 252,577 cubic feet (Newman et al., 2022). The difference between these results demonstrates that green infrastructure's performance can vary substantially depending on the magnitude and characteristics of the rainfall event being evaluated. The Sunnyside study therefore provides evidence that strategically distributed green infrastructure can reduce runoff during extreme storms while complementing the larger storage and conveyance capacity provided by Houston's regional flood-control system. Together, these developments suggest that Houston's future strategy is increasingly characterized by layered infrastructure, with green infrastructure managing runoff closer to where it falls while detention basins, reservoirs, and channels provide the larger-scale storage and conveyance capacity needed across the watershed.

Current Issues and Limitations

Despite Houston's extensive investment in flood mitigation, the region continues to face significant limitations because the scale and intensity of rainfall can exceed the capacity of engineered systems. Hurricane Harvey demonstrated this problem particularly clearly: extreme rainfall filled Addicks and Barker Reservoirs to unprecedented levels, illustrating that even large regional storage systems have finite capacity. Peer-reviewed climate-attribution research found that human-induced climate change likely increased the intensity of Hurricane Harvey's precipitation, with one analysis estimating that global warming made the rainfall approximately 15% more intense and an event of this magnitude roughly three times more likely (van

Oldenborgh et al., 2017). A separate analysis estimated that human-induced climate change increased the probability of the observed precipitation by at least a factor of 3.5 and increased precipitation amounts by at least 18.8%, with a best estimate of 37.7% (Risser & Wehner, 2017). These findings illustrate a fundamental limitation of infrastructure designed around finite storage and conveyance capacity: while reservoirs and detention basins can substantially reduce flood risk, they cannot eliminate flooding when rainfall volumes exceed the capacity available throughout the watershed. As extreme precipitation increases, infrastructure designed using historical rainfall conditions may therefore require increasingly frequent and expensive expansion to maintain the same level of protection.

A second limitation is the interaction between urban development and runoff. Houston's large regional watersheds allow the construction of extensive detention and conveyance infrastructure, but continued urbanization can simultaneously increase the amount and speed of runoff entering those systems. A study found that urbanization exacerbated both Hurricane Harvey's rainfall and its flood response in Houston, demonstrating that development can influence flood risk through more than simply the replacement of permeable land with impervious surfaces (Zhang et al., 2018). The conversion of forests, grasslands, and other vegetated land into roads, parking lots, and buildings reduces infiltration and increases the amount of rainfall that becomes surface runoff, placing additional pressure on channels and detention systems (Bidadian et al., 2025). This creates a continuing challenge for Houston's traditional approach: as development increases runoff, additional drainage infrastructure may be needed to manage the greater volumes of water produced by the urban landscape. Research therefore suggests that managing flood risk requires consideration of both the infrastructure used to store and convey water and the land-use changes that determine how much runoff enters the system.

A third limitation is that Houston's flood risk and the benefits of mitigation are not distributed equally across the region. Location within a watershed, elevation, historical development patterns, and access to infrastructure can affect the level of flood protection experienced by different communities. Research examining floodplain buyouts in the Houston area found that buyouts can produce broader patterns of displacement and disinvestment rather than functioning solely as a technical solution to flood risk (Nance et al., 2024). The researchers found that socially vulnerable and predominantly Black and Latinx communities were more likely to experience patterns in which property buyouts coincided with comparatively limited infrastructure investment, while wealthier and whiter areas often received greater infrastructure investment (Nance et al., 2024). More recent research examining twenty-five years of floodplain buyouts in Harris County also found that buyouts contributed to demographic change within affected neighborhoods, with the strongest effects occurring in wealthier neighborhoods with larger numbers of buyouts (Seong et al., 2025). These findings demonstrate that voluntary buyouts can reduce exposure to repeated flooding while also producing consequences for neighborhood stability, displacement, and demographic composition. Consequently, Houston's future flood management challenge is not only determining how much infrastructure to

construct, but also considering where investments are located, which communities receive protection, and how mitigation strategies affect the neighborhoods in which they are implemented.

Analysis

Runoff and Peak-Flow Reduction

The comparison between New York City and Houston shows that both cities increasingly use green infrastructure to reduce runoff, but the scale and role of these systems differ substantially. A 2025 meta-analysis of 50 field-scale bioretention systems found average reductions of approximately 63% in stormwater runoff volume and 74% in peak flow, demonstrating the potential of source-level infrastructure to reduce the amount and intensity of water entering conventional drainage systems (Sabbagh et al., 2025). In New York City, this approach is particularly valuable because limited space requires stormwater to be managed within existing streets, parks, and public spaces. The city's Cloudburst projects illustrate how this strategy can be scaled beyond individual green-infrastructure installations: South Jamaica can capture approximately 330,000 gallons during an individual rainfall event and manage up to 3.5 million gallons annually, while the planned $95 million Homecrest project is expected to manage approximately 30 million gallons annually (New York City Department of Environmental Protection, 2025, 2026). Houston's Sunnyside study, however, demonstrates a different pattern: modeled green infrastructure reduced average annual runoff by only approximately 3%, but reduced runoff by approximately 37% during a modeled 100-year storm (Newman et al., 2022). This suggests that green infrastructure can provide meaningful peak-flow reductions during extreme events even when its effect on total annual runoff is relatively modest. Overall, the evidence indicates that NYC's distributed systems are particularly suited to managing stormwater within highly developed areas, while Houston's green infrastructure functions more effectively as a supplementary layer to its much larger reservoirs, detention basins, and channel network; neither city's green infrastructure is intended to replace its larger-scale flood-control system.

Water-Management Capacity

The difference between Houston and New York City is not simply the amount of water each city can store, but how effectively storage capacity can be distributed across the urban landscape. Research on Houston's Buffalo Bayou watershed found that storage-based drainage strategies were substantially more effective at reducing flood losses than conveyance-based approaches, suggesting that creating space for water before it reaches vulnerable areas can provide greater protection than relying primarily on channels to move water away (Sohn et al., 2020). However, Houston's experience also demonstrates that increasing storage alone does not eliminate risk. The extreme rainfall associated with Hurricane Harvey exceeded the capacity of existing systems, while research has shown that climate change is increasing the intensity and likelihood of extreme rainfall in the region (van Oldenborgh et al., 2017; Risser & Wehner,

2017). This suggests that Houston's large-scale storage strategy is highly effective under conditions within its design range, but its long-term resilience depends on continually reassessing whether available capacity remains sufficient as rainfall extremes change. NYC faces a different challenge: its limited land availability makes it difficult to expand storage through large centralized facilities, meaning that distributing smaller storage opportunities throughout the existing city may be more practical than attempting to replicate Houston's model. Research on urban stormwater systems supports this distributed approach, finding that smaller bioretention systems can achieve substantial reductions in runoff and peak flow when appropriately designed and located (Sabbagh et al., 2025). The broader takeaway is therefore that storage capacity should not be evaluated by volume alone. Houston's centralized system provides greater absolute capacity, but NYC's distributed approach may offer greater adaptability within a dense urban environment; for both cities, the most resilient strategy is one that places storage where it can most effectively reduce downstream flood exposure while allowing capacity to be expanded or adapted as future flood conditions change.

Economic Tradeoffs

Houston's flood-control strategy demonstrates the economic advantages of investing in large-scale infrastructure when substantial areas and high concentrations of development are exposed to flooding. Research on the Buffalo Bayou watershed found that storage-based drainage systems produced substantially greater reductions in flood losses than conveyance-based approaches, indicating that strategically increasing storage can generate significant economic benefits by preventing damage before floodwaters reach vulnerable areas (Sohn et al., 2020). Houston's experience also shows, however, that avoided damages should not be treated as guaranteed returns on investment. The billions of dollars in damages prevented by the region's flood-control infrastructure demonstrate its value under major storm conditions, but Hurricane Harvey showed that an event exceeding existing system capacity can still produce catastrophic losses (U.S. Army Corps of Engineers, 2015; van Oldenborgh et al., 2017). This means Houston's cost-effectiveness depends not only on how much damage infrastructure prevents today, but on whether investments provide sufficient additional capacity as the probability and intensity of extreme rainfall change. Large projects can therefore benefit from economies of scale, but continually expanding centralized infrastructure may become increasingly expensive if extreme events continue to exceed historical design assumptions.

NYC faces a different economic trade-off because the opportunity cost of land is much higher and flood mitigation must compete with existing residential, commercial, transportation, and recreational uses. Consequently, an expensive project can still be cost-effective if it protects a large concentration of people and high-value infrastructure within a relatively small area. Distributed green infrastructure may also provide economic value that is not captured by measuring flood protection alone, since the same spaces can provide stormwater management, environmental, and public-space benefits. Research showing substantial runoff and peak-flow reductions from bioretention systems supports the potential for smaller interventions to provide

meaningful flood-management benefits when strategically located (Sabbagh et al., 2025). The comparison therefore suggests that cost-effectiveness should be evaluated relative to the risk and broader benefits produced, rather than by project size or construction cost alone. Houston can justify very large investments because centralized infrastructure protects extensive watersheds and can prevent substantial damage, while NYC may achieve greater value by strategically combining smaller interventions with major infrastructure in locations where land constraints make large-scale storage impractical. In both cities, the most economically sustainable approach is therefore one that maximizes risk reduction and additional benefits while maintaining the flexibility to adapt to changing flood conditions.

Infiltration and Environmental Impacts

Infiltration highlights an important difference in how NYC and Houston can achieve benefits beyond simply reducing flood levels. NYC's use of green infrastructure and Bluebelts allows stormwater to be retained, infiltrated, or absorbed by natural systems, creating the potential for simultaneous reductions in surface runoff, combined sewer overflow, and water pollution. Research on urban green infrastructure finds that these systems can provide meaningful hydrologic and water-quality benefits, although their effectiveness depends on factors such as soil conditions, design, and maintenance (Esraz-Ul-Zannat et al., 2024; Sabbagh et al., 2025). This makes infiltration particularly valuable in NYC, where managing stormwater before it enters the combined sewer network can address multiple environmental problems at once. Houston's conventional flood-control infrastructure provides a different set of benefits: its reservoirs, detention basins, and channels prioritize moving and temporarily storing large volumes of water rather than treating runoff at its source. Houston's increasing use of green infrastructure can therefore complement its centralized system by adding localized runoff reduction and water-quality benefits that large-scale storage facilities do not provide as directly. The comparison suggests that environmental effectiveness should be evaluated not only by how much floodwater a system manages, but also by how many environmental functions it can perform simultaneously.

The difference also reflects the physical conditions under which each city's infrastructure operates. NYC's dense development and extensive impervious surfaces limit opportunities for natural infiltration, making strategically placed green infrastructure particularly valuable where even small areas can intercept runoff before it enters the drainage network. Houston has greater opportunities to incorporate green infrastructure across larger areas, but its relatively flat landscape, low-infiltration soils, and dependence on large drainage channels and detention facilities make infiltration alone insufficient for managing major storms (Juan et al., 2020). This means neither approach is inherently superior: NYC gains greater value from infrastructure that combines flood reduction with water-quality and ecological functions within limited space, while Houston requires a combination of source-level infiltration and large-scale storage and conveyance to manage the magnitude of its flood risk. Overall, the research supports a layered approach in which infiltration provides environmental and localized hydrologic benefits while

larger systems provide the capacity needed when rainfall exceeds what natural or distributed infrastructure can absorb.

Policy Recommendations

Research supports making NYC's flood management more adaptive through continuous monitoring, targeted investment, and financial incentives. NYC's FloodNet program plans to deploy approximately 500 ultrasonic flood-depth sensors by mid-2027, providing real-time data that can improve flood modeling and identify areas where existing infrastructure is insufficient (Rosenzweig et al., 2024). Research also recommends targeting adaptation investments toward areas where multiple flood hazards overlap, rather than applying the same strategies citywide (Royte & Sanderson, 2026). Because long-term sea-level rise may eventually exceed the capacity of existing coastal defenses, planning should also incorporate stronger defenses and, where necessary, managed relocation (Gornitz et al., 2020). Financial incentives can further encourage local mitigation: FEMA's Community Rating System (CRS) provides flood-insurance discounts of up to 45% for communities that implement floodplain-management measures, while New York's Environmental Facilities Corporation provides $75 million in grants for flood mitigation and resiliency projects (Federal Emergency Management Agency, 2024; New York State Environmental Facilities Corporation, 2026). These policies suggest that NYC should combine real-time monitoring with targeted infrastructure funding and incentives that reward communities for implementing proactive mitigation.

Houston should continue expanding large-scale flood-control investment while increasing incentives for proactive, property-level mitigation. Research finds that storage-based drainage systems can substantially reduce flood losses, supporting continued investment in detention basins and expanded flood-storage capacity (Sohn et al., 2020). Harris County's $2.5 billion 2018 Flood Bond provides a major funding mechanism for expanding channel capacity and detention, and by 2026 it had been matched with approximately $2.8 billion in additional partnership funding (Harris County Flood Control District, 2018; Harris County Flood Control District, 2026). Research also supports acquiring flood-prone properties before repeated damage occurs, finding that proactive acquisition of vacant land can produce avoided losses nearly 2 to 10 times greater than acquisition costs in selected Houston areas (Atoba et al., 2021). Harris County's buyout and relocation programs already provide financial incentives for property owners to move out of high-risk areas, while acquired land can be converted to floodplain, retention, or green space (Harris County Housing and Community Development, 2026). Houston should therefore prioritize continued bond and federal/state grant funding for detention and storage, expand proactive buyouts in repeatedly flooded areas, and direct incentives toward properties and neighborhoods where mitigation can produce the greatest long-term reduction in flood exposure.

Conclusion

Overall, the comparison suggests that NYC and Houston should pursue different but layered approaches to long-term flood resilience. NYC should prioritize a combination of major coastal protection with distributed stormwater infrastructure, expanding Cloudburst projects, Bluebelts, permeable surfaces, and neighborhood-scale green infrastructure where multiple flood hazards overlap. Rather than treating each project independently, the city should use programs such as FloodNet to continually identify where risk is increasing and direct investment toward the neighborhoods where it can have the greatest effect. Houston, meanwhile, should continue prioritizing large-scale detention and flood storage because of the sheer volume of rainfall it must manage, but pair this infrastructure with floodplain preservation, proactive buyouts, green infrastructure, and stronger restrictions on development in repeatedly flooded areas. In both cities, future investment should focus more on creating multiple layers of protection so that the failure or overload of one strategy does not leave communities unprotected. Ultimately, the research suggests that the most sustainable form of flood mitigation is not to attempt to control every drop of water, but to strategically determine where water should be stored, where development should be protected, and where giving water more space is the safer and more economically sustainable choice.

Works Cited

A Buyout Displacement Index for Uncovering the Effects of Disinvestment in Greater Houston Watersheds - Earthea Nance, Sheri L. Smith, Jiwan Pun Thapa, Liza T. Powers, 2024. https://journals.sagepub.com/doi/10.1177/0739456X221116359?utm_source\. Accessed 8 Aug. 2026.

"An Introduction to Rain Gardens." https://extension.psu.edu/an-introduction-to-rain-gardens. Accessed 15 June 2026.

"Army Corps of Engineers Projects Prevent $13.3 Billion in Flood Damages during Spring Rain." U.S. Army Corps of Engineers Headquarters, https://www.usace.army.mil/Media/News/NewsSearch/Article/616269/army-corps-of-engineers-projects-prevent-133-billion-in-flood-damages-during-sp/. Accessed 13 July 2026.

Assessing the Triple Bottom Line Co-Benefits and Life Cycle Cost Tradeoffs of Cloudburst Infrastructure in New York City. ProQuest, https://www.proquest.com/openview/b7eb9b8f5cc3e26b72d7640e7dbdc746/1?pq-origsite=gscholar&cbl=18750&diss=y. Accessed 21 June 2026.

Attributable Human-Induced Changes in the Likelihood and Magnitude of the Observed Extreme Precipitation during Hurricane Harvey. Risser, 2017. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017GL075888?utm_source=. Accessed 11 Aug. 2026.

Attribution of Extreme Rainfall from Hurricane Harvey, August 2017. IOPscience, https://iopscience.iop.org/article/10.1088/1748-9326/aa9ef2. Accessed 8 Aug. 2026.

Ayalew, Tibebu B., et al. "Insights into Expected Changes in Regulated Flood Frequencies Due to the Spatial Configuration of Flood Retention Ponds." Journal of Hydrologic Engineering, vol. 20, no. 10, 2015, p. 04015010. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001173.

"City Completes Reconstruction of Rockaway Corridor." The Official Website of the City of New York, 13 June 2023, http://www.nyc.gov/site/dep/news/23-025/city-completes-reconstruction-rockaway-corridor-dramatic-safety-upgrades-new-porous.

Ceres, Robert L., et al. "Trade-Offs and Synergies in Managing Coastal Flood Risk: A Case Study for New York City." Journal of Flood Risk Management, vol. 15, no. 1, 2022, p. e12771. https://doi.org/10.1111/jfr3.12771.

Comparing Floodplain Evolution in Channelized and Unchannelized Urban Watersheds in Houston, Texas. Juan, 2020. https://onlinelibrary.wiley.com/doi/10.1111/jfr3.12604?utm_source. Accessed 17 Aug. 2026.

East Side Coastal Resiliency. https://www.nyc.gov/site/escr/index.page. Accessed 15 June 2026.

Emerson, Clay H., et al. "Watershed-Scale Evaluation of a System of Storm Water Detention Basins." Journal of Hydrologic Engineering, vol. 10, no. 3, May 2005, pp. 237–42. ASCE, https://doi.org/10.1061/(ASCE)1084-0699(2005)10:3(237).

Esraz-Ul-Zannat, Md., et al. “A Review of Nature-Based Infrastructures and Their Effectiveness for Urban Flood Risk Mitigation.” WIREs Climate Change, vol. 15, no. 5, 2024, p. e889. https://doi.org/10.1002/wcc.889.

Federal Emergency Management Agency. “Planning Considerations for Flood-Prone Areas.” Electronic Code of Federal Regulations, 44 CFR § 60.22, https://www.ecfr.gov/current/title-44/chapter-I/subchapter-B/part-60/subpart-A/section-60.22.

Galveston District. “About The Reservoirs.” https://www.swg.usace.army.mil/Missions/Dam-Safety-Program/About-The-Reservoirs/utm_source/. Accessed 15 July 2026.

Glas, Robin, et al. “Hazard Potential of Compound Flooding from Rainfall, Storm Surge, and Groundwater in Coastal New York and Connecticut.” Natural Hazards and Earth System Sciences, vol. 26, no. 5, 2026, pp. 2169–88. https://doi.org/10.5194/nhess-26-2169-2026.

Harris County Flood Resilience Plan. https://www.hcfcd.org/FloodResiliencePlan?utm_source. Accessed 7 Sept. 2026.

Heavier Rainfall in 129 U.S. Cities. Climate Central, https://www.climatecentral.org/climate-matters/heavier-rainfall-in-129-us-cities. Accessed 10 June 2026.

Hoss, Frauke, et al. “Effectiveness of Best Management Practices for Stormwater Treatment as a Function of Runoff Volume.” Journal of Water Resources Planning and Management, vol. 142, no. 11, 2016, p. 05016009. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000684.

“How Climate Change Is Fueling More Deadly and Destructive Floods.” Center for American Progress, 25 Nov. 2025, https://www.americanprogress.org/article/how-climate-change-is-fueling-more-deadly-and-destructive-floods/.

Integrated Hydrological Model-Based Assessment of Stormwater Management Scenarios in Copenhagen’s First Climate Resilient Neighbourhood Using the Three Point Approach. https://www.mdpi.com/2073-4441/9/11/883?utm_source. Accessed 28 June 2026.

Jayawardena, H. M. I. D. P., and Marjorie van Roon. “Water Sensitive Planning and Design as an Ecologically Inspired Approach to Delivering Flood Resilient Urban Environment in Sri Lanka.” Water Practice and Technology, vol. 12, no. 4, 2017, pp. 964–77. https://doi.org/10.2166/wpt.2017.100.

LL70 of 2025 Update. ArcGIS StoryMaps, 24 July 2026, https://storymaps.arcgis.com/stories/72eb2ba9e0a34a528c338b4139435b21.

“Mayor Adams Completes First Section of East Side Coastal Resiliency Project.” NYC.gov, 17 Oct. 2024, http://www.nyc.gov/site/dep/news/772-24/mayor-adams-completes-first-section-east-side-coastal-resiliency-project-taking-steps-to.

National Centers for Environmental Information. “Billion-Dollar Weather and Climate Disasters.” NOAA, https://www.ncei.noaa.gov/access/billions/state-summary/US. Accessed 5 Jun. 2026.

Nature-Based Solutions. European Commission, 9 Sept. 2025, https://research-and-innovation.ec.europa.eu/research-area/environment/nature-based-solutions_en.

Newman, Galen, et al. “A Framework for Evaluating the Effects of Green Infrastructure in Mitigating Pollutant Transferal and Flood Events in Sunnyside, Houston, TX.” Sustainability, vol. 14, no. 7, 2022, p. 4247. https://doi.org/10.3390/su14074247.

NPCC4: Climate Change and New York City’s Flood Risk. Rosenzweig, 2024. https://nyaspubs.onlinelibrary.wiley.com/doi/10.1111/nyas.15175?utm_source. Accessed 8 July 2026.

Peeples, Amelia, et al. “Flood Mitigation Using Channelization and Detention Given Changing Rainfall Conditions in Houston.” Natural Hazards Review, vol. 24, no. 4, 2023, p. 05023012. https://doi.org/10.1061/NHREFO.NHENG-1861.

Pickett, Steward T. A., et al. “The Relational Shift in Urban Ecology.” Ambio, vol. 53, no. 6, 2024, pp. 845–70. https://doi.org/10.1007/s13280-024-02001-y.

Price, Jason C., et al. “Calibrated Methodology for Assessing Climate Change Adaptation Costs for Urban Drainage Systems.” Urban Water Journal, vol. 13, no. 4, 2016, p. 10.1080/1573062x.2014.991740.

Project Background – ESCR. https://www.nyc.gov/site/escr/about/project-background.page?utm_source. Accessed 20 June 2026.

Sarhadi, Ali, et al. Climate Change Contributions to Increasing Compound Flooding Risk in New York City. Bulletin of the American Meteorological Society, Feb. 2024, https://doi.org/10.1175/BAMS-D-23-0177.1.

Sabbagh, Mohamed, et al. “Urban Stormwater Bioretention Reduces Runoff and Improves Water Quality.” Journal of Hydrology, vol. 663, 2025, p. 134163. https://doi.org/10.1016/j.jhydrol.2025.134163.

Seong, Kijin, et al. “Racial and Ethnic Change in Floodplain Buyout Neighborhoods.” Journal of Planning Education and Research, Dec. 2025, p. 0739456X251397310. https://doi.org/10.1177/0739456X251397310.

Serra-Llobet, Anna, et al. “Flood Diversions and Bypasses: Benefits and Challenges.” WIREs Water, vol. 9, no. 1, 2022, p. e1562. https://doi.org/10.1002/wat2.1562.

Shetty, Nandan H., et al. “Studying the Effect of Bioswales on Nutrient Pollution in Urban Combined Sewer Systems.” Science of The Total Environment, vol. 665, 2019, pp. 944–58. https://doi.org/10.1016/j.scitotenv.2019.02.121.

Sohn, Wonmin, et al. “How Effective Are Drainage Systems in Mitigating Flood Losses?” Cities, vol. 107, 2020, p. 102917. https://doi.org/10.1016/j.cities.2020.102917.

The Bluebelt Program – DEP. https://www.nyc.gov/site/dep/water/the-bluebelt-program.page?utm_source. Accessed 3 July 2026.

"The Role of Greater Copenhagen Utility in Implementing the City's Cloudburst Management Plan." ResearchGate, https://doi.org/10.2166/wpt.2017.039. Accessed 21 June 2026.

VanWoert, Nicholaus D., et al. "Green Roof Stormwater Retention." Journal of Environmental Quality, vol. 34, no. 3, 2005, pp. 1036–44. https://doi.org/10.2134/jeq2004.0364.

Wahl, Thomas, et al. "Increasing Risk of Compound Flooding from Storm Surge and Rainfall for Major US Cities." Nature Climate Change, vol. 5, no. 12, 2015, pp. 1093–97. https://doi.org/10.1038/nclimate2736.

Zhang, Wei, et al. "Urbanization Exacerbated the Rainfall and Flooding Caused by Hurricane Harvey in Houston." Nature, vol. 563, no. 7731, 2018, pp. 384–88. https://doi.org/10.1038/s41586-018-0676-z.

Protecting Children in Online Gaming: Legal Responses to Grooming and the Limits of Platform Accountability By Dongwon Shin

Abstract

Children are more exposed to danger in video games caused by an increase in digital platform accessibility, specifically child grooming in video games that include live chatting, private chatting, a gifting system, and more. This paper examines how children are exploited on video game platforms through grooming, trafficking-related predation, and other harmful platform practices. The paper then analyzes the legal approaches used to address these harms. It investigates recent lawsuits by plaintiffs against video game companies, such as those in Louisiana and Kentucky, to show their allegations, reforms sought, and the effectiveness of these legal strategies in addressing the broader structural causes of exploitation while balancing online safety with legal protections for digital platforms. Further, this paper evaluates past legal approaches and platform reforms aimed at reducing child exploitation on video game platforms, arguing that while current lawsuits and regulations have improved accountability and safety measures, significant gaps remain that require further legal and regulatory action. Particularly, the Louisiana and Kentucky complaints demonstrate how consumer-protection and negligence claims challenge misleading safety representations, inadequate warnings, and allegedly unsafe platform design. Moving forward, lawmakers may choose to follow a more proactive approach, as seen in South Korea, where harmful conduct is targeted during the grooming process, rather than waiting for every form of abuse to be completed.

Introduction

In July 2021, allegations of online grooming received renewed attention when League of Legends streamer Joedat ‘Voyboy’ Esfahani released a video discussing past allegations against Robert ‘Hashinshin’ Brotz, a prominent League of Legends personality and former streamer (Masters). The allegations included claims of inappropriate online interactions with underage girls, while the most serious accusations concerned alleged inappropriate relations with girls below the age of consent (Masters). Hashinshin denied the majority of the allegations, and no legal action had been taken against him at the time of the article’s publication (Masters). Although the allegations remained contested, the incident illustrates how grooming concerns can emerge from within gaming communities and through the online communication spaces surrounding them. This paper examines gaming platforms' responsibility for the risks children face when using their communication features.

Many children use online games to spend time with friends and meet other players. Players can chat with friends and strangers, join communities, and create virtual identities. These features support creativity and friendship, but can also expose children to harmful contact. Faraz et al. review threats associated with children’s participation in online gaming, including harmful content, predatory contact, and limitations in existing safeguards (p. 115895). Preventing harmful

contact in these settings requires attention to the communication features that platforms provide and the rules governing their use.

Roblox combines user-generated games with social communication. The company reported 123 million average daily active users for the three months ended June 30, 2026, and noted that many of its users were under thirteen (Roblox Corporation 27, 56). In its petition against Roblox, the State of Louisiana describes the service as an "interactive online gaming platform" that encourages real-time interaction through gameplay, user hubs, direct messaging, and voice chat (State of Louisiana 3). The State also claims that the platform is easily accessible to children and that an account can be created without parental involvement (State of Louisiana 3). A later Kentucky complaint raises similar allegations, including claims about Roblox's child audience, parental consent, safety controls, warnings, and platform design (Commonwealth of Kentucky 1–3).

This paper focuses on online sexual grooming and related sexual solicitation of minors in gaming environments and associated communication services. Grooming involves developing a relationship with a child to facilitate sexual contact or abuse, including conduct occurring online; it is related to, but distinct from, sexual exploitation and trafficking (Drejer et al. 3; Greijer and Doek 95–96). Public and private chat, voice communication, avatars, in-game gifts, weak age-verification systems, and connections to external services can all shape how users interact and can create risks when harmful users exploit those features (Drejer et al. 27–28). Platforms aren't automatically responsible whenever a user causes harm. The question is whether they've taken reasonable steps to address risks their communication systems create or increase.

The reviewed literature gives less attention to how recent legal strategies address these risks. Existing research largely explains how risks develop, where safety measures may fail, and what general safeguards could reduce those risks (Faraz et al. 115895–115896). Lawsuits, however, must translate these broad problems into specific legal theories such as negligence, deceptive trade practices, failure to warn, consumer protection, public nuisance, unjust enrichment, or other forms of platform liability (Commonwealth of Kentucky 64–66; State of Louisiana 37–40). A legal remedy may address one risk while leaving others unresolved. For example, requiring stronger warnings may address a lack of transparency but do little about cross-platform communication, while stronger age verification may limit some forms of identity misrepresentation without addressing every method offenders may use to contact, deceive, isolate, or manipulate minors (Drejer et al. 27–29).

To what extent do the selected legal approaches address the grooming mechanisms identified in the literature? This paper examines that question by comparing documented risks with the claims and requested remedies in the Louisiana and Kentucky complaints.

Ultimately, this paper argues that the selected legal approaches address several platform practices associated with grooming risks, while leaving unresolved questions about enforceability and protection across services. Its central concern is how legal claims and potential remedies correspond to the mechanisms described in the literature. The literature identifies mechanisms and enabling conditions that include identity misrepresentation,

non-textual interaction, weak age verification, and communication across services, some of which extend beyond a single company's moderation system (Drejer et al. 27–28; Merry and Whitfield 3321–3322). Using doctrinal legal analysis, which involves critical analysis and synthesis of relevant legislation and case law, this paper examines lawsuits, statutes, and platform-liability doctrines to evaluate whether current legal approaches address those mechanisms (Hutchinson 131). This paper finds that legal complaints against platforms target several important platform-level problems, particularly communication controls, warnings, safety representations, age verification, and safety design (Commonwealth of Kentucky 1–3, 65–66; State of Louisiana 3, 21, 29). However, identifying these risks in a complaint does not establish that the claims are legally enforceable or that the requested remedies would provide comprehensive protection.

Methodology

Doctrinal Analysis

This study used doctrinal legal analysis, supplemented by interdisciplinary evidence, to evaluate existing legal protections and identify potential reforms to strengthen child safety on gaming platforms (Hutchinson). Doctrinal analysis guided the identification and interpretation of legal rules, claims, defenses, and remedies relevant to platform responsibility (Hutchinson; Theil). The study compared the Louisiana and Kentucky lawsuits against Roblox to examine how the two states applied these legal principles to the company's practices (Commonwealth of Kentucky; State of Louisiana). Roblox was selected because it combines gaming, user-generated content, and social communication, serves a large population of young users, and has faced multiple legal actions concerning child safety (Commonwealth of Kentucky 1–3; Roblox Corporation 19–20; State of Louisiana 3–4). The study examines how the selected lawsuits and regulatory approaches target inadequate safeguards, misleading safety representations, age-verification practices, and other platform decisions, and identifies limitations in the legal and practical protection they may provide.

The research design was adapted from legal and child-safety studies in the source collection. Hutchinson (2015) explains that reform-oriented doctrinal research analyzes and synthesizes legal authorities while incorporating evidence from other disciplines to evaluate existing law and support reform recommendations. Similarly, Theil (2025) argues that empirical research can help evaluate how law operates but stresses that legal, empirical, and normative conclusions must remain separate. Following these approaches, this study used legal authorities to determine the relevant legal framework, lawsuit complaints to identify the states' arguments, and interdisciplinary research to evaluate whether those arguments addressed documented platform risks. Faraz et al. (2022) mapped threats within online gaming and reviewed existing protections, while Drejer et al. (2024) compared ten games using four risk categories: verbal-interactive, environmental-visual, behavioral-economic, and cybersecurity risks. This study adapted their category-based structure into five categories for analyzing the complaints: 1)

age assurance and parental consent concerned allegations about identifying users' ages and obtaining parental permission; 2) communication safeguards concerned restrictions on text, voice, and private communication; 3) parental warnings and controls concerned risk information and account-supervision tools; 4) content moderation concerned detecting, reporting, and removing harmful content and users; and 5) movement to external services concerned interactions continuing beyond Roblox's safeguards. For each complaint, the study identified the factual allegations, legal claims, alleged platform failures, and requested remedies. Within each category, the analysis compared the alleged failures and requested remedies with prior research and the reforms discussed in the complaints. It considered which risks the claims addressed, whether the requested remedies could be enforced, and what evidence supported their likely effects. The analysis also considered relevant legal limitations, including Section 230, while distinguishing allegations from judicial findings, settlement terms, and voluntary platform changes.

Data Collection

Sources were purposely selected according to their credibility, direct relevance to the research question, and usefulness in establishing legal authority or explaining how platform risks operate. The legal sources included the Louisiana and Kentucky complaints, the Louisiana Unfair Trade Practices Act, the Kentucky Consumer Protection Act as invoked in the Kentucky complaint, and Section 230 of the Communications Decency Act of 1996. Section 230 generally limits platforms' liability for information supplied by third-party users, subject to statutory limits and exceptions (Brannon and Holmes 9, 29–32; Communications Decency Act, sec. 230(c)(1), 230(e)). The study also examined legal scholarship concerning the Children's Online Privacy Protection Act (COPPA), a federal law regulating how covered online services collect, use, and disclose personal information from children under thirteen (Johnson). Interdisciplinary sources included Faraz et al. (2022), Drejer et al. (2024), Merry and Whitfield's (2025) empirical study of reported sexual harm in gaming and streaming communities, and Jarvie and Renaud's (2024) review of age-verification laws and practices. Reports from ECPAT International (2024), a global child-protection network focused on ending the sexual exploitation of children, and Emma S. (2024) were used to identify recommended child-safety measures. The complaints' descriptions of Roblox's safety changes were reviewed to identify reforms acknowledged by the states and the shortcomings they continued to allege; the study did not independently evaluate Roblox's current safeguards. Primary legal and official sources were prioritized, while investigative journalism and other secondary reporting were used mainly to provide context and were not treated as proof of legal liability.

Limitations

This research has several limitations. As a high school student based in South Korea, I examined U.S. legal disputes remotely through publicly available sources, such as those found on Google Scholar. Limited time for this project restricted the number of cases and publications I

could review, resulting in a focused comparison rather than an exhaustive survey. The study provides a doctrinal analysis of legal protections and potential reforms, not an empirical test of whether particular laws or safeguards reduce child exploitation. The core case analysis concerns the Louisiana and Kentucky complaints and the applicable state and federal legal frameworks. I then look at US federal enforcement, South Korean law, and international standards, which provide supplementary illustrations rather than a comprehensive comparative-law assessment. Therefore, these findings cannot automatically be generalized to other platforms or jurisdictions.

Also, the materials reviewed did not include final judgments resolving the selected claims, so allegations were not treated as established facts (because a federal district court or equivalent state court establishes the facts in its ruling), and the study could not determine Roblox's liability or the effectiveness of final remedies. Finally, source collection concluded on September 7, 2026, but the analysis is limited to the selected authorities and versions discussed and does not constitute an exhaustive review of all developments through that date. Later court decisions, legislative changes, or platform updates may alter its conclusions. The findings therefore identify potential strengths and gaps in the examined legal approaches and support reform recommendations without establishing causal effects or predicting litigation outcomes.

Literature Review

The Problem: How Are Children Exploited on Video Game Platforms?

Research on child safety in online gaming identifies risks arising from both the content children encounter and the relationships they form. Faraz et al. (2022) review the literature using three categories: content risks, contact risks, and conduct risks. Content risks concern exposure to inappropriate material; contact risks concern harmful interactions with other people; and conduct risks concern harmful behavior within online environments (Faraz et al. 115897–115898). This framework helps explain why sexual exploitation cannot be understood only as a problem of removing explicit material. Children may also face risks through apparently ordinary communication and participation in gaming communities. Accurate terminology is therefore important. The terminology guidelines by Greijer and Doek (2025) describe grooming as establishing a relationship with a child, in person or through digital technologies, to facilitate sexual contact (pp. 95–96). They distinguish sexual exploitation from sexual abuse through an underlying exchange, gain, or benefit, which need not be monetary (Greijer and Doek 40–41). Their discussion of international trafficking instruments explains that recruiting, transporting, transferring, harboring, or receiving a child for exploitation may constitute child trafficking without proof of the coercive means generally required in adult-trafficking definitions (Greijer and Doek 109–110). Drejer et al. (2024), meanwhile, distinguish unwanted sexual solicitation from online sexual grooming but include solicitation within grooming for purposes of their own study (p. 3). These distinctions matter because evidence of grooming or online sexual harm should not automatically be described as trafficking.

Merry and Whitfield (2025) analyzed 55 publicly available victim statements describing 45 alleged offenders in gaming, streaming, and esports communities. The sample included 22 cases involving children and 33 involving adults (pp. 3312–3313). The reported harms included sexual assault, sexual communication with children, grooming, and image-based offenses (Merry and Whitfield 3321). These publicly disclosed accounts cannot establish how common such harms are among gamers. Some statements also lacked corroborating material or details needed for the analysis (Merry and Whitfield 3326). Adding to this analysis of victim accounts, Drejer et al. (2024) examine game features. Their preprint, *It's All in the Game*, analyzes ten games and identifies four risk categories involving communication, visual environments, in-game exchanges, and cybersecurity (p. 1). The authors explain how these features could support contact, deception, isolation, and coercion (Drejer et al. 27). The study identifies features that may enable exploitation. It doesn't measure how often victimization occurs or whether a particular feature causes it. The authors acknowledge the limited sample and explicitly state that they did not examine the effectiveness of content moderation and reporting systems (Drejer et al. 30).

Trust can develop through repeated play. Profiles and avatars may also allow an adult to conceal their age or intentions. Drejer et al. (2024) describe how avatars can be used to appear younger or convey a shared identity with a child (p. 27). Rosen (2005) describes how players present alternative versions of themselves and form emotional connections in virtual worlds (pp. 4, 12). The essay discusses these relationships but doesn't study child grooming. Alongside the identities presented through avatars, an individual's status in the gaming community can shape trust. Merry and Whitfield (2025) found that some accounts described victims feeling fortunate to receive attention from a prominent streamer or esports professional; the authors argue that fame could allow an offender to bypass parts of the usual process of establishing trust, particularly in cases involving children (pp. 3321–3322). Across the sample, they also describe using moderator roles on Twitch or Discord to maintain compliance (p. 3322). A supporting example appears in Masters's (2021) reporting on grooming allegations against the streamer Hashinshin, which Voyboy publicized. The report also records Hashinshin's denials of most allegations. This example illustrates public controversy involving community influence; the news report does not establish a judicial finding or the prevalence of such conduct.

Another mechanism is isolation through private communication or less visible areas within games. Drejer et al. (2024) describe private chats, direct messaging, and secluded virtual locations as settings in which contact can become less observable (pp. 27–28). Merry and Whitfield (2025) also found that accounts involving children described using voice chat and video calls. The authors suggest that live conversations may leave less recoverable evidence than messages or images. They did not directly test why offenders used these channels (Merry and Whitfield 3322). Safety systems that examine only text may miss harmful interactions through voice chat and video calls. These communication risks also extend across several services. Merry and Whitfield (2025) identify Discord and Twitch as important sites of initial contact in their sample, with other services used for communication and image sharing (pp. 3321–3322). Initial

contact sometimes occurred on associated services such as Discord and Twitch rather than in-game, so the study does not establish a uniform progression from in-game contact to external communication. Drejer et al. (2024) also discuss external links and cases in which contact through gaming continued through other communication services (pp. 27–28). The Exodus Road's awareness article describes related risks involving gaming and Discord (Emma S.). It offers additional illustrations, but it is an advocacy article rather than an academic study of trafficking prevalence.

Virtual gifts and currency can also be used for manipulation. Drejer et al. (2024) identify in-game money and tokens as resources that can build trust or pressure children into unwanted behavior (p. 27). Carville and D'Anastasio (2024) conducted an investigation that supplied supporting examples from police records involving Robux and Roblox gift cards exchanged for sexual material from children. These reports describe how virtual rewards can be used to manipulate children, but don't show how often this happens on Roblox.

The literature also identifies limitations in safeguards that depend on parental involvement. Faraz et al. (2022) explain that parental controls require parents to understand their functions, and that limited awareness can undermine their use (p. 115896). Their review also discusses the difficulty of restricting harmful contact while retaining communication features central to multiplayer games (Faraz et al. 115903–115904). Drejer et al. (2024) identify areas that remain outside parental-control settings and describe ways that users may circumvent restrictions (p. 28). Having parental controls doesn't show that parents can use them effectively or that the settings cover every risky interaction. Age verification presents a related problem of determining who can access particular spaces and features. Drejer et al. (2024) criticize reliance on self-reported ages in the platforms they studied because users can enter false information (p. 28). Jarvie and Renaud (2024), examining age-verification legislation and public perceptions more broadly, identify concerns about privacy, affordability, and effectiveness. They also identify several implementation problems: service providers may receive insufficient official guidance, risk-based compliance may become fragmented and inconsistent, available technologies may not satisfy both legal requirements and users' privacy needs, and public distrust may reduce acceptance of age-verification systems (Jarvie and Renaud 15–17). Their public-perception analysis draws on UK-related Reddit and YouTube comments, which limits its generalizability (Jarvie and Renaud 16–17). Jarvie and Renaud's (2024) study describes implementation challenges as they stood in 2024; it does not test whether age verification prevents grooming in games or document every platform's current practices.

Faraz et al. (2022) identify limited access to gaming datasets as a barrier to research on predatory behavior. Much of the research they reviewed focused on chat logs (p. 115909). Drejer et al.'s (2024) discussion of avatars, virtual spaces, and in-game exchanges suggests that a broader range of behavior may require attention, while their study leaves moderation and reporting effectiveness untested (pp. 27, 30). Carville and D'Anastasio (2024) report that current and former Roblox safety employees described receiving hundreds of escalated child-safety reports each day, an uncleared backlog of incident reports, unanswered requests for additional

staff, rejected proposals for measures such as pop-up safety notices, and automated systems that could miss subtle signs of grooming. Roblox disputed the claims concerning inadequate resources and backlogs, stating that moderator numbers do not determine moderation quality and that it continually evaluates and develops new safety tools (Carville and D'Anastasio). Because the article presents conflicting accounts from employees and the company, it documents concerns at the time of publication rather than independently measuring Roblox's current moderation performance.

ECPAT International's (2024) Seoul roundtable report adds a child-centered perspective. It describes gaming as a source of belonging, social connection, and identity exploration, while recognizing that those same environments can expose children to harm (pp. 1–2). Its recommendations emphasize sustained participation by children and cooperation among gaming companies, civil society, and law enforcement (ECPAT International 4–5). The roundtable offers recommendations, but it doesn't test whether they reduce harm. Shared interests, misleading identities, private conversations, community status, and virtual rewards recur throughout the literature. How often these factors lead to abuse, and how well specific safeguards prevent it, remains less clear.

Legal Approaches So Far

Legal responses to child exploitation in gaming address different parts of the problem. Criminal law targets offenders; privacy and consumer-protection law regulates companies' data practices and representations; tort law tests whether a platform's design, warnings, or safeguards created an unreasonable risk; and international frameworks state broader duties to protect children. This division matters because the mechanisms discussed in the first section do not all present the same legal question. A prosecution for sexually soliciting a child addresses an offender's conduct, while a deceptive-practices claim addresses a company's own speech. A claim demanding better screening of user messages may instead be characterized as imposing liability for publishing third-party content. ECPAT International (2024) similarly treats regulation as one part of a system that also requires prevention, education, child participation, and cooperation among companies, law enforcement, and civil society (pp. 2, 4–5).

Recent state lawsuits against Roblox illustrate how consumer-protection and tort law are being used to address alleged platform-level failures. State of Louisiana's August 2025 petition alleges that Roblox marketed its service as safe for children while failing to disclose risks and implement adequate controls. The petition alleges violations of Louisiana's Unfair Trade Practices and Consumer Protection Law (LUTPA) and brings claims for negligence, public nuisance, and unjust enrichment. It seeks restitution, penalties, damages, disgorgement, and injunctive relief (State of Louisiana 37–41). LUTPA prohibits unfair competition and unfair or deceptive practices in trade or commerce (Louisiana State Legislature, sec. 51:1405(A)). Commonwealth of Kentucky's October 2025 complaint asserts violation of the Kentucky Consumer Protection Act (KCPA), unjust enrichment, negligence, and negligent failure to warn, seeking similar relief and up to $2,000 for each willful KCPA violation (Commonwealth of

Kentucky 61–67). These pleadings contain allegations, not findings, that Roblox caused or knowingly permitted the alleged harm. Roblox's July 2026 filing with the Securities and Exchange Commission groups the Kentucky and Louisiana actions among state child-safety proceedings described as being at an early stage and states that the company anticipated filing, or had filed, motions to dismiss in the remaining cases (Roblox Corporation 19–20). The materials reviewed did not include a merits ruling resolving either selected case.

Louisiana's consumer-protection theory focuses most directly on Roblox's own commercial conduct. LUTPA declares unfair or deceptive acts or practices in trade or commerce unlawful. Relying on LUTPA's enforcement provisions, the petition requests an injunction, civil penalties, restitution, and other relief. It alleges that Roblox misrepresented the platform as safe, omitted material information about predators, and used those representations to increase usership and purchases (State of Louisiana 37–39). This fits LUTPA more directly than a theory based only on harmful user communications because it identifies statements and omissions allegedly made by Roblox itself. The requested remedies may be available if Louisiana proves a violation, but the complaint does not establish that entitlement. Its negligence count presents additional questions about duty and causation because it challenges age and identity verification, parental consent, moderation, messaging, and warnings; the same count labels the alleged environment a public nuisance (State of Louisiana 39–40). The petition does not cite a separate nuisance statute, so that label does not remove the need to establish a viable theory under Louisiana's general tort articles. It also pleads unjust enrichment only as an alternative if no remedy at law is available (State of Louisiana 40). Kentucky follows a comparable strategy but separates negligent design from failure to warn (Commonwealth of Kentucky 61–66).

Section 230 of the Communications Decency Act may limit these theories, and applying it requires analyzing the duty and conduct underlying each claim. Section 230(c)(1) provides that an interactive computer service may not be treated as the publisher or speaker of information supplied by another content provider (Communications Decency Act, sec. 230(c)(1)). Courts generally ask whether the defendant provides or uses an interactive computer service, whether the claim treats the defendant as a publisher or speaker, and whether the relevant information came from another provider. If any answer is no, Section 230(c)(1) does not bar the claim (Brannon and Holmes 9). Roblox is likely to satisfy the first part. The dispute concerns the conduct targeted by each claim. Section 230 permits states to enforce laws consistent with the federal provision while precluding state-law causes of action or liability that are inconsistent with it (Communications Decency Act, sec. 230(e)(3)). Courts therefore examine the particular application of consumer-protection or tort law rather than automatically barring every state-law claim (Brannon and Holmes 29).

Claims faulting Roblox for failing to monitor, screen, or delete users' communications face the strongest Section 230 defense. The Fifth Circuit has treated a failure-to-protect negligence claim as an attempt to impose publisher liability when the alleged harm depended on third-party communications (*Doe v. MySpace* 419–420). Section 230 does not necessarily cover claims based on a platform's independent conduct, however. Courts have allowed claims

involving a platform's own duty to warn or product design when liability did not depend on monitoring or altering user content (*Doe v. Internet Brands* 851–852; *Lemmon v. Snap* 1092–1094). These decisions illustrate that Roblox's defense will depend on whether each Louisiana claim targets its treatment of users' communications or its own representations, warnings, transactions, or independently actionable design choices.

The Allow States and Victims to Fight Online Sex Trafficking Act of 2017 (FOSTA) creates a limited exception to Section 230. Communications Decency Act, added by FOSTA in 2018, preserves specified federal civil sex-trafficking claims and certain state criminal prosecutions (47 U.S.C. § 230(e)(5)). Louisiana, however, pleaded state civil claims under LUTPA, negligence, public nuisance, and unjust enrichment (State of Louisiana 39–40). Louisiana's pleaded state civil counts are not among the categories preserved by Section 230(e)(5), so their treatment turns on the ordinary Section 230 analysis, including whether a particular claim seeks to impose publisher liability for third-party content (Brannon and Holmes 30–32).

Federal enforcement supplies another approach. Johnson (2020) explains that COPPA requires covered operators to provide notice and obtain verifiable parental consent before collecting personal information from children under thirteen (pp. 420–421). She also identifies limits created by COPPA's age threshold, its knowledge requirement, and age screens that children can answer inaccurately (Johnson 449, 453–454). Arneaud (2022) argues more broadly that the Federal Trade Commission (FTC) should use Section 5 of the Federal Trade Commission Act (FTC Act) against unfair or deceptive video-game practices. Because Arneaud's examples concern game-industry marketing and design rather than grooming, the article is most useful here as a doctrinal analogy for safety representations. The FTC's enforcement against Epic Games provides a closer child-safety example. In December 2022, the FTC announced settlements under which Epic Games would pay a $275 million COPPA penalty, adopt protective voice and text communication defaults for children and teenagers, and pay $245 million for consumer refunds through a separate proposed administrative order concerning alleged deceptive billing practices (Federal Trade Commission). This shows that regulators can address a game company's data practices, communication defaults, and deceptive design without first holding it liable for a predator's speech. Yet COPPA principally regulates data, and Section 5 still requires an unfair or deceptive business practice; neither establishes a general duty to prevent every instance of grooming.

Age verification illustrates the limits of relying on a single preventive measure. Jarvie and Renaud (2024) identify concerns about privacy, affordability, effectiveness, and implementation guidance in age-verification laws (pp. 15–17). Their findings suggest that requiring an age check does not establish that the chosen system will accurately identify children or reduce grooming. International child-rights standards supply a broader framework. Committee on the Rights of the Child (2021) explains that states' obligations in the digital environment include protecting children from violence, sexual exploitation, trafficking, and other exploitation. General Comment No. 25 calls for protective mechanisms, investigation, remedies, and regularly

reviewed safeguards, and specifically recognizes digital solicitation, livestreamed abuse, sexual extortion, and coerced self-generated material (Committee on the Rights of the Child, pars. 25, 80–82). This framework guides state policy but does not provide a cause of action pleaded in the Louisiana or Kentucky cases.

South Korea's 2021 legislation provides a historical illustration of criminal intervention directed at grooming communications. Abd Jalil's (2013) presentation at the 13th Australasian Conference on Child Abuse and Neglect provides historical background on Korea's earlier combination of youth-protection, communications, gaming, and criminal laws, but it should not be treated as a statement of current Korean law. A 2021 amendment to Korea's Act on the Protection of Children and Youth Against Sex Offenses added Article 15-2; paragraph (1) criminalized specified persistent or repeated sexual communications and enticement through information and communications networks by adults aged nineteen or older for the purpose of sexually exploiting a child or young person (Act on the Protection of Children and Youth Against Sex Offenses, arts. 15-2(1)). The 2021 amendment to the same Act also authorized concealed-identity and undercover investigations of specified digital sexual offenses, subject to procedural safeguards (Act on the Protection of Children and Youth Against Sex Offenses, arts. 25-2 through 25-4). The 2021 provision allowed intervention during specified exploitative communications and solicitation. Its primary target was the alleged offender to curtail this behavior. Questions about a platform's safety statements or civil liability require a separate analysis.

Neither Abd Jalil's historical presentation nor the statutory text evaluates whether these measures have reduced exploitation (Abd Jalil; Act on the Protection of Children and Youth Against Sex Offenses). Louisiana's petition presents a plausible LUTPA theory and seeks restitution, civil penalties, and injunctive relief, but those remedies depend on proving the alleged unfair or deceptive conduct (State of Louisiana 37–39). Still, Section 230 may bar claims that functionally impose liability for failing to police third-party communications while leaving room for claims based on Roblox's own speech or conduct (Brannon and Holmes 13–16). Negligence depends on duty and causation, public nuisance raises an additional question of doctrinal fit, and unjust enrichment is pleaded only as an alternative if no other legal remedy is available (State of Louisiana 39–40). The procedural assessment here relies on Roblox's July 2026 disclosure and does not establish the subsequent development of either selected case (Roblox Corporation 19–20). Accordingly, these remain contested legal theories, so we can not draw definitive conclusions until fully litigated in court. The wider literature reveals a continuing division among offender-focused criminal law, platform-focused consumer and privacy regulation, and international child-rights standards (Committee on the Rights of the Child, pars. 25, 80–82; Federal Trade Commission). The materials reviewed here provide limited evidence that any particular legal intervention actually reduces exploitation in gaming.

Results and Discussion

Comparing the Louisiana and Kentucky Cases

Both complaints challenge platform practices associated with the grooming risks described in the literature. Whether they can lead to effective safeguards remains uncertain: some claims face potential defenses, and restrictions on Roblox cannot directly control communication on other services. Roblox's July 2026 disclosure described the selected state proceedings as being at an early stage; the materials reviewed contain no final judgment resolving Louisiana's or Kentucky's claims (Roblox Corporation 19–20). The Louisiana and Kentucky complaints illustrate this partial alignment between consumer-protection law and the mechanisms of exploitation.

Louisiana's LUTPA allegations connect allegedly misleading safety representations with parents' decisions to permit use and purchases. Kentucky similarly challenges representations and omissions, but separately pleads negligent design and failure to warn (Commonwealth of Kentucky 61–66; State of Louisiana 37–40). Warnings can help parents understand risks; communication restrictions can limit contact. Correcting a misleading advertisement won't, by itself, prevent an adult from contacting a child. The limits of these approaches become clearer when considering risks that extend beyond Roblox. Louisiana describes the platform as a point of initial contact before communication moves elsewhere; Kentucky makes similar allegations (Commonwealth of Kentucky 31; State of Louisiana 21). These concerns correspond to Drejer et al.'s (2024) account of contact, deception, isolation, and exchanges across gaming environments (pp. 27–28).

Louisiana requests the implementation of effective safeguards, while Kentucky seeks broad injunctive relief against the alleged unlawful conduct; neither request establishes which specific interventions a court would order or whether they would interrupt movement between services (Commonwealth of Kentucky 66–67; State of Louisiana 39–41). Restrictions within Roblox cannot directly govern another service's communications. Accessible reporting and lawful cooperation between services may help address this gap. Whether Roblox is liable for harm occurring elsewhere remains a separate legal question.

Section 230 and the Scope of Platform Responsibility

Section 230 requires attention to the duty alleged in each claim. Claims that require a platform to screen or remove users' messages may impose publisher liability, even when framed as negligence. By contrast, a company's own misleading statements can present a different question (Brannon and Holmes 13–16). In *Lemmon v. Snap (2021)*, the Ninth Circuit allowed a design claim to survive Section 230 because the alleged duty was independent of publishing user content (pp. 1092–1094). That reasoning offers an analogy, not a guaranteed outcome for Roblox. Calling moderation a design defect does not, by itself, avoid immunity. Courts should assess each claim's actual duty and alleged conduct, then evaluate whether claims that fall outside Section 230 satisfy duty, causation, and other legal requirements; therefore, surviving Section 230 would not establish liability.

For that reason, the appropriate approach in the Louisiana and Kentucky cases is to evaluate responsibility claim by claim. If actionable deception is proven, relief should correct the representations and omissions that misled families. If an independently actionable safety failure is established, prospective relief should address that failure and provide a way to assess compliance. Hence, Louisiana's public-nuisance and unjust-enrichment theories should not substitute for establishing their distinct legal requirements. Any remedy should address the conduct for which the platform is found legally responsible.

Strengthening FTC Enforcement and COPPA Protections

Federal enforcement demonstrates how legal action can reach communication settings. The FTC's 2022 Epic Games settlement required voice and text communications to remain off for children and teenagers unless the specified consent was provided (Federal Trade Commission). Such a remedy is more closely connected to unwanted contact than an advertising correction alone. Where their authority permits, regulators should identify the feature alleged to be unfair or deceptive and seek an order specifying what must change and how compliance will be checked. The Epic order does not establish duties for every other game. Its requirements alone also do not show whether victimization fell. Alongside enforcement directed at communication settings, further COPPA reform should address gaps that remain beyond the FTC's 2025 amendments, particularly the limited coverage of teenagers under the statute's under-thirteen framework (Federal Trade Commission; Johnson 420–421). Johnson (2020) argues that extending COPPA-style protections to everyone would preserve basic privacy safeguards even when users misstate their ages (p. 449). Extending privacy protections to teenagers could reduce the gaps created by inaccurate age declarations. Such reform should account for adolescents' increasing autonomy. Yet expanding privacy coverage alone would not address manipulative relationships formed through chat, gifts, or community status. Privacy reform should be paired with protections that address how users establish and maintain contact.

Age Verification and Its Limits

Age verification addresses one weakness identified in existing research, but it cannot fully address grooming. Drejer et al. (2024) explain how self-reported age and limitations in parental controls can leave children exposed to risky interactions (pp. 27–28). Kentucky's complaint acknowledges changes to Roblox's communication settings and parental controls in November 2024 while alleging that protections remained inadequate (Commonwealth of Kentucky 24–25). Reforms must therefore be evaluated separately to understand the problems particular to each of them. More accurate age checks may help platforms apply communication restrictions. They can't establish a user's intentions. Jarvie and Renaud (2024), whose analysis includes the UK context, identify concerns about privacy, affordability, and effectiveness that complicate age-verification requirements (pp. 15–17). These concerns suggest that a legal requirement to check age should be evaluated according to the protection it provides and the burdens it creates for users. Age verification should be judged by whether it helps prevent

unwanted contact and manipulation. It should therefore complement communication safeguards, understandable parental controls, and education about the risks described in the literature. The existing sources support this combined approach, but they do not establish that any particular verification method prevents grooming.

Combining Criminal Enforcement with Community Safeguards

The original 2021 version of South Korea's Article 15-2 provides a complementary historical example of criminal intervention against specified exploitative communications and enticement through communications networks (Act on the Protection of Children and Youth Against Sex Offenses, arts. 15-2(1)). Its relevance lies in proactively addressing conduct during the grooming process, rather than waiting for every form of abuse to be completed. This does not mean U.S. law lacks comparable offenses or that it should copy the Korean provision without examining existing law and constitutional limits. Still, US law may attempt to have criminal enforcement target grooming conduct, while platform safeguards can restrict opportunities for harmful contact. Alongside criminal intervention, prevention must also address the relationships and community influence through which exploitation can develop. Merry and Whitfield (2025) identify risks involving influential streamers and recommend safeguarding education and a clearer understanding of the apparent familiarity between streamers and young viewers (pp. 3324–3325). ECPAT International (2024) similarly emphasizes children's participation and cooperation among industry, civil society, and law enforcement (pp. 4–5). Education and community safeguards could supplement platform duties by addressing manipulation involving virtual gifts and influential figures. Whether the remedies requested by the states would prevent that manipulation remains untested.

Conclusion

The legal approaches examined in this study address several mechanisms of grooming and child sexual exploitation in online gaming, but they do not fully cover all mechanisms. The Louisiana and Kentucky complaints demonstrate how consumer-protection and negligence claims challenge misleading safety representations, inadequate warnings, and allegedly unsafe platform design (Commonwealth of Kentucky 61–66; State of Louisiana 37–40). Federal approaches include FTC enforcement against unfair or deceptive practices and COPPA's protections for personal information collected from children under thirteen (Federal Trade Commission; Johnson 420–421). Criminal provisions, shown through South Korea's 2021 approach to specified exploitative communications, target offenders, while international child-rights principles provide a broader framework for protection (Act on the Protection of Children and Youth Against Sex Offenses, arts. 15–2; Committee on the Rights of the Child, pars. 25, 80–82). The reviewed sources identify possible grounds for legal action, but do not establish whether these interventions have reduced exploitation.

Legal claims reflect important weaknesses in age verification, communication safeguards, and parental warnings, but addressing these weaknesses does not necessarily prevent

grooming. Correcting safety representations can improve families' understanding without preventing manipulative contact. Age verification can support communication restrictions without establishing that a user is safe, and safeguards within one platform cannot directly prevent dangers elsewhere. Section 230 may also limit claims that impose publisher liability for users' communications. Claims based on independently actionable platform conduct may fall outside that protection, but must still satisfy the requirements for liability (Brannon and Holmes 13–16).

Protective communication defaults, stronger privacy protections for teenagers, accessible reporting, and lawful cooperation between services could help address these gaps. These safeguards should accompany enforceable legal duties for platforms. Age verification should add to these protections, with criminal enforcement and education addressing manipulation and influential community relationships. Remedies should target failures and provide ways to assess whether the resulting changes improve children's safety.

Future research should follow the Louisiana and Kentucky lawsuits in order to verify which claims survive Section 230 and what remedies they produce. Moreover, researchers should distinguish between reductions in the number of victims and changes in reporting or detection. Researchers could also examine how services organize interventions while protecting children's privacy. Age checks and communication restrictions need to be assessed for their effects on safety and adolescent autonomy. Another question is whether safeguards involving streamers and community organizers can address manipulation that account restrictions miss.

Works Cited

Abd Jalil, Juriah. "Korean Approach to Online Protection for Children in the Digital Era." Conference presentation. 13th Australasian Conference on Child Abuse and Neglect: Protecting Children: New Solutions to Old Problems, Melbourne, Australia, https://irep.iium.edu.my/39992/.

Act on the Protection of Children and Youth Against Sex Offenses. Act No. 17972, Korean Legislation Research Institute, 23 Mar. 2021, Statutes of the Republic of Korea, https://elaw.klri.re.kr/eng_service/lawViewContent.do?hseq=56569.

Arneaud, Sydny K. "Video Games and the Federal Trade Commission Act: An Analysis of the Practices of Video Game Developers and Their Effects on Players." Florida A&M University Law Review, vol. 16, no. 1, 2022, p. 47.

Brannon, Valerie C., and Eric N. Holmes. Section 230: An Overview. CRS Report no. R46751, Congressional Research Service, 4 Jan. 2024, https://www.congress.gov/crs-product/R46751.

Carville, Olivia, and Cecilia D'Anastasio. "Roblox's Pedophile Problem." Bloomberg Businessweek, 22 July 2024. www.bloomberg.com, https://www.bloomberg.com/features/2024-roblox-pedophile-problem/.

Committee on the Rights of the Child. General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment. General Comment CRC/C/GC/25, United Nations, 2 Mar. 2021, UN Treaty Body Database, https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation.

Commonwealth of Kentucky. "Commonwealth of Kentucky v. Roblox, Inc. and Roblox Prepaid Company, LLC: Complaint." Complaint. Madison Circuit Court, 6 Oct. 2025, https://media.reclaimthenet.org/docs/ky-roblox-complaint-final-2025-10-06.pdf.

Communications Decency Act of 1996. U.S.C., vol. 47, nos. 104–104, 8 Feb. 1996, https://www.law.cornell.edu/uscode/text/47/230.

Doe v. Internet Brands, Inc. F.3d, vol. 824, no. 12-56638, 31 May 2016, p. 846, https://scholar.google.com/scholar_case?case=3822187665709132656.

Doe v. MySpace, Inc. F.3d, vol. 528, no. 07-50345, 16 May 2008, p. 413, https://scholar.google.com/scholar_case?case=7723885977876502124.

Drejer, Catharina, et al. It's All in the Game - an Exploration of Extensive Communication on Gaming Platforms and the Risks of Online Sexual Grooming. Jan. 2024. ResearchGate, https://doi.org/10.2139/ssrn.4671140.

ECPAT International. Action Report: Shifting the Narrative to Build a Culture of Child Safety in Online Gaming. Action Report. ECPAT International, 2024, https://ecpat.org/resource/action-report-shifting-the-narrative-child-safety-online-gaming/.

Emma S. "Online Gaming and Human Trafficking." The Exodus Road, 17 May 2024, https://theexodusroad.com/online-gaming-human-trafficking/.

Faraz, Anum, et al. “Child Safety and Protection in the Online Gaming Ecosystem.” IEEE Access, vol. 10, 2022, pp. 115895–913. IEEE Xplore, https://doi.org/10.1109/ACCESS.2022.3218415.

Federal Trade Commission. “Fortnite Video Game Maker Epic Games to Pay More than Half a Billion Dollars over FTC Allegations of Privacy Violations and Unwanted Charges.” Federal Trade Commission, Federal Trade Commission, 19 Dec. 2022, https://www.ftc.gov/news-events/news/press-releases/2022/12/fortnite-video-game-maker-epic-games-pay-more-half-billion-dollars-over-ftc-allegations.

Greijer, Susanna, and Jaap Doek. Terminology Guidelines for the Protection of Children from Sexual Exploitation and Sexual Abuse. Terminology guidelines. 2nd ed., ECPAT International, Mar. 2025, https://ecpat.org/wp-content/uploads/2025/04/Second-Edition-Terminology-Guidelines-final.pdf.

Hutchinson, Terry. “The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law.” Erasmus Law Review, edited by Sanne Taekema, vol. 8, no. 3, Dec. 2015, pp. 130–38. DOI.org (Crossref), https://doi.org/10.5553/ELR.000055.

Jarvie, Chelsea, and Karen Renaud. “Online Age Verification: Government Legislation, Supplier Responsibilization, and Public Perceptions.” Children [Basel, Switzerland], vol. 11, no. 9, Aug. 2024, p. 1068. Global Pediatric Health. PubMed Central, https://doi.org/10.3390/children11091068.

Johnson, Ariel Fox. “13 Going on 30: An Exploration of Expanding COPPA’s Privacy Protections to Everyone.” Seton Hall Legislative Journal, vol. 44, no. 3, 2020. Zotero, https://scholarship.shu.edu/cgi/viewcontent.cgi?article=1168&context=shlj.

Lemmon v. Snap, Inc. F.3d, vol. 995, no. 20-55295, 4 May 2021, p. 1085, https://scholar.google.com/scholar_case?case=6983497523645535721.

Louisiana State Legislature. Unfair Trade Practices and Consumer Protection Law. Louisiana Revised Statutes, https://www.legis.la.gov/legis/Law.aspx?p=y&d=104029.

Masters, Tim. “‘Exposing Hashinshin the Predator’: New Voyboy Video Brings Hashinshin’s Grooming Allegations Back to Light.” InvenGlobal, 12 July 2021, https://www.invenglobal.com/articles/14508/hashinshin-voyboy-drama-grooming-allegations-resurfacing.

Merry, Oliver J., and Kate C. Whitfield. “The Landscape of Sexual Harm in the Video Game, Streaming, and Esports Community.” Journal of Interpersonal Violence, vol. 40, nos. 13–14, July 2025, pp. 3307–32. SAGE Journals, https://doi.org/10.1177/08862605241271349.

Roblox Corporation. Quarterly Report for the Quarterly Period Ended June 30, 2026 (Form 10-Q). Form 10-Q nos. 001–39763, U.S. Securities and Exchange Commission, 30 July 2026, SEC EDGAR, Accession No. 0001628280-26-051082, p. 89, https://www.sec.gov/Archives/edgar/data/1315098/000162828026051082/rblx-20260630.htm

Rosen, Christine. “Playgrounds of the Self.” The New Atlantis, no. 9, 2005, pp. 3–27.

State of Louisiana. “State of Louisiana v. Roblox Corporation.” Petition for Injunctive Relief, Restitution, and Other Relief. 21st Judicial District Court, Parish of Livingston, State of Louisiana, 14 Aug. 2025, https://dcer237tfveol.cloudfront.net/img/wbrz/files/roblox_lawsuit.pdf.

Theil, Stefan. “Carefully Tailored: Doctrinal Methods and Empirical Contributions.” Oxford Journal of Legal Studies, vol. 45, no. 4, Aug. 2025, pp. 1047–75. DOI.org (Crossref), https://doi.org/10.1093/ojls/gqaf029.

An Exploration of the Link Between Music Preference, Self-Esteem, and Interpersonal Connection: Collective Self-Esteem By Tiffany Wang

Abstract

Music is undoubtedly a prevalent part of everyday life and has shown to be important to personal identities. Two areas of research, self-esteem and interpersonal connection, have separately emerged in relation to music preference (Rentfrow & Gosling, 2003; Shepherd & Sigg, 2015; Launay & Dunbar, 2015). This integrative literature review examined whether these two fields may be linked by a common mechanism. A random-effects meta-analysis conducted in this paper revealed no consistent relationship between personal self-esteem and preference for intense and rebellious music. Other findings suggest that this inconsistency may be explained by differences in gender, as men and women showed relationships in opposite directions. Separately, music preference was found to be a good predictor of the initiation, though not necessarily the strength and sustainment, of interpersonal relationships (Baym and Ledbetter, 2009). The review proposes that collective self-esteem shows a more consistent relationship with music preference and explains why music preference may initiate interpersonal relationships. Thus, music preference would also function as a way to show identification with a group. The paper recommends that future research on music preference directly examines personal self-esteem, collective self-esteem, and the outcomes of interpersonal relationships in a single study.

Introduction

Music is a universal aspect of everyday life. Across cultures, people often express themselves through the music they listen to. As studied by Shepherd and Sigg (2015), fans of classical music "may not feel motivated to wear clothes and regalia expressing their preference…while those who identify with the rock or rap genres may" (*p*. 512). Although public attention to the relationship between music preference and identity has existed for a long time, research on music preferences has remained limited until relatively recently.

Despite the longstanding relevance of music in culture and media, researchers only began studying the relationships between music preferences and human traits in the early 2000s, largely because social and personality psychology has historically overlooked everyday behaviors (Rentfrow & Gosling, 2003).

Previous research on the relationship between self-esteem and music preferences has produced mixed findings: some studies have found no significant association, whereas others have reported evidence of a correlation. For example, Swami and colleagues (2013) found a significant correlation between preferences for heavy metal and lower self-esteem among a population of 414 participants from the United Kingdom with survey responses collected from Facebook. However, Rentfrow and Gosling (2003) found no significant correlation between preferences for intense/rebellious music and self-esteem among undergraduates at the University of Texas at Austin, although they did find small but significant correlations between preferences

for upbeat/conventional and energetic/rhythmic music. The differences in the two studies' populations' cultures and ages may have affected the inconsistency; while Swami et al. (2013) included an English sample more diverse in age and background, Rentfrow and Gosling (2003) utilized a sample of American undergraduate students. This may suggest that there are additional factors, such as gender, affecting the findings, which this paper will go more in depth on.

The purpose of this integrative literature review is to examine the relationship between self-esteem and music preference and to explore how music preference can potentially predict connections between individuals, in order to evaluate the common link behind these two relationships: collective self-esteem. The research questions that will guide this literature review are: (1) What is the relationship between self-esteem and music preference?; (2) How does shared music preference influence interpersonal connections?; and (3) Is there an underlying link behind the relationship between self-esteem and music preference, and the relationship between interpersonal connections and music preference?

Music Preferences and Personal Self-Esteem

A recurring theme across several sources was the relationship between musical preferences and self-esteem. Although these constructs may initially appear unrelated, individuals often use their musical preferences to express, construct, and communicate aspects of their identity. Self-esteem is broadly defined as one's positive or negative evaluation of their self-worth (Rosenberg, 1965). Almost all studies included in this paper (Swami et al., 2013; Rentfrow & Gosling, 2003; Shepherd & Sigg, 2015; North, 2010) measured self-esteem through the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965), a 10-item self-report inventory rated on a four-point scale, which are summed up to provide a measure of self-esteem.

Several studies suggest a negative relationship between self-esteem and preferences for intense and rebellious music. Swami et al. (2013) measured self-esteem with a sample of 414 participants through the RSES. The authors found that those who exhibited a stronger preference for heavy metal correlated with lower self-esteem. The authors suggest that this may be because of the intense nature of heavy metal, which "allows for a purge of negative feelings…[to] help boost self-worth and promote positive self-evaluations" (Swami et al., 2013, *p*. 381). North (2010) similarly used the RSES and found a negative correlation between preferences for alternative rock music and self-esteem across a sample of $n = 36{,}518$ participants. Findings from this study align with Swami and colleagues' explanation of catharsis; ratings of self-esteem were related negatively to managing negative emotions, including "listening to music to relieve boredom, get through difficult times, be trendy, relieve tension, create an image, express feeling and emotions, please friends, and reduce loneliness" (North, 2010, *p*. 207). This agreement amongst the two papers may indicate that the relationship between music preferences and self-esteem may not be solely due to the genre itself, but also to psychological reasons that listeners may use to regulate emotions. However, it is worth noting that other studies such as Rentfrow and Gosling (2003), which also used the RSES, found no significant relationship between self-esteem and preferences for intense and rebellious music across two samples of $n =$

1,704 and n = 1,383 participants. This inconsistency between the studies suggests that the relationship found by Swami et al. (2013) and North (2010) may not be entirely reliable or generalizable, or that there are other variables affecting the studies.

Nevertheless, this relationship appears in other genres as well. Although Rentfrow and Gosling (2003) found no significant relationship for intense and rebellious music, the study did show small but significant relationships for upbeat and conventional, as well as energetic and rhythmic genres. North (2010) reports a positive correlation between self-esteem and genres such as rock and music of Black origin. Although North uses his own classification of ten distinct genres instead of Rentfrow and Gosling's four dimensions of music, this shows that other genres which may be categorized as intense and rebellious may not share the same negative relationship with self-esteem as alternative rock and heavy metal do.

The relationship between music preferences and self-esteem is further moderated by gender. Shepherd and Sigg (2015) observed a significant negative correlation among women between self-esteem and preference for intense and rebellious genres relative to reflective and complex genres, indicating that women with higher self-esteem tended to favor reflective and complex music. In contrast, men showed no significant relationship between self-esteem and preference for intense and rebellious genres relative to reflective and complex genres. In the study, men only showed a significant relationship for another genre pair, where self-esteem and relative preferences for intense and rebellious over upbeat and conventional music had a positive correlation. These results show that men and women may also differ in their basic genre preferences, not just in the relationship between music preferences and self-esteem. Altogether, the discrepancy indicates that gender may explain the inconsistent results discussed earlier, especially when men and women are pooled together. For example, although no significant relationship is found when the two genders are combined in Rentfrow and Gosling (2003), significant relationships between self-esteem and music preferences may exist within each gender separately.

Music Preferences and Interpersonal Connections

Discussions of music preferences are common in both initial acquaintance interactions and ongoing social exchanges. Music preferences are featured as standard profile prompts on virtually all dating platforms, while social media sites such as Instagram and TikTok provide mechanisms for users to share music through stories, short-form videos, and similar content formats.

Rentfrow and Gosling (2006) conducted a study on n = 60 University of Texas at Austin undergraduates over the course of 6 weeks to examine what topics people naturally gravitated towards as they got acquainted, including "books, clothing, movies, music, television shows, football, and sports other than football" as the study took place during football season (Rentfrow & Gosling, 2006, *p*. 237). Findings reported that music was the most frequently talked about topic across every week, and that it was not the most discussed topic only once in the first five weeks of the study. Although music may not be as conventional a conversation starter as

hobbies, these findings suggest that it serves as an accessible topic through which strangers can begin to get acquainted. This may imply that the vast number of artists, songs, and genres, as well as extended conversations on shared or differing opinions over the aforementioned, may explain why music was so commonly discussed among the students.

Another study conducted by Launay and Dunbar (2015) experimentally demonstrated the link between music preferences and interpersonal liking of strangers. The study introduced n = 294 participants to virtual partners that shared various, some, or no traits, which included music taste, hobbies, ethical statements, religion, and sense of humor, then measured the participants' fondness for each partner. Results from two separate experiments indicate that shared music tastes, alongside religion and ethical statements, "play a consistent role in determining the likeability of a similar stranger" (Launay & Dunbar, 2015, *p*. 10). The experimental nature of this study provides strong evidence that sharing music preferences causally, not only correlationally, leads to increased liking for a stranger. The study is consistent with findings from Rentfrow and Gosling (2006), suggesting that music taste is an important factor that people consider while forming friendships. Future studies should be conducted on whether certain shared genres can predict the ease of forming such friendships.

However, Baym and Ledbetter (2009) appear to show that while shared music taste may be important for initiating relationships, it is less important for the length or closeness of these relationships. The study surveyed n = 559 participants on Last.fm, "a music-focused…social network site," to study whether shared music taste can predict the strength of relations developed on the site (Baym & Ledbetter, 2009, *p*. 408). Findings show that friendship pairs did most often appear to have shared music tastes, meaning that shared music preferences fostered communication between users of Last.fm–which agrees with Rentfrow and Gosling (2006) and Launay and Dunbar (2015). However, this shared music preference did not correlate with the development of relationships. Instead, the development of relationships was most strongly predicted by the number of ways people communicated outside of Last.fm, whether it be through other apps, phone calls, or direct face-to-face conversations. These results may demonstrate that although shared music preferences can spark a connection between users, the actual depth of that connection does not hinge on this one singular factor. Building a meaningful relationship with others requires more investment and means of communication rather than solely what initially started that relationship.

Collective Self-Esteem as a Bridge Between Personal Identity and Interpersonal Relationships

Past studies have shown that personal self-worth and social connections each show relationships with music preference. This section of the paper aims to argue that personal self-worth and social connections are connected by the collective self-esteem a person gains from identifying with a community with shared music preference. This is especially relevant for adolescents, whom Frith (1983) argues "use music as a badge" to express their social identity (Frith, 1983, *p*. 217).

Clark and Lonsdale (2023) pointed out that much of the existing literature on this topic focused on personal self-esteem and "seems to ignore the links widely thought to exist between musical preferences, group membership, and social identity" (Clark & Lonsdale, 2023, *p*. 1121). Through testing this idea, the study found that collective self-esteem, rather than personal self-esteem, significantly predicted certain components of the 139 participants' music preferences. Specifically, results showed that private collective self-esteem, defined as an individual's appraisal of their social group, positively predicted preferences for intense and rebellious music. Additionally, the importance participants placed on the group identity aspect of their identity negatively predicted preferences for reflective and complex music. In addition to gender, these findings may further explain the inconsistency discussed in previous research on personal self-esteem and music preference. Thus, music preferences may show a stronger relationship to the value one places on their music-based social group.

These findings also support the relationship between forming interpersonal connections and shared music preferences discussed by Rentfrow and Gosling (2006), Launay and Dunbar (2015), and Baym and Ledbetter (2009). Repeating Frith's (1983) claim that adolescents use their music preferences as a badge of the groups they belong to, one's level of self-esteem may also naturally affect how they evaluate others based on shared or differing music tastes. Bakagiannis and Tarrant (2006) studied how music can influence the growth of more positive intergroup feelings through 97 high school students, manipulating whether they perceived their in-group as having similar or different music tastes to the out-group. Findings reported that participants who were told that the two groups shared similar music preferences showed "more positive intergroup perceptions" (Bakagiannis & Tarrant, 2006, *p*. 134). This study applies a more group-oriented perspective to the interpersonal studies discussed previously. Tarrant et al. (2001) bridge these findings to the self-esteem aspect discussed by Swami et al. (2013), Rentfrow and Gosling (2003), Shepherd and Sigg (2015), and North (2010), investigating English male adolescents to measure their evaluations of the in-group and out-group, as well as music preferences of the groups. Supporting the social identity theory, results show that participants with lower self-esteem expressed "greater differentiation between groups and greater derogation of the out-group" (Tarrant et al., 2001, *p*. 565). This suggests that self-esteem may not only shape one's music preferences, but it may also predict how strongly they separate themselves from others based on these preferences.

Overall, these findings suggest a link between the relationship of music preferences and self-esteem and the relationship of music preferences and interpersonal connections. Self-esteem, whether personal or collective, does not simply predict the music one enjoys; it may reveal a broader process that begins in an individual's personal identity, then moves to their group identifications, and culminates in the friendships and connections they form through a musical community.

Theoretical Frameworks

This paper is guided by two theoretical frameworks: social identity theory and similarity-attraction theory.

Social identity theory (SIT; Tajfel, 1978; Tajfel & Turner, 1986) explains how an individual's sense of identity is derived from the social groups they belong to. The theory suggests that individuals tend to evaluate their own group more favorably than other groups, a tendency known as in-group favoritism, to protect their self-esteem and social identity. Although the theory was originally designed to explain identity formed around categories, such as ethnicity, it can be applied to groups formed around shared interests as well. SIT informs the discussion surrounding personal self-esteem and music preference in this paper by explaining how an individual's self-esteem may be related to their musical groups. The theory likewise informs the discussion of collective self-esteem by showing why identifying with a musical group may lead the individual to show favoritism to others in that group.

Similarity-attraction theory (Byrne, 1961) suggests that an individual is more attracted to those whom they perceive as similar to themselves, whether it be personality traits, interests, or views. According to the theory, similarity between individuals increases interpersonal liking, even without any previous relationship or interaction. The theory explains the broader principle of homophily, the tendency for people to bond with those similar to themselves. This theoretical framework informs the discussion on the relationship between interpersonal connections and music preference, which explores how shared music taste affects interpersonal liking, including the strength and duration of this liking.

Methods

For this literature review paper, sources were identified on Google Scholar or directly through JSTOR. Search terms included, but were not limited to, "music preferences", "self-esteem", "social identity", and "interpersonal connections." Date limits were not imposed in this paper, and included sources ranging in publication year from 1961 to 2023. 20 sources were originally identified, and 15 were ultimately included. The sources used were required to be peer-reviewed or established theoretical works. The only exclusion criteria of this paper was non-English studies. Most included sources used cross-sectional correlational designs by collecting responses through questionnaires regarding music preferences and self-esteem; Swami et al. (2013) specifically had participants rate preferences after listening to audio clips instead of only basing ratings on genre names, and Rentfrow and Gosling (2006) used naturalistic observation to examine conversation topics. Two included sources, Bakagiannis and Tarrant (2006) and Launay and Dunbar (2015), used experimental designs in their studies on the relationship between music preference and interpersonal connection. Studies were organized by their correspondence to the relationship between personal self-esteem and music preference, the relationship between interpersonal connections and music preference, the relation of collective self-esteem or group-based identities to music preference, and theoretical works.

Results

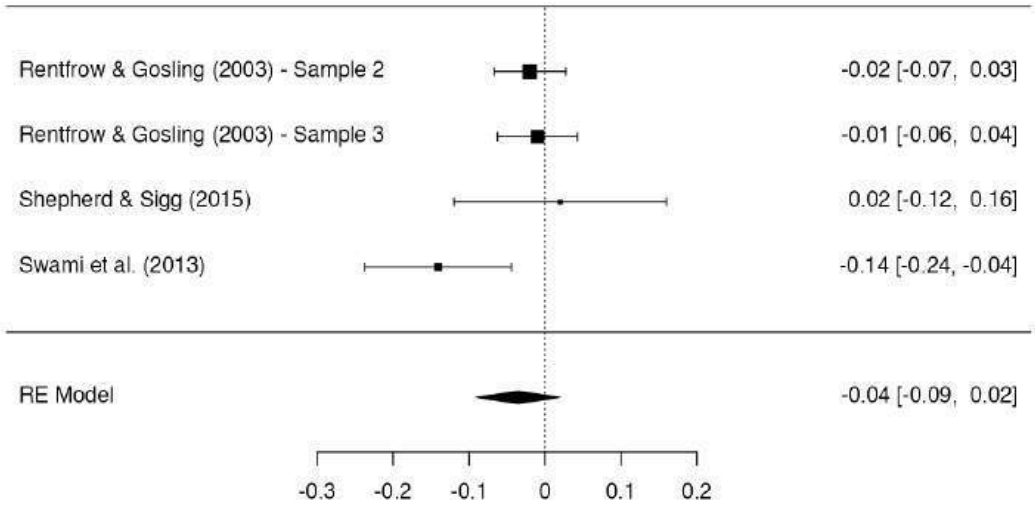
This literature review examined the relationships between personal self-esteem and music preference, interpersonal connections and music preference, and how collective self-esteem may explain and relate these two relationships. The studies reviewed in this paper found an inconsistent relationship between self-esteem and music preferences, as well as the ability of shared music preference to predict the initiation of interpersonal connections, though not the closeness, strength, or maintenance of the connections. More details can be found in Tables 1 and 2 below.

Table 1. Summary of studies examining the relationship between self-esteem and music preferences.

Study	Sample Size	Binary Gender	Genres Tested	Findings
Rentfrow & Gosling (2003)	n = 1,704 & 1,383	All	Intense/Rebellious	r = -.02, -.01
			Upbeat/Conventional	r = .07*, -.05
			Energetic/Rhythmic	r = .06*, -.04
North (2010)	n = 36,518	All	Alternative Rock	β = -.021
			Rock	β = .002
			Music of Black Origin	β = .030
Swami et al. (2013)	n = 414	All	Heavy Metal	r = -.14*
Shepherd & Sigg (2015)	n = 199	Males	Intense/Rebellious vs. Upbeat/Conventional	r = .34*
		Females	Intense/Rebellious vs. Reflective/Complex	r = -.15*
		All	Intense/Rebellious vs. Reflective/Complex	r = 0.02

Note. All studies measure self-esteem through the RSES. For Rentfrow & Gosling (2003), the first value in the "Finding" column reflects Sample 2 (n = 1,704), and the second value reflects Sample 3 (n = 1,383). North (2010) does not report significance values.
* $p < .05$

Figure 1. Forest Plot of Random-Effects Meta-Analysis of Personal Self-Esteem and Music Preference.



Note. Squares represent the effect size of each individual study. Sizes of the squares represent relative weight in the pooled estimate. Horizontal lines represent 95% confidence intervals. The diamond represents the pooled random-effects estimate.

To further examine the relationship between personal self-esteem and music preference, a random-effects meta-analysis was conducted using four correlations from three studies (Rentfrow & Gosling, 2003; Shepherd & Sigg, 2015; Swami et al., 2013). These studies measured self-esteem using the RSES and reported zero-order correlations with preference for intense/rebellious music. Findings from the analysis suggest that there is no statistically significant relationship between personal self-esteem and preference for intense/rebellious music, with a pooled estimate of r = -0.04 (SE = 0.028), 95% Confidence Interval [-0.09, 0.02], z = -1.26, p = 0.208. Heterogeneity among the studies is moderate (I^2 = 56.26%), but the Q-test for heterogeneity is not statistically significant ($Q(3)$ = 6.25, p = 0.100). This may be due to the limited number of studies. A 95% prediction interval for the true outcomes ranged from -0.13 to 0.06. This demonstrates that the true relationship between personal self-esteem and preference for intense/rebellious music may be either positive or negative in different populations or contexts. This corroborates the idea found by Shepherd and Sigg (2015) that gender may explain the relationship's inconsistent results.

Table 2. Summary of studies examining the relationship between interpersonal connections and music preferences.

Study	Sample Size	Outcome Tested	Finding
Rentfrow and Gosling (2006)	*n* = 60	Frequency of conversation topics	Music was the most frequently talked about topic in 5 of 6 weeks
Launay and Dunbar (2015)	*n* = 294	Liking of strangers	Shared music tastes, alongside religion and ethics, were the strongest predictors of liking
Baym and Ledbetter (2009)	*n* = 559	Strength of relations developed on Last.fm	Shared music taste did not predict how close the relationships became

Discussion

This paper's evidence suggests that the relationship between self-esteem and music preference is inconsistent in direction and significance across different factors such as genre and gender. Shared music taste was found to influence the initiation of interpersonal connections through increased liking towards strangers and more available conversation topics, though it did not necessarily predict the strength or sustainment of these connections. Overall, the evidence found in this review suggests that the relationship among self-esteem, music preference, and interpersonal connections may have one possible explanation in collective self-esteem.

Findings from Rentfrow and Gosling (2003), North (2010), Swami et al. (2013), and Shepherd and Sigg (2015) demonstrated an inconsistent relationship between self-esteem and music preference. The inconsistent relationship between self-esteem and music preference may be influenced by factors such as gender, the type of self-esteem measured, musical genre, or the social context in which music is used. Findings from Rentfrow and Gosling (2006), Launay and Dunbar (2015), and Baym and Ledbetter (2009) showed that while shared music taste predicted the initiation of friendships, it could not predict the closeness, strength, or maintenance of the relationships. These findings suggest that the relationship between music preference and interpersonal connections is strongest during the relationship's initial stages due to shared music taste influencing increased liking towards strangers and increased conversation topics. Lastly, findings from Clark and Lonsdale (2023), Bakagiannis and Tarrant (2006), and Tarrant et al. (2001) revealed that collective self-esteem may help explain these two areas of research. These studies show that music preference may relate more to how individuals appraise their belonging to a musical group rather than how individuals feel about themselves. This group identity also shapes how favorably an individual views other groups based on shared or differing music tastes, with personal self-esteem affecting the strength of this favorability.

Interpretation of Findings

The inconsistent relationship between personal self-esteem and music preference determined in prior studies reflects a limitation in how music preference relates to personal self-esteem. This is supported by the random-effects meta-analysis (see Figure 1), which indicated no overall statistical significance. The meta-analysis' wide prediction interval suggests that the expected effects substantially vary across different populations or future studies; this variability may be explained by factors such as gender or differences in musical genres. Furthermore, Clark and Lonsdale's (2023) study found that collective self-esteem, instead of personal self-esteem, meaningfully predicted preference for intense and rebellious music. This suggests that measures of self-esteem that do not take into account group membership may be the reason for the inconsistencies seen between personal self-esteem and music preference. Studies conducted by Rentfrow and Gosling (2006) and Launay and Dunbar (2015) show that music is not only a common topic of discussion among strangers getting acquainted, but shared music taste is also a predictor of liking towards these strangers. However, Baym and Ledbetter (2009) show that although shared music taste does predict the initiation of these relationships, it may not be sufficient to explain the closeness, strength, or maintenance of the relationships. Other factors, such as sustained communication, shared experiences, or similarity in values, provide a better predictor for the quality of relationships.

These findings support the current literature review's position that self-esteem and interpersonal relationships are linked to music preference through shared musical identities and the collective self-esteem arising from these identities. This conclusion is further supported by Bakagiannis and Tarrant (2006) and Tarrant et al. (2001). Bakagiannis and Tarrant (2006) demonstrated that believing one's music taste aligned with a different group's music taste improved intergroup attitudes, and Tarrant et al. (2001)'s study revealed that personal self-esteem determined how the improved liking is expressed or not expressed. Together, these studies suggest how personal identity, group identity, and interpersonal relationships are all threaded by music preference.

The non-significant pooled relationship found in this literature review is consistent with previous research on the topic by Rentfrow and Gosling (2003), who also found little to no correlation between music preference and personal self-esteem depending on the musical genre studied. However, these results differ from Swami et al. (2013) and North (2010), whose studies both reported small but statistically significant relationships. This disagreement may be due to the differing methods used; Swami et al. asked participants to rate audio clips of metal music, while Rentfrow and Gosling used the Short Test of Music Preference (STOMP; Rentfrow & Gosling, 2003), in which participants rated their enjoyment for a variety of music genres.

This paper suggests that collective self-esteem may better relate to music preference than personal self-esteem, which supports the social identity theory, which says that identity and self-esteem are derived from one's social groups. Thus, Clark and Lonsdale (2023), which measured collective self-esteem, is a more accurate evaluation of the social identity theory than the prior studies that used personal self-esteem. Likewise, the study's findings agree with the similarity-attraction theory, which argues that an individual shows more attraction towards those

they perceive similar to themselves. Because shared music taste predicted the initiation of but not the strength or sustainment of relationships in Baym and Ledbetter (2009), the similarity-attraction theory's boundaries may need to be further clarified.

The findings of this paper may benefit research through a better understanding of the psychological and social effects of music preference, especially in relation to personal and collective self-esteem and interpersonal relationships. Additionally, this paper can benefit educators and others who work with adolescents, whose musical identities studies show to be particularly important, by informing how shared musical tastes can serve as a potential way to foster initial social connections with others. More broadly, this paper can contribute to a better understanding of how personal and social identities overlap with everyday behaviors, such as the music people listen to.

Strengths and Limitations

The main strength of this paper includes the variety and magnitude of sample sizes in the studies included, such as the sample size of n = 36,518 participants in North (2010). The paper also included a large set of English-speaking countries, such as the United States, United Kingdom, and New Zealand, throughout the studies, which may improve generalizability in these countries.

There are several limitations that must be taken into account in this review. First, while most studies in this paper used the Rosenberg Self-Esteem Scale to measure self-esteem, Clark and Lonsdale (2023) used the Collective Self-Esteem Scale. This leaves the findings of this paper between personal and collective self-esteem more inferential, rather than direct. Second, within the meta-analysis conducted in this paper, Shepherd and Sigg (2015) measured music preference differently than the other studies by using relative difference scores between two music factors (such as Intense/Rebellious vs. Energetic/Rhythmic). Third, the meta-analysis was based only on four correlations from three studies, resulting in a sensitive pooled estimate and lower statistical power. North (2010) was not able to be included due to reporting mean standardized beta weights instead of zero-order correlations, and this difference could not be resolved due to insufficient reporting of other test statistics. Fourth, many included studies included self-report measures for self-esteem and music preference, which may potentially affect results due to factors such as social desirability bias. Lastly, although the backgrounds of participants examined in this paper ranged throughout many different countries, they were primarily English-speaking populations, in countries such as the United Kingdom, the United States, and New Zealand. Thus, the findings of this paper may be limited when imposed onto different cultures.

Directions for Future Research

The findings of this paper recommend that collective self-esteem should be more closely examined in future research on the relationship between music preference and personal self-esteem. Additionally, findings can inform those working with adolescents, whose musical

identities have shown to be particularly important to individuals of this age group, as shared music tastes can serve as a way to build relationships with peers.

Future research should examine personal and collective self-esteem, interpersonal connection, and music preference within one study in order to clarify whether collective self-esteem truly mediates the two research areas of personal self-esteem's relationship with music preference and music's role in forming social connections. This paper proposes a longitudinal or mediation study for future research on the topic, the latter of which could test if collective self-esteem statistically explains the relationship between music preference and personal self-esteem. To determine whether the process differs by age, a longitudinal design could examine if outcomes change over an individual's lifespan. To more closely examine the moderating variable of gender as examined by Shepherd and Sigg (2015), future research should conduct direct comparisons between subgroups of men and women in order to address whether collective self-esteem may operate differently for the two binary genders. As studies in this paper focused on Western and English-speaking populations, future studies should identify if findings can be generalized to other contexts through cross-cultural research with culturally appropriate measures and more diverse samples.

Conclusions

Overall, the findings suggest that collective self-esteem may represent a potential mechanism linking music preference with both personal self-esteem and interpersonal connection. Because previous research has generally identified weak or nonsignificant associations between personal self-esteem and music preference, future research should examine whether music contributes more strongly to self-evaluation through group identification, belonging, and collective self-esteem. Studies should also investigate how collective self-esteem relates to the formation, strength, and maintenance of interpersonal connections, rather than focusing solely on whether shared musical preferences initiate relationships. In addition, future work should consider whether these processes vary across gender groups, cultures, age groups, and social contexts. Such research would provide a more comprehensive understanding of how music functions not only as a form of individual identity expression but also as a means of fostering social belonging and connection.

Works Cited

Bakagiannis, S., & Tarrant, M. (2006). Can music bring people together? Effects of shared musical preference on intergroup bias in adolescence. Scandinavian Journal of Psychology, 47(2), 129–136. https://doi.org/10.1111/j.1467-9450.2006.00500.x

Baym, N. K., & Ledbetter, A. M. (2009). Tunes that Bind?: Predicting Friendship Strength in a Music-Based Social Network. Information, Communication & Society, 12(3), 408–427. https://doi.org/10.1080/13691180802635430

Byrne, D. (1961). Interpersonal attraction and attitude similarity. The Journal of Abnormal and Social Psychology, 62(3), 713–715. https://doi.org/10.1037/h0044721

Clark, A. B., & Lonsdale, A. J. (2023). Music preference, social identity, and collective self-esteem. Psychology of Music, 51(4), 1119-1131. https://doi.org/10.1177/03057356221126202

Frith, S. (1983). Sound effects: Youth, leisure and the politics of rock “n” roll. Constable.

Launay, J., & Dunbar, R. I. M. (2015). Playing with strangers: which shared traits attract us most to new people? PLoS ONE, 10(6), Article e0129688. https://doi.org/10.1371/journal.pone.0129688

North, A. C. (2010). Individual Differences in Musical Taste. The American Journal of Psychology, 123(2), 199–208. https://doi.org/10.5406/amerjpsyc.123.2.0199

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. Journal of Personality and Social Psychology, 84(6), 1236–1256. https://doi.org/10.1037/0022-3514.84.6.1236

Rentfrow, P. J., & Gosling, S. D. (2006). Message in a Ballad: The Role of Music Preferences in Interpersonal Perception. Psychological Science, 17(3), 236–242. http://www.jstor.org/stable/40064524

Rosenberg, M. (1965). Society and the adolescent self-image. Princeton University Press.

Shepherd, D., & Sigg, N. (2015). Music Preference, Social Identity, and Self-Esteem. Music Perception: An Interdisciplinary Journal, 32(5), 507–514. https://doi.org/10.1525/mp.2015.32.5.507

Swami, V., Malpass, F., Havard, D., Benford, K., Costescu, A., Sofitiki, A., & Taylor, D. (2013). Metalheads: The influence of personality and individual differences on preference for heavy metal. Psychology of Aesthetics, Creativity, and the Arts, 7(4), 377–383. https://doi.org/10.1037/a0034493

Tajfel, H. (1978). Differentiation between social groups. Academic Press.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behaviour. In S. Worchel & W. Austin (Eds.), Psychology of intergroup relations (pp. 7–24). Nelson-Hall.

Tarrant, M., North, A. C., & Hargreaves, D. J. (2001). Social categorization, self-esteem, and the estimated musical preferences of male adolescents. The Journal of Social Psychology, 141(5), 565–581. https://doi.org/10.1080/00224540109600572

Evaluating Visual Perception Architectures for Autonomous Vehicles
By Vihaan Siramdas

Abstract

While Convolutional Neural Networks have been the favored architecture for computer vision, developers have been shifting to Vision Transformers (ViTs) for their higher accuracy. However, ViTs have higher computational complexity and struggle with less training data. For autonomous vehicles (AVs), this trade-off is critical as misclassifications or latency delays can directly impact passenger safety and real-time decision-making. Existing work has focused on benchmarks rather than real-world AV deployment constraints and does not compare architectures relative to embedded vehicle systems. This paper aims to bridge this gap by determining which architecture best balances accuracy, latency, and computational efficiency for autonomous vehicles. We did this by pinpointing AI failures in autonomous vehicle accidents using the AI incident database. Then we evaluated the speed and accuracy of architectures by reviewing papers. This review of thirteen papers shows that while ViTs provide better object detection accuracy and spatial context, hybrid CNN-Transformer architectures offer the best balance of low latency and low power consumption required. By establishing trade-offs and benefits between perception accuracy and latency, this paper provides developers and researchers with a framework for selecting and optimizing visual perception models for autonomous systems.

Section 1: Introduction

Autonomous vehicles (AVs) operate in high-stakes environments where their real-time perception is directly tied to the safety of passengers and other cars. Vision systems need to process various input streams such as cameras, LiDAR, radar, and ultrasonic sensors to detect and distinguish pedestrians, vehicles, hazards, and road signs in fractions of a second. At freeway speeds of 65 mph, a car moves at 95 feet per second. If the car had to brake in an emergency, having just 50ms of model latency means that a vehicle would travel nearly 5 feet before braking. When these delays are paired with occlusion, bad weather, or edge-case obstacles, the consequences can be deadly.

Additionally, these systems are constrained by the limitations of the hardware itself. Autonomous cars process around 1-5 terabytes of data every hour through cameras, LiDAR, radar, and more, depending on the model.[1] Offloading this to cloud infrastructure is not possible due to variable network latency and connectivity across the world. Due to this, perception systems must run on onboard edge hardware.

For years, Convolutional Neural Networks (CNNs) have been used as the primary architecture for vision. By processing image regions with sliding kernels, CNNs have linear time complexity $O(N)$ relative to the pixel resolution N. This makes CNNs fast and optimal for low-latency edge deployment. However, developers have recently shifted architectures with the

[1] Liu et al., “Edge Computing for Autonomous Driving.”

emergence of Vision Transformers (ViTs) to address limitations of convolution. CNNs are less effective at modeling long-range dependencies than Transformers and struggle with complex driving environments. Instead of using convolutions, ViTs use global self-attention, allowing them to find relationships across an entire image. This is valuable for vehicles as ViTs can connect two events happening on opposite sides of the AV's vision, like a cyclist swerving on the left side and a delivery truck blocking the cyclist, instantly. In contrast, a convolutional neural network would need several layers to realize they are related. Comparing these two architectures shows that given sufficient pre-training data, ViTs outperform CNNs when fine-tuned for classification tasks. However, standard ViTs have the drawback of having quadratic computational complexity, $O(N^2)$, and are worse on smaller datasets[2].

To evaluate how modern models balance perception accuracy against latency and real-time constraints, this review evaluates four distinct architectures:

- YOLOv8: A real-time CNN
- Standard ViT: A pure global Transformer
- EfficientViT: An edge hybrid that combines CNN feature extraction with self-attention
- BEVFormer: A spatiotemporal Transformer that applies spatial and temporal attention and maps camera feeds into a 3D Bird's Eye View (BEV) space.

This paper aims to find an architecture that AVs can use to mitigate accidents with minimal latency, high accuracy, and low hardware burden. In Section 2, we will explain how CNNs and ViTs work and how they handle computer vision. In Section 3, we will explore AI ethics in AVs and pinpoint AI failures in autonomous vehicle accidents using the AI incident database. In Sections 4 and 5, we evaluate the speed and accuracy of architectures by reviewing papers and benchmarks. Finally, using our findings, we can accurately find a model that balances all three elements. By establishing trade-offs and benefits between perception accuracy and latency, this paper aims to provide developers and researchers with a framework for selecting and optimizing visual perception models for autonomous systems.

Section 2: CNNs vs ViTs

2.1 How CNNs and ViTs work

CNNs and Vision Transformers are the two dominant architectures in computer vision. In fact, nearly every modern vision model is a direct implementation of one, or a hybrid of both. CNNs focus more on efficiency and assume that nearby pixels are related, allowing them to learn on small datasets with lower computational complexity. ViTs focus on accuracy and instead use self-attention to connect pixel patches of an image to each other simultaneously. Feeding an

[2] Kaya et al., "A Comparative Study of Vision Transformers and CNNs for Few-Shot Rigid Transformation and Fundamental Matrix Estimation."

image into a standard Multi-Layer Perceptron Model (MLP) requires flattening the 2D image into a 1D vector, with each pixel as an input feature. Connecting thousands of pixels to neurons in hidden layers creates millions or billions of trainable weights that the computer must compute, making the process very inefficient. Additionally, flattening destroys 2D spatial context, preventing the model from learning spatial patterns.

To solve this, CNNs use a process called a convolution. In a convolution, the CNN slides a matrix of weights (e.g., 3x3 or 5x5), called a filter, across the image. By calculating the dot product between the filter and the input pixels at each step, the network reuses weights to detect features like edges and corners in the image. By using a sliding matrix of neurons rather than a neuron per pixel, this process drastically reduces computational complexity and parameter count while preserving the image's 2D spatial structure. Fig. 1 shows a convolution in which an input vector is passed through the filter to produce an output matrix. By stacking several convolutional layers, CNNs can progress from seeing basic shapes like lines and dots to complex objects like faces and cars.

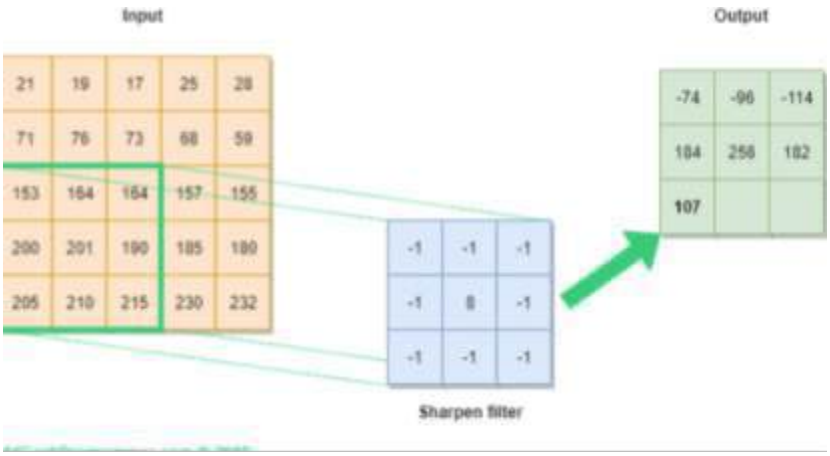


Fig. 1: CNN visualization

The Vision Transformer, introduced by Dosovitskiy et al.[3], uses a transformer architecture for computer vision. This process is identical to using transformers in LLMs, where one could think of each pixel as a letter and 16x16-pixel patches as words. The model flattens each patch into a 1D vector, which is passed through a linear layer to turn the raw pixel numbers into features the transformer can interpret; this is called patch embedding. Positional embeddings and classification tokens are also added to the sequence before it is passed to the encoder.

The encoder has two main parts: Multi-Head Self-Attention (MHSA) and the MLP. In CNNs, the field of view grows layer by layer as the sliding filters overlap. In contrast, Multi-Head Self-Attention allows patches to compare themselves with all other patches directly, providing a global view from the first layer. After MHSA, each token passes through the MLP. This process repeats through several layers (typically 12-32) of encoders. Once the token passes through the last encoder layer, spatial tokens are dropped. Finally, this vector is fed into a classification head, which outputs the final class probabilities.

2.2 Classification and Localization

[3] Dosovitskiy et al., “An Image Is Worth 16x16 Words.”

Classification is predicting the class of an object. In autonomous vehicles, this would be the computer labelling people, cars, and signs. Localization is predicting the location of the object in the environment. These are the two main goals of computer vision.

CNNs handle classification and localization by splitting into two branches of connected layers. The classification head outputs a probability distribution over the given object categories. At the same time, the localization head performs regression to predict four numbers representing a box around the object (The center coordinates, width, and height). YOLOv8 simplifies this process by dividing the image into a grid, allowing the network to process the whole image in a single forward pass[4].

ViTs were originally used for classification but have been adapted for localization by preserving spatial tokens rather than removing them. By passing the tokens into a transformer decoder, the model can find relationships across all patches, predicting the class labels and bounding box coordinates.

Section 3: Ethics and Real World Events

3.1 AI Ethics in Autonomous Vehicles

The integration of AI in autonomous vehicles promises to revolutionize transportation by enhancing safety, efficiency, and reclaiming time otherwise spent driving. In April 2026, the CEO of Tesla, Elon Musk, made the claim that Tesla's Full Self-Driving reduces traffic fatalities globally by 90% on X. However, claims like these have drawn scrutiny that reveals a disconnect between marketing promises and the reality of autonomous driving. Tesla's own reported data to the NHTSA shows that its unsupervised robotaxi fleet has a crash rate approximately four times higher than the average human driver[5] (One crash every 57,000 miles for the Robotaxi vs one minor crash every 229,000 miles and one major crash every 699,000 miles for humans). Additionally, an availability heuristic shapes public perception, as AV crashes receive disproportionately more media coverage than human-driven accidents, amplifying ethical concerns. Ultimately, the issue of AI driving is a perception and decision-making failure. For a vehicle to navigate safely, its visual perception system must accurately detect and classify objects with low latency to allow for sufficient reaction time. Therefore, the ethical concerns of AVs are fundamentally a concern of the technical capabilities of vehicles, as the choice of perception architecture has a direct impact on the safety outcomes.

3.2 Accountability

The NHTSA (National Highway Traffic Safety Administration) decides accountability based on 5 levels of driving systems[6]. Level 1 is a driver-support feature that's for either steering or braking/accelerating, but not both simultaneously. An example of this is adaptive cruise control, where a car maintains a safe distance from other vehicles by controlling its speed. Level

[4] Jocher, "YOLOv8 Documentation."
[5] Lambert, *Tesla 'Robotaxi' Adds 5 More Crashes in Austin in a Month — 4x Worse than Humans*.
[6] "Standing General Order 2021-01."

2 is when a car can control braking/accelerating and steering at the same time. Most AVs, such as Tesla vehicles with FSD, are classified as Level 2 by NHTSA. In both Level 1 and 2, the human driver must remain fully and continuously engaged in the driving task. In Level 3, the vehicle has full control over driving under some circumstances. However, the human driver must be ready to take back control at any time if the AV requests the human driver to do so. The Mercedes-Benz Drive Pilot is a Level 3 system. Level 4 vehicles can perform all driving tasks and monitor the driving environment without the need for a human driver to take over within the confines of specific cities and mapped zones. Waymo and Zoox both operate as Level 4 AVs within certain cities like San Francisco. Finally, Level 5 systems have total control of the vehicle and can operate in any location; humans are treated as passengers. Currently, there are no Level 5 systems yet.

Accountability varies between levels. In Levels 1-2, the human driver is almost always at fault, unless the hardware or software clearly malfunctioned. In Level 3, the liability is shared. If the system gives inadequate warning, the manufacturer is at fault; otherwise, it will be the driver. For Levels 4-5, the manufacturer is always at fault.

3.3 Real World Cases

In April 2026, six reported cases of Waymo AVs did not recognize and drove past ramp closure signs into pre-planned freeway construction zones in Phoenix, AZ. On June 17, 2026, Waymo recalled over three thousand vehicles due to AVs entering and driving at speed in freeway construction zones. Waymo attributes this issue to the software prioritizing the avoidance of other freeway hazards and/or failing to recognize the construction zone[7].

In another incident on June 20, 2026, an unoccupied Zoox AV encountered heavy smoke that obscured an active emergency fire scene that was not cordoned off with cones. The Zoox vehicle entered the scene and then braked hard while attempting to steer away before coming to a stop. After a recall from June-July, Zoox identified the problem as the software being unable to detect and respond to heavy smoke[8].

Table 1: More real-world cases

Manufacturer and Recall Date	Description	Possible Cause
Waymo LLC 5/1/2025 [1]	On May 13th 2024, the NHTSA opened an evaluation on Waymo's fleet, citing 7 events between 12/4/2022 and 4/25/2024 where the AVs would collide with stationary/semi-stationary objects like gates and chains. Waymo cited 9 more from 2/28/2024 to 12/11/2024.	Sensors like LiDAR and cameras struggle to classify thin objects like chains as impassable, especially under poor lighting, occluded conditions, and low contrast. Even when the car detected the barrier, the perception model failed to label the object and instead classified it as noise.

[7] "Part 573 Safety Recall Report."
[8] "Part 573 Safety Recall Report - Zoox.

Waymo LLC 12/10/2025 [2]	Waymo vehicles that were stopped or stopped for a school bus with its red lights flashing and/or stop arm extended would proceed again before the school bus had deactivated its lights/arm. There are 12 citations related to this behavior between 9/25/2025 and 11/14/2025	The software prioritized moving to avoid impeding vehicles over remaining stationary for a school bus.
Waymo LLC 5/6/2026 [3]	On 4/20/2026, an unoccupied Waymo AV encountered an untraversable flooded section of a roadway that has a 40 mph speed limit. The Waymo detected the untraversable floodwater but still proceeded at reduced speed.	The perception system correctly identified the flood but failed to accurately calculate the water depth and boundaries. This caused the driving system to assign the flood as low danger.
ZOOX 5/22/2025 [4]	In cases where the Zoox vehicles start moving, and a prone VRU (Vulnerable Road User, e.g., a biker) is adjacent to the vehicle, there may be an increased risk of injury. On May 8, 2025, in San Francisco, an unoccupied, stopped Zoox robotaxi was struck by an e-scooterist. The e-scooterist fell to the ground next to the vehicle. The robotaxi then started to move, but did not make contact with the e-scooterist	The perception system experienced a tracking state latency when the VRU fell into a horizontal position directly adjacent to the vehicle's body. This caused the perception system to fail to classify the fallen road user as an active obstacle
ZOOX 12/29/2025 [5]	In some lane line crossings, a Zoox vehicle may come to a stop in front of the oncoming lane. On 8/26/2024, a Zoox robotaxi made a wide right turn, crossed partially into the opposing travel lane, and temporarily stopped in front of the oncoming travel lane.	Zoox cited several reasons, including unexpected routing changes, mislabeled double-parked vehicles, and poorly executed attempts to avoid blocking traffic in intersections.

In all the cases, one or more of three failures can be found: sensor failures, perception failures, and planning errors. These real-world cases show the importance of vision software in autonomous navigation. In each instance, the vehicle's perception system failed to correctly classify or understand its environment. From these incidents, we can see that a robust vision system must resolve three key challenges: maintain high accuracy for static obstacles (e.g., chains, floods, and stop arms), contextualize complex scenes (e.g., construction zones, emergencies), and maintain spatial awareness against faulty inputs (e.g., heavy smoke, flooding), all while operating under latency and computational constraints.

Section 4: Accuracy and Spatial Perception

4.1 Accuracy

In general, ViTs have similar accuracy to CNNs. In the paper by Dosovitskiy et al.[9], the researchers compared three ViTs (ViT-B with 12 layers, ViT-L with 24 layers, and ViT-H with 32 layers) against the leading CNN, ResNet. While transformers consistently scored higher than CNNs in terms of accuracy, the difference wasn't large. The biggest difference was 94.55% for ViT-H and 93.25% for ResNet, and the smallest was 99.74% for ViT-L and 99.63% for ResNet.

[9] Dosovitskiy et al., "An Image Is Worth 16x16 Words."

A big caveat of this difference is that ViTs only beat CNNs when pre-trained on massive datasets. Without large-scale pre-training, standard ResNets outperform ViTs.

4.2 Occluded Objects

At night in 2018, a test AV owned by Uber crashed and hit a woman jaywalking[10]. According to the National Transportation Safety Board (NTSB), the vehicle misclassified the woman several times as she was walking her bicycle and holding bags. Additionally, the on-board software had no classification for pedestrian jaywalking. Since the woman was partially covered by the bicycle and the bags, the AI system couldn't correctly classify what was in front of it. This is object occlusion, when a person or object is partly or fully hidden from a camera's view by another object. There are three types of object occlusion. Multiple objects overlapping is called Inter-object Occlusion. For example, when a truck obscures a car. Background Occlusion is when a stationary element obscures the object, like a car passing behind a tree. Self-Occlusion refers to one part of an object blocking another part of the same object, such as when a person covers their face. Fig. 2 shows visual examples of occlusion and explains why they happened.

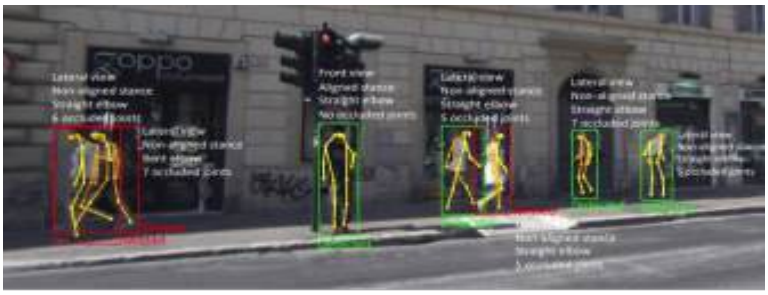


Fig. 2: Occluded People from ECP-DP

4.3 ViTs vs CNNs in Object Occlusion

In a study conducted by Muzammal Naseer et al.[11], the researchers used a method called PatchDrop, where pixel patches of an image are omitted, to observe performance under occlusion. They tested three variations: Random PatchDrop (removing random patches), Salient PatchDrop (omitting the foreground), and Non-salient PatchDrop (omitting the background). Fig. 3 shows a visual representation from the paper. Under 80% Random PatchDrop, a CNN (ResNet50) has 0.1% accuracy, whereas DeiT-S, a lightweight transformer from Meta, retains 40% top-1 accuracy.

This performance gap is due to differences in how each model processes spatial data. Because CNNs rely on local sliding filters, when an occlusion masks a critical region, a CNN can't

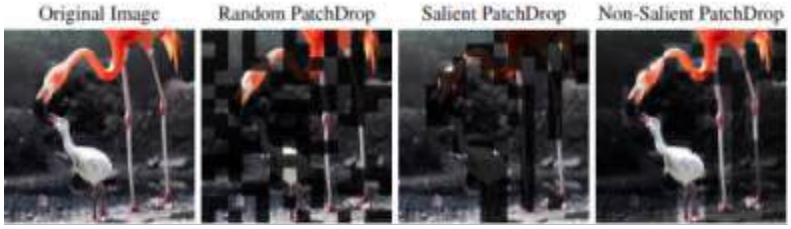


Fig. 3: PatchDrop Examples

[10] "Project Summary: Highway Investigation - 44 Docket Items - HWY18MH010."
[11] Naseer et al., "Intriguing Properties of Vision Transformers."

easily compensate by using visual cues from other regions. Conversely, ViTs process the entire image globally using MHSA in the first layer. This allows the network to have accurate classification even when portions of the object are missing.

4.4 BEVFormer

Vision Transformers are good at occlusion when the target is partially obscured, but this doesn't work if an object is fully obscured. If a pedestrian walks past a car, it's reasonable to assume that the person is still there even if they aren't visible to the driver. However, Standard ViTs and CNNs don't have spatial memory, so if a person walks into a blind spot, the AV will have no idea. BEVFormer, built upon transformer architecture by Li et al.[12], was made to solve this issue by maintaining a continuous spatial memory of the vehicle's surroundings, allowing it to provide a top-down bird's-eye view (Hence the name BEV) of objects near the vehicle regardless of visibility.

BEVFormer is called a spatiotemporal model as it uses spatial data (cameras) and temporal data (past time frames) to map out the environment. BEVFormer starts by creating a blank 2D grid of queries that represents the 3D space around the vehicle in a top-down format. The queries are then updated using two different attention mechanisms.

Spatial Cross-Attention links each query to its location in the camera images. Then, for each query, a set of predefined reference points is sampled at varying heights above its grid cell. These points are then projected into the features of all camera images, and deformable attention (Deformable attention is like MHSA but samples specific locations rather than the whole image) aggregates the most relevant image features into the BEV query. This gives each cell a feature representation based on what the cameras see.

Temporal Self-Attention is the memory portion of the model. Using speed and odometry data from the vehicle, it takes the BEV representation from the previous frame and aligns it with the present frame. Then deformable attention is used to extract features from the old frame, giving the model knowledge of objects that may be occluded.

The final BEV grid is fed into task-specific heads, which transform the representation into an output of bounding boxes of obstacles in the AV's environment.

Section 5: Computational Complexity and Edge Latency

5.1 Hardware Baseline

Since AVs can't use cloud computing to offload their work, they must rely on local in-vehicle processing, also known as edge hardware. To see how edge hardware handles complex AI architectures in autonomous vehicles, we need to determine which hardware is currently used. NVIDIA holds several contracts with major manufacturers such as BYD[13], Mercedes-Benz[14],

[12] Li et al., "BEVFormer."
[13] *BYD Partners With NVIDIA for Mainstream Software-Defined Vehicles*.
[14] *Mercedes-Benz*, "Mercedes-Benz and NVIDIA to Build Software-Defined Computing Architecture for Automated Driving Across Future Fleet."

Toyota[15], and Volvo[16] to use the NVIDIA Orin SoC (system-on-a-chip) as their foundational edge computer.

Table 2 below shows the specs and performance of the NVIDIA DRIVE AGX Orin (The full computer system with the Orin SoC, power management, circuit boards, etc.)[17].

NVIDIA Integrated GPU	Ampere architecture-class — up to 254 INT8 TOPS, including DLA with 87 INT8 TOPS | 5.2 FP32 TOPS
CPU	12x Cortex®-A78A CPU
Programmable Vision Accelerators (PVAs)	Up to 2048 INT8 GMACS | 512 INT16 GMACS
Image Signal Processor (ISP)	1.85 Gigapixels/s
Memory Bandwidth (256-bit LPDDR5)	Up to 200 GB/s
Cameras	90 Gb/s over 16x GMSL(R) ports
Ethernet	Up to 30 Gb/s for data transmission (including lidar/radar)
Vehicle I/O	6 CAN interfaces

The main performance specs to keep in mind are the 254 INT8 TOPS (meaning the computer can make 254 trillion integer operations a second), 200 GB/s memory, and the 90 Gb/s of information from 16 GMSL cameras, with 30 Gb/s from lidar and radar (Note that memory is in GB, gigabytes, and data is in Gb, gigabits). It is also good to know that NVIDIA actually has a better computer system, the DRIVE AGX Thor, with much better specs such as 1000 INT8 TOPS, but manufacturers still haven't adopted it. So we will be using the DRIVE AGX Orin specs as a baseline.

5.2 Workload Demands

The accuracy of Vision Transformers comes at a cost. To determine whether the DRIVE AGX Orin can run a Vision Transformer in real time, we need to calculate the model's TOPS requirement and compare it against Orin's 254 INT8 TOPS. The computational cost of a ViT forward pass can be estimated using FLOPs (Floating-Point Operations) introduced by Kaplan et al.[18], where each transformer layer has $24d^2_{model}$ operations per token, where d_{model} is the hidden dimension or the size of the vector that represents each token. Assuming each camera has

[15] *Toyota, Aurora and Continental Join Growing List of NVIDIA Partners Rolling Out Next-Generation Highly Automated and Autonomous Vehicle Fleets*.
[16] *From Car to Cloud: Volvo Cars Expands Collaboration with NVIDIA*.
[17] *DRIVE AGX Developer Kits*.
[18] Kaplan et al., "Scaling Laws for Neural Language Models."

a resolution of 1600x900 and each patch is 16x16 pixels, there will be around 5,600 patches/tokens. For a ViT-Base architecture[19] with 12 layers and a hidden dimension of 768, the number of operations per frame (OPS per frame) can then be calculated by $5,600 \times 12 \times 24 \times 768^2$, which is 951 GOPS (giga-operations) per frame. Assuming that there are 16 cameras that each run at 30 FPS, the computational demand of a standard ViT in an AV would be 456 TOPS, significantly more than the 254 TOPS ceiling of the AGX Orin. CNNs have a far lighter computational burden. According to Dosovitskiy et al., the GFLOPS for a ViT-Base is 16.8 GFLOPS at a resolution of 224x224. By comparison, He et al.[20] shows that ResNet has only 3.8 GFLOPs at 224x224, roughly 4 times computationally cheaper. Additionally, CNNs use the same filters per convolution, meaning that FLOPS grows linearly with pixel count. ViTs process every patch of pixels as an individual token, meaning that the computational complexity increases exponentially, so scaling from 224x244 to 1600x900 would be significantly more. Therefore, while the AGX Orin would not be able to run a ViT, a CNN would be significantly easier to run.

5.3 Hybrid Models

A goal of computer vision is to figure out how to balance the speed of CNNs with the accuracy of ViTs. One way of doing this is with hybrid models like EfficientViT made by Cai et al.[21] The issue with a standard transformer is that MHSA consumes up to 80% of inference time despite holding only a few parameters. To fix this, EfficientViT puts MHSA between two fast feed-forward network layers (FFNs). The initial FFN shrinks the dimensional space of the input, reducing the cost during attention. The second then expands the output back. The FFNs act like a 1x1 convolution layer in a CNN. The paper shows up to 15x latency reduction on mobile hardware versus the transformer-based architecture SegFormer.

Conclusion

Traditional autonomous vehicle perception systems frequently fail due to static object misclassification, scene misinterpretation, and occlusion. While Vision Transformers (ViTs) theoretically overcome these limitations through global context modeling, their computational overhead exceeds standard edge hardware constraints. CNNs have the efficiency to satisfy the real-time latency constraints on edge hardware such as the NVIDIA AGX Orin, yet their local filters limit global spatial reasoning and fail under occlusion. Therefore, AVs require hybrid architectures like EfficientViT, which achieves up to a 15x latency reduction compared to standard transformers while preserving high accuracy. However, even advanced hybrid models remain limited by immediate sensory occlusion. To bridge this gap, integrating frameworks such as BEVFormer provides crucial spatiotemporal memory across sequential frames. Ultimately, a

[19] Dosovitskiy et al., “An Image Is Worth 16x16 Words.”
[20] He et al., “Deep Residual Learning for Image Recognition.”
[21] Cai et al., “Efficient iT.”

robust autonomous driving vision system requires not a single model but rather multiple architectures working together.

To further bridge the gap between accuracy, computational complexity, and vehicle safety, future research needs to focus on advancing three major fields. One is advancing current architectures, as future research should explore attention mechanisms that scale linearly according to pixel resolution while retaining the global context of standard MHSA. Developing efficient attention algorithms specifically for visual perception can allow transformer architectures to match the speed of classic CNNs. In addition, autonomous vehicles can benefit from advancements in edge hardware. While better chips than the NVIDIA Orin exist, like NVIDIA Thor, manufacturing companies are slow to implement them. Finally, improving BEVFormers to use multi-camera feeds along with LiDAR, radar, and other sensor technology into a single representation space will ensure that perception systems remain fail-safe under extreme driving conditions.

Works Cited

"BYD Partners With NVIDIA for Mainstream Software-Defined Vehicles." BYD, 22 Mar. 2023, www.byd.com/us/news-list/BYD-Partners-With-NVIDIA-for-Mainstream-Software-Defined-Vehicles. Accessed 7 Sept. 2026.

Cai, Han, et al. "EfficientViT: Multi-Scale Linear Attention for High-Resolution Dense Prediction." arXiv, 6 Feb. 2024, https://doi.org/10.48550/arXiv.2205.14756. Preprint.

Dosovitskiy, Alexey, et al. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, 3 June 2021, https://doi.org/10.48550/arXiv.2010.11929. Preprint.

"DRIVE AGX Developer Kits." NVIDIA Developer, NVIDIA Corporation, developer.nvidia.com/drive/agx. Accessed 7 Sept. 2026.

"From Car to Cloud: Volvo Cars Expands Collaboration with NVIDIA." Volvo Cars Media Newsroom, Volvo Car Corporation, 5 Sept. 2024, www.volvocars.com/intl/media/press-releases/9DFE4A251542FAF0/. Accessed 7 Sept. 2026.

He, Kaiming, et al. "Deep Residual Learning for Image Recognition." arXiv, 10 Dec. 2015, https://doi.org/10.48550/arXiv.1512.03385. Preprint.

Jocher, Glenn. "YOLOv8 Documentation." Ultralytics, github.com/ultralytics/ultralytics. Accessed 7 Sept. 2026.

Kaplan, Jared, et al. "Scaling Laws for Neural Language Models." arXiv, 23 Jan. 2020, https://doi.org/10.48550/arXiv.2001.08361. Preprint.

Kaya, Alon, Igal Bilik, and Inna Stainvas. "A Comparative Study of Vision Transformers and CNNs for Few-Shot Rigid Transformation and Fundamental Matrix Estimation." arXiv, 6 Oct. 2025, https://doi.org/10.48550/arXiv.2510.04794. Preprint.

Lambert, Fred. "Tesla 'Robotaxi' Adds 5 More Crashes in Austin in a Month — 4x Worse than Humans." Electrek, 17 Feb. 2026, electrek.co/2026/02/17/tesla-robotaxi-adds-5-more-crashes-austin-month-4x-worse-than-humans/. Accessed 7 Sept. 2026.

Li, Zhiqi, et al. "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers." arXiv, 13 July 2022, https://doi.org/10.48550/arXiv.2203.17270. Preprint.

Liu, Shaoshan, et al. "Edge Computing for Autonomous Driving: Opportunities and Challenges." Proceedings of the IEEE, vol. 107, no. 8, Aug. 2019, pp. 1697–716, https://doi.org/10.1109/JPROC.2019.2915983.

Mercedes-Benz. "Mercedes-Benz and NVIDIA to Build Software-Defined Computing Architecture for Automated Driving Across Future Fleet." Mercedes-Benz USA Media Center, 23 June 2020, media.mbusa.com/releases/release-9e110a76b364c518148b9c1ade1077da-mercedes-benz-and-nvidia-to-build-software-defined-computing-architecture-for-automated-driving-across-future-fleet. Accessed 7 Sept. 2026.

Naseer, Muzammal, et al. "Intriguing Properties of Vision Transformers." arXiv, 25 Nov. 2021, https://doi.org/10.48550/arXiv.2105.10497. Preprint.
"Part 573 Safety Recall Report." National Highway Traffic Safety Administration, U.S. Department of Transportation, 2026, static.nhtsa.gov/odi/rcl/2026/RCLRPT-26E035-7637.pdf. Report no. 26E035.
"Part 573 Safety Recall Report - Zoox." National Highway Traffic Safety Administration, U.S. Department of Transportation, 2026, static.nhtsa.gov/odi/rcl/2026/RCLRPT-26E044-2993.pdf. Report no. 26E044.
"Project Summary: Highway Investigation - 44 Docket Items - HWY18MH010." NTSB Docket Management System, National Transportation Safety Board, data.ntsb.gov/Docket?ProjectID=96894. Accessed 7 Sept. 2026.
"Standing General Order 2021-01." National Highway Traffic Safety Administration, U.S. Department of Transportation, 29 June 2021.
"Toyota, Aurora and Continental Join Growing List of NVIDIA Partners Rolling Out Next-Generation Highly Automated and Autonomous Vehicle Fleets." NVIDIA Newsroom, NVIDIA Corporation, 6 Jan. 2025, nvidianews.nvidia.com/news/toyota-aurora-continental-nvidia-drive. Accessed 7 Sept. 2026.

The Role of Ultra-Processed Foods in Neurodegenerative Diseases through Gut Dysbiosis in Adolescents By Hannah Wang

Abstract

As the prevalence of neurodegenerative diseases (NDs) becomes an increasingly significant burden to the health care system, the necessity of understanding the early-life triggers of NDs becomes more important than ever before. Although they typically manifest later in life, emerging evidence attributes their origins to processes that begin much earlier. A structure significant to the development of NDs is the gut-brain axis, which is supported by microbial activity stemming from the gut microbiome. The gut microbiome is heavily influenced by dietary intake, and ultra-processed foods (UPFs), which are known to be nutritionally lacking, degrade microbiome health. These foods, which are linked to gut dysbiosis and, in turn, neurodegeneration, are consumed most frequently by youth, including adolescents. Therefore, this paper examines whether high consumption of UPFs during adolescence, a key developmental period, influences neurodegeneration in later years. Given the heightened physiological sensitivity during adolescence, it is hypothesized that early-life UPF consumption more readily induces gut dysbiosis that is likely to persist longer, leading to an earlier onset of neurodegenerative disease in adulthood. By synthesizing age-dependent differences in gut-brain responses, this paper highlights adolescence as a potentially important window and demonstrates the importance of preventive strategies targeting excessive UPF consumption early on.

Introduction

Neurodegenerative diseases (NDs) affect 15% of the global population and are one of the primary causes of physical and cognitive disabilities (Giri et al.). These diseases are characterized by progressive loss of neurons in the brain, which manifests as gradual dysfunction of motor skills, cognitive function, behavior, and/or sensory function (Wilson et al.). Although the most common NDs, like Alzheimer's and Parkinson's disease, manifest in the later half of life, evidence suggests that neurodegenerative diseases may result from processes that begin much earlier in life (Landrigan et al.).

Central to the development and progression of neurodegenerative diseases is the gut-brain axis, a bidirectional pathway linking the gut microbiome directly to the brain (Cryan and O'Mahony; Sekirov et al.). This axis is heavily modulated by gut microbiota, and diet can substantially alter the composition of the gut microbiota. Therefore, ultra-processed foods (UPFs), a classification of foods made mostly from food substances and additives and often characterized by poor nutritional quality, are important to consider when examining the relationship between gut health and neurodegeneration (Leo and Campos).

Current research focuses on the mechanisms through which UPFs cause neurodegeneration. For example, Tristan Asensi et al. have discussed the effects of UPFs on systemic inflammation, which contributes to the progression of neurodegenerative diseases (Cunningham; Tristan Asensi et al.). Moreover, reduced gut microbiome diversity increases

blood-brain barrier (BBB) permeability, allowing neurotoxic substances to enter the brain (Braniste et al.). Research has also found that a UPF-centric diet reduces the abundance of microbiota species that produce short-chain fatty acids (SCFAs), which are essential for supporting neuronal health (Majumdar et al.; Rondinella et al.).

Much of this research focuses on diet-induced mechanisms of neurodegeneration in a general context and does not account for how the brains of different age groups respond to UPFs. It remains unclear whether alterations associated with UPFs during adolescence translate into increased risk of neurodegenerative disease decades later; current literature centers on adult brain function and fails to account for the unique plasticity of the developing brain and body.

Thus, this paper addresses the role of high UPF intake during adolescence in inducing dysbiosis and, in turn, influencing the risk of neurodegeneration. Adolescence represents a critical developmental window during which the body may be particularly susceptible to external influences, including dietary intake. Therefore, it is hypothesized that early-life consumption of UPFs induces greater dysbiosis due to the body's vulnerability, potentially resulting in persistent changes to the gut microbiome that could contribute to an increased risk of neurodegeneration later in life. Younger people are the primary consumers of UPFs, making it particularly important to understand how the developing body may respond differently to high UPF intake. Greater knowledge of these effects supports the enforcement of preventive measures, such as dietary guidance aimed at reducing excessive UPF consumption, and better equips us to develop targeted interventions.

This paper will first examine the evidence between UPFs and neurodegeneration, followed by the mechanisms underlying this relationship as established in the more extensively studied adult population. It will then consider how these pathways may operate in adolescents.

Evidence of the association between UPFs and neurodegeneration

A significant amount of work has been done that demonstrates a relationship between neurodegeneration and UPF consumption. This section will focus on the evidence linking neurodegeneration to UPF consumption.

Greater consumption of UPFs in adults is associated with increased risk of incident cognitive impairment (Bhave et al.). In a 2024 study, Bhave and colleagues assessed cognitive impairment by measuring cognitive performance relative to a normative sample on memory and fluency assessments (Bhave et al.). They found that a 10% increase in the relative intake of UPFs was associated with a 16% higher risk of developing cognitive impairment. Furthermore, another article states that those who reported UPF consumption of more than 19.9% of their daily calories had a 28% faster rate of global cognitive decline than those who reported consumption of UPFs less than 19.9% (Gomes Goncalves et al.). The group that consumed a larger amount of UPFs also had a 25% faster rate of executive function decline. Both global and executive function decline were observed after a median follow-up of 8 years. Although cognitive impairment and neurodegeneration are different, with cognitive impairment describing troubles

with memory and neurodegeneration describing brain cells dying off, poor cognition can be a marker of neurodegeneration (Mitchell and Shiri‐Feshki).

In another study, adult participants who consumed a larger amount of UPFs were associated with a higher risk of developing dementia. Among the participants, 287 people developed Alzheimer's disease and 119 developed vascular dementia (Li et al.). Moreover, the same study found that replacing 10% of UPF weight in diet with an equivalent proportion of unprocessed or minimally processed foods was estimated to be associated with a 19% lower risk of dementia in adults. There has also been an association between the consumption of artificially sweetened beverages and an increased risk of all-cause dementia and Alzheimer's disease in adults with a mean age of 69 (Pase et al.). Artificially sweetened drinks are generally classified as UPFs under the NOVA food classification system (Monteiro et al.).

Overall, numerous studies have demonstrated that in adulthood, consuming UPFs, as well as ingredients that make up a large presence of UPFs, is associated with neurodegenerative diseases, cognitive impairment, and faster rates of cognitive decline. The following sections will examine the specific physiologic pathways that underlie these associations.

The induction of gut dysbiosis through UPF consumption

This section highlights current knowledge of the mechanisms linking UPF consumption to changes in gut health, based mostly on evidence from adult populations, and serves as a reference point for understanding how the body reacts to UPF consumption.

UPFs are engineered with a specific combination of fat, sodium, and carbohydrates at levels that are not found in nature (Monteiro et al.). According to the NOVA food classification system proposed by Monteiro and colleagues, UPFs are made mostly from food additives, lacking essential nutrients, such as fiber and protein, and are designed for convenience or palatability (Monteiro et al.). UPFs are a significant component of the "western-style diet," characterized by the consumption of highly refined foods with poor nutritional quality (McBurney et al.).

Although certain bacteria are essential to holistic health, the gut microbiome does not require a specific formula of taxa to be considered "ideal" (McBurney et al.). In fact, the adult gut microbiome houses a highly individualized microbial community, and characteristics such as microbial diversity play larger roles in shaping gut health (McBurney et al.). High microbial diversity in the gut is often associated with a healthy gut microbiome, whereas low diversity has been linked to greater susceptibility to certain diseases. The resilience of the microbiota, defined as its ability to return to equilibrium following a chemical or physical perturbation, is another important characteristic of a healthy gut microbiome and appears to be associated with greater microbial diversity (McBurney et al.).

Gut dysbiosis is described as an imbalance or disruption in the composition and function of the gut microbial community (Hrncir). Its most distinct features include decreased gut microbial diversity, fewer beneficial bacteria, and increased harmful microbiota. Harmful microbiota can trigger inflammation and produce toxins (Hrncir).

UPFs can promote gut dysbiosis and contribute to systemic inflammation through several mechanisms. The two main pathways include increasing gut permeability and introducing pro-inflammatory markers into the body.

The nutritional profile of UPFs often lacks fiber, which is an important food source for many gut bacteria (McBurney et al.; Rondinella et al.). Dietary fiber supports the growth and diversity of the gut microbiota, whereas low-fiber diets may reduce microbial diversity and are associated with the degradation of the mucus barrier of the gut epithelial lining (McBurney et al.; Rondinella et al.). This layer overlies the gut epithelium and is a primary defense against enteric pathogens (Desai et al.). In the absence of sufficient dietary fiber, certain gut bacteria use the glycoprotein-rich mucus layer as an energy source (Desai et al.). Continued consumption of this mucus layer over time thins and weakens it, allowing pathogens greater access to the gut epithelium and contributing to intestinal inflammation (Desai et al.). Increased gut permeability allows bacterial metabolites and endotoxins into the bloodstream, contributing to systemic inflammation (Di Vincenzo et al.).

The microbial fermentation of dietary fiber also produces short-chain fatty acids (SCFAs), such as acetate, propionate, and butyrate, which play key roles in maintaining gut health (Rondinella et al.). SCFAs support the integrity of the gut mucosal barrier (Tan et al.). They also regulate inflammation by reducing pro-inflammatory signals, like tumor necrosis factor alpha (TNF-α), while increasing anti-inflammatory signals (Tan et al.). Higher UPF intake often reduces fiber intake, leading to lower levels of SCFAs in the gut, potentially compromising mucosal barrier integrity and inflammation regulation (Rondinella et al.; Tan et al.). Higher UPF intake also displaces nutrient-dense foods, such as fruits, that are associated with anti-inflammatory properties (Tristan Asensi et al.).

Additives commonly found in UPFs are linked with systemic inflammation. A 2017 mouse study conducted by Bian and colleagues found that six months of sucralose consumption, an artificial sweetener widely used in UPFs, increased the abundance of the bacterial genes involved in the production of pro-inflammatory mediators, such as lipopolysaccharide (LPS) in the gut, which can initiate the creation of pro-inflammatory cytokines such as interleukin-6 (IL-6) and tumor necrosis factor (TNF)-α (Bian et al.). LPS can also translocate into circulation and tissues, contributing to systemic inflammation (Di Vincenzo et al.). Sucralose-treated mice showed elevated levels of inflammatory mediators in the liver, such as MMP-2, an enzyme that can activate pro-inflammatory cytokines by cleaving them. UPFs also often use emulsifiers, such as carboxymethylcellulose (CMC). In a double-blind, controlled-feeding trial, healthy adults consuming 15 g/day of CMC for 11 days showed reduced microbiota diversity and lower fecal SCFAs (Chassaing et al.).

In addition to contributing to systemic inflammation, UPFs alter microbial composition and diversity. Increased consumption of UPFs reduces advantageous species that support the gut microbiome (Rondinella et al.). Studies have found that a “western-style diet” decreases gut diversity and changes gut microbiome composition (McBurney et al.). A study using a humanized mouse model fed a low-fat, polysaccharide-rich diet, a plant-based diet that

emphasizes the consumption of dietary fibers and complex carbohydrates, showed a significant increase in *Erysipelotrichia* in the gut microbiome after being put onto the western-style diet (Turnbaugh et al.). Although studies have yielded inconsistent results, *Erysipelotrichacea*e has been associated with pro-inflammatory signals when present in excess (Kaakoush). An additional study compared the gut microbiome of adults who consumed fewer than three servings of UPFs daily with that of adults who consumed more than five servings of UPFs daily. It showed that individuals who consumed a high amount of UPFs had a larger amount of potentially harmful bacterial groups compared to those who consumed a low amount of UPFs (Cuevas-Sierra et al.). Additionally, this heavy consumption reduced the amount of beneficial bacteria, like *Lachnospira* and *Roseburia*, which are known to produce SCFAs (Cuevas-Sierra et al.).

Overall, UPFs impact the gut through two main mechanisms. The first is that UPFs use the gut to contribute to systemic inflammation by increasing the permeability of the gut and introducing pro-inflammatory markers into the body. The second is that UPFs alter the gut's microbial composition by promoting harmful bacteria and decreasing beneficial bacteria. The following section discusses how this inflammation and gut dysbiosis relate to neurodegeneration.

General mechanisms of neurodegeneration as a product of the gut-brain axis

The primary connection between the gut and the brain is through the gut-brain axis (Carabotti et al.). This is a bidirectional pathway, meaning the brain affects gut function and the gut influences the brain. The gut-brain axis works through the vagus nerve, which sends signals from the gut to the brain, and vice versa. Gut microbiota can influence the brain’s health and mood through the gut-brain axis by producing metabolites such as SCFAs and modulating neurotransmitter signaling (Carabotti et al.).

Lower gut diversity that contains more harmful than beneficial microbiota leads to a reduced production of SCFAs. SCFAs are crucial for regulating inflammation and maintaining the integrity of the blood-brain barrier (BBB) (Fock and Parnova). Gut dysbiosis can also increase gut permeability, allowing bacteria, LPS, and other toxic products to enter the circulation and initiate systemic inflammation, which has been linked to neurodegenerative diseases (Fock and Parnova).

The BBB is a filter that acts as a barrier between the brain and the circulatory system, regulating the transfer of substances between the two systems (Daneman and Prat). Increased permeability in the BBB can lead to neuroinflammation in the brain by facilitating the infiltration of leukocytes, the passage of bacteria and their toxic products, and the entry of pro-inflammatory mediators (Fock and Parnova). Additionally, evidence links Alzheimer's to BBB permeability in both animal models and clinical studies in Alzheimer patients (Zenaro et al.).

Prior research suggests that the gut microbiome plays a role in regulating the BBB, and that a lack of gut microbiota increases BBB permeability. This is supported by Braniste and colleagues, who found that germ-free mice–which have no microbiomes–exhibited higher BBB permeability than pathogen-free mice–which have healthy microbiomes–beginning in intrauterine life (Braniste et al.). This difference persisted after birth and into adulthood. The

study also showed that exposing germ-free adult mice to a pathogen-free microbiome by administering SCFAs decreased BBB permeability, suggesting a link between BBB integrity and SCFAs, with SCFAs having a protective effect on the BBB (Fock and Parnova; Braniste et al.).

UPFs cause systemic inflammation through the gut, increasing its permeability and decreasing the abundance of end products that regulate inflammation. This systemic inflammation exacerbates current neuroinflammation in the brain (Cunningham). In healthy brains, microglia lie "quiescent." However, in a brain experiencing neurodegenerative diseases like Alzheimer's, microglia transition to a primed state. These primed microglia are hypersensitive to inflammatory stimuli and release more pro-inflammatory cytokines such as IL-1β and TNF-α, initiating neutrophil infiltration directly into the brain, thereby contributing to neuroinflammation (Cunningham).

Neuroinflammation is linked to neurodegeneration. Prolonged inflammatory reactions perform neuronally detrimental functions in the brain by mediating the deleterious onset and progression of neurodegenerative diseases (Kempuraj et al.). A study in rats showed that systemic LPS triggered TNF-α production in the liver, which then crossed the BBB to activate microglia (Qin et al.). Once the microglia were activated by the LPS, they entered a "self-propelling" cycle of inflammation. Following LPS exposure, neuronal death in the substantia nigra, a characteristic of Parkinson's disease, occurred, though neuronal death did not happen until 7 to 10 months following LPS exposure. This may suggest why neurodegenerative diseases most often manifest much later in life despite early life exposure to factors that are linked to neurodegeneration.

In summary, gut dysbiosis induces systemic inflammation and decreases the production of SCFAs that are linked to maintaining the integrity of the BBB and regulating inflammation. This leads to neuroinflammation through inflammatory markers breaching the BBB and priming the microglia in the brain. Neuroinflammation is linked to neurodegeneration by performing neuronally detrimental functions.

Neurodegeneration in the context of adolescents

Having established the general mechanisms underlying the relationship between UPF consumption and neurodegeneration, this paper will now examine evidence specific to adolescents.

The adolescent brain is still in development, and the gut microbiome is not stable, unlike its adult-like form. These concurrent developmental processes mean that disruptions during this window may yield different results compared to the same disruption in adulthood, with a fully developed system. The adolescent gut microbiome continues to develop into its adult-like profile during adolescence (Carson et al.). Repeated shifts in the gut microbiota during critical windows of development may impair its ability to recover into a stable state. External factors that alter the composition of the gut microbiome may lead to persistent gut dysbiosis, potentially resulting in lasting effects on the host. In a mouse study, Carson et al. found that antibiotic treatment during

adolescence altered the gut microbiome and impaired skeletal maturation even after treatment ended (Carson et al.)

Several studies have pointed to how UPF affects the adolescent brain differently. A study found that there was a 23% reduction in doublecortin (DCX)-positive cells in mice when a high-fat diet started at weaning and continued into adolescence, whereas mice exposed to the same diet during adulthood did not show a significant reduction (Boitard et al.). Though high-fat diets are not equivalent to UPFs, many UPFs are characterized by high levels of added fats. DCX-positive cells are neuronal precursor cells that are important for neurogenesis, allowing new neurons to migrate to their correct locations in the brain and form functional connections. This finding suggests that exposure to a high-fat diet, and thus perhaps UPFs, during a developmental window may stunt neurogenesis. A lack of DCX-positive cells could suggest that the brain is more susceptible to age-related decay due to the brain failing to build a structurally robust network initially.

Additionally, adolescent rats given 30 days of access to a 11% high fructose corn syrup solution showed significant elevations in pro-inflammatory cytokines IL-6 and IL-1β in the hippocampus, which is often observed before the onset of hippocampal dysfunction (Hsu et al.). However, adult rats show no difference in these pro-inflammatory cytokines following the same exposure. Such a difference in the adult rats' and the adolescent rats' response points to the increased vulnerability of the adolescent brain, being more susceptible to producing pro-inflammatory signals (Hsu et al.).

A human study showed a similar increase in pro-inflammatory cytokines. Adolescents in the highest tertile of UPF consumption had a 79% increase in pro-inflammatory cytokine IL-8 compared to adolescents who consumed the least amount of UPFs, suggesting that greater UPF consumption is associated with increased inflammation during adolescence (dos Santos Martins et al.). This inflammation plays a vital part in triggering neurodegeneration by promoting neuroinflammation in the brain. Several human studies have reported associations between higher UPF consumption and increased IL-6 levels in adults, whereas IL-6 generally did not vary with UPF intake in children and adolescents (Ciaffi et al.). IL-8 in adults remains little studied, and comparable increases associated with UPF consumption have not yet been clearly demonstrated. These findings suggest that the inflammatory responses associated with UPF consumption may differ across age groups; however, direct comparisons between adolescents and adults are currently lacking. Further investigation is therefore needed.

A number of studies have examined associations between UPF consumption, cognitive performance, and academic performance among children and adolescents. Nguyen et al.'s systematic review and meta-analysis of 35 studies involving 84,062 children and adolescents found that higher UPF consumption was associated with poorer overall cognitive performance, particularly in attention, executive functioning, fluid intelligence, language, and visuospatial ability (Nguyen et al.). In the Eating Healthy and Daily Life Activities (EHDLA) study, López-Gil et al. found, among 788 youths aged 12-17, significant dose–response associations between UPF consumption and all the different academic performance indicators, showing that

higher UPF consumption is consistently associated with poorer academic performance (López-Gil et al.).

Adolescents have an increased reward sensitivity and reduced behavioral control relative to the adult brain, in part because dopamine is especially abundant during this period while the prefrontal cortex does not reach functional maturity until the early twenties (Reichelt). Rodent studies have shown that although basal dopamine release may be lower during adolescence, stimulation by environmental or pharmacological challenges can produce greater dopamine release in adolescents than in adult brains (Reichelt). Thus, consuming palatable high-fat, high-sugar foods, including many UPFs, may produce a greater dopamine release in adolescents than in adults. This means that adolescents are more likely to seek out UPFs and consume more of them on a regular basis. In a 2025 report based on National Health and Nutrition Examination Survey (NHANES) data collected from August 2021 to August 2023, US adolescents aged 12 to 18 obtained 63% of total calories from UPFs, compared with 53% among adults 19 and older (Williams et al.). Similarly, in a large European Cohort, adolescents aged 10 to 19 obtained 47% of total calories from UPFs, compared with 40% among adults 20 and older (Formisano et al.).

A few longitudinal studies have shown the associations between UPF consumption during childhood and adolescence and later adverse health outcomes. In a nationally representative German cohort, Figueiredo Barata et al. followed 4,762 participants aged 3 to 17 for an average of 11 years. They found a dose-response association between UPF intake and incident overweight; each 10% increase in the energy from UPFs was associated with 12% higher odds of becoming overweight, and participants in the highest UPF quartile had 49% greater odds of becoming overweight than those in the lowest quartile (Figueiredo Barata et al.). Li and Shi analyzed longitudinal data from 3,437 Chinese children and adolescents aged 6 to 18 in the China Health and Nutrition Survey between 1997 and 2011. Participants consuming 50 to 99 g of UPFs per day had approximately twice the odds of BMI-defined overweight or obesity compared with non-consumers (Li and Shi).

Overall, both the brain and the gut continue to develop through adolescence. Evidence from animal studies has suggested disrupting the gut microbiome during adolescence may have lasting consequences. Animal studies have also shown that adolescents exposed to diets high in fat or added sugar are more prone to experience reduced neurogenesis and an elevation of pro-inflammatory cytokines in the hippocampus compared to adults exposed to the same diets, which may make them more susceptible to neuroinflammation and neurodegenerative processes. Human studies have shown that higher UPF consumption in adolescence has been associated with increased pro-inflammatory cytokines and suggest that the inflammatory response may vary across different age groups. However, whether these developmental effects translate to an increased risk of neurodegenerative disease later in life remains unknown. Additionally, studies have reported that adolescents who derive more of their calories from UPFs have lower cognitive and academic performance. It is also important to note that adolescents are more at risk because they are more sensitive to dopamine, meaning they are more likely to consume large amounts of UPFs to trigger the dopamine release that comes with consuming hyperpalatable foods. Reports

from different countries have confirmed that adolescents obtain a higher percentage of daily calories from UPFs than adults. Longitudinal studies have shown that UPF consumption in childhood and adolescence is associated with adverse health outcomes later in life. These studies mainly focused on adiposity outcomes. No longitudinal study has been conducted to evaluate whether UPF consumption during adolescence is associated with neurodegenerative disease later in life.

Adult evidence linking UPF consumption to neurodegeneration makes up the bulk of the research, while the adolescent evidence is limited to cognitive and academic performance in cross-sectional studies. Adolescents, especially those in early adolescence who are experiencing the onset of puberty, are in a critical window of development. UPF consumption may contribute to conditions that precede neurodegeneration later in life, in a similar way to what's seen in adults, where UPF consumption is linked to factors such as gut dysbiosis and inflammation that contribute to neurodegeneration. Additionally, adolescents are among the highest consumers of UPFs. Thus, it is critical for us to give greater research attention to the adolescent population. Future studies should include comparative human research examining how inflammatory responses to UPF consumption differ between adolescents and adults. Longitudinal studies that follow adolescent populations into later adulthood are especially important to understand the long-term effects of early UPF consumption. The study could include an initial screening to test for inflammatory markers or gut dysbiosis during adolescence, followed by repeated assessment over time to determine the longevity of the inflammatory markers, such as how long the increased cytokines prevail and whether they can still be observed in adulthood, and whether they are associated with later neurodegenerative outcomes.

With a potential causal link between UPFs and neurodegeneration comes the need for public health and dietary policy interventions. The focus should be on educating the public on the risk of UPFs in the context of neurodegeneration, as well as in the context of nutrition in general. Regulatory measures, such as requiring clear warning labels on food packaging, may also help consumers make more informed dietary choices. Not only should we be more mindful of the UPFs we consume to reduce the risks of adverse brain health, but we should also do so to optimize our body's general well-being in the future.

Conclusion

Animal studies have suggested that the adolescent gut microbiota may be more susceptible to external factors compared to the adult gut microbiota. The association between adult UPF consumption and neurodegeneration, as well as its underlying mechanisms, has been supported by numerous studies. Some factors that appear in the pathways of neurodegeneration, as researched in adults, also appear in excess following UPF consumption in adolescents, which may suggest a potential link between UPF consumption and neurodegeneration. For example, inflammatory markers sometimes increase following adolescent UPF consumption, and inflammation is linked to inciting and accelerating neurodegeneration in adults. However, diet-induced neurodegeneration manifests later in life, which means that longitudinal studies are

needed to measure the long-term effects of early UPF consumption. The main knowledge gap is centered around how UPF-induced inflammation markers and gut dysbiosis in adolescents influence neurodegeneration and how much of a long-lasting impact they have on the body following adolescence. In understanding more about this topic, preventative measures for future neurodegeneration can be put into place and implemented during one's youth.

Acknowledgements

I would like to acknowledge my mentors, Dr. Jorge Avila and Bre Calhoun, MS, for guiding me through this process, as well as my parents who have been nothing but supportive.

Works Cited

Bhave, Varun M., et al. “Associations between Ultra-Processed Food Consumption and Adverse Brain Health Outcomes.” Neurology, vol. 102, no. 11, 2024, p. e209432.

Bian, Xiaoming, et al. “Gut Microbiome Response to Sucralose and Its Potential Role in Inducing Liver Inflammation in Mice.” Frontiers in Physiology, vol. 8, 2017, p. 272733.

Boitard, Chloe, et al. “Juvenile, but Not Adult Exposure to High‐fat Diet Impairs Relational Memory and Hippocampal Neurogenesis in Mice.” Hippocampus, vol. 22, no. 11, 2012, pp. 2095–100.

Braniste, Viorica, et al. “The Gut Microbiota Influences Blood-Brain Barrier Permeability in Mice.” Science Translational Medicine, vol. 6, no. 263, 2014, pp. 263ra158-263ra158, https://doi.org/10.1126/scitranslmed.3009759.

Carabotti, Marilia, et al. “The Gut-Brain Axis: Interactions between Enteric Microbiota, Central and Enteric Nervous Systems.” Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology, vol. 28, no. 2, 2015, p. 203.

Carson, Matthew D., et al. “Adolescence and the Microbiome: Implications for Healthy Growth and Maturation.” The American Journal of Pathology, vol. 193, no. 12, 2023, pp. 1900–09.

Chassaing, Benoit, et al. “Randomized Controlled-Feeding Study of Dietary Emulsifier Carboxymethylcellulose Reveals Detrimental Impacts on the Gut Microbiota and Metabolome.” Gastroenterology, vol. 162, no. 3, 2022, pp. 743–56.

Ciaffi, Jacopo, et al. “Ultra-Processed Food Consumption and Systemic Inflammatory Biomarkers: A Scoping Review.” Nutrients, vol. 17, no. 18, 2025, p. 3012.

Cryan, John F., and Siobhain M. O’Mahony. “The Microbiome‐gut‐brain Axis: From Bowel to Behavior.” Neurogastroenterology & Motility, vol. 23, no. 3, 2011, pp. 187–92, https://doi.org/10.1111/j.1365-2982.2010.01664.x.

Cuevas-Sierra, Amanda, et al. “Gut Microbiota Differences According to Ultra-Processed Food Consumption in a Spanish Population.” Nutrients, vol. 13, no. 8, 2021, p. 2710.

Cunningham, Colm. “Microglia and Neurodegeneration: The Role of Systemic Inflammation.” Glia, vol. 61, no. 1, 2013, pp. 71–90, https://doi.org/10.1002/glia.22350.

Daneman, Richard, and Alexandre Prat. “The Blood–Brain Barrier.” Cold Spring Harbor Perspectives in Biology, vol. 7, no. 1, 2015, p. a020412.

Desai, Mahesh S., et al. “A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility.” Cell, vol. 167, no. 5, 2016, pp. 1339–53.

Di Vincenzo, Federica, et al. “Gut Microbiota, Intestinal Permeability, and Systemic Inflammation: A Narrative Review.” Internal and Emergency Medicine, vol. 19, no. 2, 2024, pp. 275–93.

dos Santos Martins, Glauciane Márcia, et al. “Intake of Ultra-Processed Foods Is Associated with Inflammatory Markers in Brazilian Adolescents.” Public Health Nutrition, vol. 25, no. 3, 2022, pp. 591–99.

Figueiredo Barata, Mayra, et al. “Association of Ultra-Processed Food Patterns with Overweight and Obesity in the German Health Interview and Examination Survey for Children and Adolescents (KiGGS): A Longitudinal Study.” BMC Public Health, vol. 25, no. 1, 2025, p. 3961.

Fock, Ekaterina, and Rimma Parnova. “Mechanisms of Blood–Brain Barrier Protection by Microbiota-Derived Short-Chain Fatty Acids.” Cells, vol. 12, no. 4, 2023, p. 657.

Formisano, Annarita, et al. “Ultra-Processed Foods Consumption and Metabolic Syndrome in European Children, Adolescents, and Adults: Results from the I. Family Study.” Nutrients, vol. 17, no. 13, 2025, p. 2252.

Giri, Paras Mani, et al. “Neuroinflammation in Neurodegenerative Disorders: Current Knowledge and Therapeutic Implications.” International Journal of Molecular Sciences, vol. 25, no. 7, 2024, p. 3995.

Gomes Goncalves, Natalia, et al. “Association between Consumption of Ultraprocessed Foods and Cognitive Decline.” JAMA Neurology, vol. 80, no. 2, 2023, pp. 142–50.

Hrncir, Tomas. “Gut Microbiota Dysbiosis: Triggers, Consequences, Diagnostic and Therapeutic Options.” Microorganisms, vol. 10, no. 3, 2022, p. 578.

Hsu, Ted M., et al. “Effects of Sucrose and High Fructose Corn Syrup Consumption on Spatial Memory Function and Hippocampal Neuroinflammation in Adolescent Rats.” Hippocampus, vol. 25, no. 2, 2015, pp. 227–39.

Kaakoush, Nadeem O. “Insights into the Role of Erysipelotrichaceae in the Human Host.” Frontiers in Cellular and Infection Microbiology, vol. 5, 2015, p. 84.

Kempuraj, Duraisamy, et al. “Neuroinflammation Induces Neurodegeneration.” Journal of Neurology, Neurosurgery and Spine, vol. 1, no. 1, 2016, p. 1003.

Landrigan, Philip J., et al. “Early Environmental Origins of Neurodegenerative Disease in Later Life.” Environmental Health Perspectives, vol. 113, no. 9, 2005, p. 1230.

Leo, Edwin E. Martínez, and Maira R. Segura Campos. “Effect of Ultra-Processed Diet on Gut Microbiota and Thus Its Role in Neurodegenerative Diseases.” Nutrition, vol. 71, 2020, p. 110609.

Li, Huiping, et al. “Association of Ultraprocessed Food Consumption with Risk of Dementia: A Prospective Cohort Study.” Neurology, vol. 99, no. 10, 2022, pp. e1056–66.

Li, Ming, and Zumin Shi. “Ultra‐processed Food Consumption and Obesity among Children and Adolescents in China—Findings from China Health and Nutrition Survey.” Pediatric Obesity, vol. 20, no. 7, 2025, p. e70012.

López-Gil, José Francisco, et al. “Investigating the Relationship between Ultra-Processed Food Consumption and Academic Performance in the Adolescent Population: The EHDLA Study.” Nutrients, vol. 17, no. 3, 2025, p. 524.

Majumdar, Atreye, et al. “Short-Chain Fatty Acids in the Microbiota–Gut–Brain Axis: Role in Neurodegenerative Disorders and Viral Infections.” ACS Chemical Neuroscience, vol. 14, no. 6, 2023, pp. 1045–62.

McBurney, Michael I., et al. “Establishing What Constitutes a Healthy Human Gut Microbiome: State of the Science, Regulatory Considerations, and Future Directions.” The Journal of Nutrition, vol. 149, no. 11, 2019, pp. 1882–95.

Mitchell, Alex J., and Mojtaba Shiri‐Feshki. “Rate of Progression of Mild Cognitive Impairment to Dementia–Meta‐analysis of 41 Robust Inception Cohort Studies.” Acta Psychiatrica Scandinavica, vol. 119, no. 4, 2009, pp. 252–65.

Monteiro, Carlos Augusto, et al. “A New Classification of Foods Based on the Extent and Purpose of Their Processing.” Cadernos de Saude Publica, vol. 26, 2010, pp. 2039–49.

Nguyen, Lan, et al. “Food for Thought: A Systematic Review and Meta‐analysis on the Effects of Ultra‐processed Foods on Cognition in Children and Adolescents.” Food Frontiers, vol. 6, no. 4, 2025, pp. 1838–66.

Pase, Matthew P., et al. “Sugar-and Artificially Sweetened Beverages and the Risks of Incident Stroke and Dementia: A Prospective Cohort Study.” Stroke, vol. 48, no. 5, 2017, pp. 1139–46.

Qin, Liya, et al. “Systemic LPS Causes Chronic Neuroinflammation and Progressive Neurodegeneration.” Glia, vol. 55, no. 5, 2007, pp. 453–62.

Reichelt, Amy C. “Adolescent Maturational Transitions in the Prefrontal Cortex and Dopamine Signaling as a Risk Factor for the Development of Obesity and High Fat/High Sugar Diet Induced Cognitive Deficits.” Frontiers in Behavioral Neuroscience, vol. 10, 2016, p. 189.

Rondinella, Debora, et al. “The Detrimental Impact of Ultra-Processed Foods on the Human Gut Microbiome and Gut Barrier.” Nutrients, vol. 17, no. 5, 2025, p. 859.

Sekirov, Inna, et al. “Gut Microbiota in Health and Disease.” Physiological Reviews, 2010.

Tan, Jian, et al. “The Role of Short-Chain Fatty Acids in Health and Disease.” Advances in Immunology, vol. 121, 2014, pp. 91–119.

Tristan Asensi, Marta, et al. “Low-Grade Inflammation and Ultra-Processed Foods Consumption: A Review.” Nutrients, vol. 15, no. 6, 2023, p. 1546, https://doi.org/10.3390/nu15061546.

Turnbaugh, Peter J., et al. “The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice.” Science Translational Medicine, vol. 1, no. 6, 2009, pp. 6ra14-6ra14.

Williams, Anne M., et al. “Ultra-Processed Food Consumption in Youth and Adults: United States, August 2021–August 2023.” NCHS Data Briefs [Internet], National Center for Health Statistics (US), 2025.

Wilson, David M., et al. “Hallmarks of Neurodegenerative Diseases.” Cell, vol. 186, no. 4, 2023, pp. 693–714.

Zenaro, Elena, et al. “The Blood-Brain Barrier in Alzheimer’s Disease.” Neurobiology of Disease, vol. 107, 2017, pp. 41–56.

Comparing Heatmap Regression, Segmentation, and Multi-Frame Temporal Context for Surgical Tool-Tip Localization Under a Unified Evaluation Protocol By Tara Tony[1]

Abstract

Accurate surgical tool-tip localization is important for robotic assistance, surgical skill assessment, workflow analysis, and other computer-assisted surgery applications, but localization remains difficult because surgical videos frequently contain motion blur, occlusion, illumination changes, and closely spaced instrument landmarks. Prior work has primarily approached this problem using single-frame heatmap-based keypoint estimation or segmentation-based representations, while more recent methods have introduced temporal information from neighboring video frames. However, these approaches have generally been evaluated using different datasets, landmark definitions, and metrics, and prior studies have not directly compared heatmap regression, segmentation-based localization, and temporal multi-frame localization under a common experimental framework.

To address this gap, this study evaluates three representative computer vision approaches for surgical tool-tip localization using the same annotated JIGSAWS data and evaluation protocol: ViTPose for heatmap-based keypoint estimation, DeepLabV3 for segmentation-based localization, and MFCNet for temporal multi-frame localization. All methods predict the same surgical landmark locations, and performance is evaluated using mean pixel error and Percentage of Correct Keypoints at 5 and 10 pixels (PCK@5 and PCK@10).

The results show a consistent performance ordering across all three metrics. ViTPose achieved a mean pixel error of 11.99 pixels, PCK@5 of 45.28%, and PCK@10 of 76.04%. DeepLabV3 substantially improved performance to 4.67 pixels, 71.02%, and 90.38%, respectively, while MFCNet achieved the best overall performance with a mean pixel error of 3.97 pixels, PCK@5 of 74.97%, and PCK@10 of 94.91%. These findings suggest that representing landmarks as spatial regions may be more effective than direct heatmap-based keypoint estimation for closely spaced surgical landmarks, while incorporating temporal context provides an additional improvement beyond single-frame segmentation. Future work should evaluate these findings on more diverse surgical datasets and assess robustness under these challenging visual conditions.

1. Introduction

Minimally invasive surgery relies heavily on endoscopic video to guide surgical instruments within the body [1]. Unlike open surgery, where surgeons can view the operative field directly, minimally invasive procedures require them to interpret the surgical scene through a camera feed [10]. As robotic-assisted surgery becomes more common, computer vision has an increasingly important role in extracting useful information from these videos. In particular, surgical tool-tip localization underpins instrument tracking and supports downstream applications such as robotic assistance, surgical skill assessment, workflow analysis, navigation,

[1] Notre Dame High School

and autonomous robotic systems [4,10]. However, accurate localization remains challenging because surgical videos frequently contain motion blur, partial occlusion, lighting changes, reflections, and camera movement [4]. These difficulties are especially important when localizing small, closely spaced landmarks such as tool tips and jaw bases, which are specifically labeled points on an instrument [4].

Early approaches to surgical instrument localization relied on hand-designed visual features such as color and edges, where a person decided in advance which visual property identified the tool and passed that measurement to a classifier. These methods were limited by the variability of surgical video, as color-based features can become unreliable under changing illumination, while edge-based features may degrade when instrument motion produces blur [5]. More broadly, limited availability of standardized surgical datasets and consistent evaluation protocols also hindered progress, making direct comparisons between methods difficult [1]. Deep learning reduced the need for manually engineered features and led to two common approaches for representing surgical landmarks. Heatmap-based keypoint estimation predicts a probability or confidence map for each landmark and extracts the landmark location from the peak response [11], while segmentation-based methods assign pixel-level labels to landmark regions and recover landmark coordinates from those segmented regions [2].

Prior studies have explored both representations. Laina et al. combined segmentation and landmark heatmap prediction within a shared network and showed that incorporating segmentation information could improve landmark localization [5]. However, they used segmentation to support heatmap prediction rather than evaluating it independently as a method for recovering landmark coordinates; every model they tested predicted landmarks directly. Du et al. instead used heatmap-based prediction to estimate surgical instrument joints and their spatial relationships, followed by a graph-matching procedure to construct complete instrument poses [3]. Both approaches demonstrated the feasibility of deep-learning-based landmark localization, but they operated on individual frames and therefore did not exploit the temporal continuity naturally present in surgical video [3,5].

Several gaps remain in this prior work. First, single-frame methods discard information from adjacent frames even though surgical instrument motion is continuous over time. Both Laina et al. and Du et al. state that their methods use no temporal information, and Du et al. report failures where joints hidden behind a second instrument could not be recovered [3,5]. This limitation may matter most during motion blur or occlusion, when a landmark difficult to identify in one frame may still be recoverable from neighboring frames [4]. Because these conditions are among the defining characteristics of surgical video, a method that discards temporal information is weakest in exactly the conditions it will encounter most often, and it remains untested how much of that weakness multi-frame context can recover [4]. Second, many keypoint-estimation architectures were originally developed and benchmarked on natural-image or human-pose datasets [2,6,9,11], where strong performance does not necessarily translate to surgical scenes with small, closely spaced instrument landmarks. Pose estimation models are built to localize a small number of labeled points, so a pose estimator is the natural first choice

for surgical tool-tip localization, but whether that choice holds up against other representations has not been tested on surgical data. Third, previous studies have not directly compared heatmap-based keypoint estimation, segmentation-based localization, and temporal multi-frame localization using the same data, landmark definitions, and evaluation criteria. Differences among prior datasets and metrics therefore make it difficult to determine which representation is most effective for surgical tool-tip localization [1]. This choice matters because the jaw base and tool tips sit only a few pixels apart in a typical frame, so small localization errors can determine whether individual landmarks are distinguishable.

These limitations motivate a controlled comparison that evaluates different localization strategies using a unified protocol under shared train and test data, landmark targets, coordinate scale, and result metrics. Such a comparison makes it possible to investigate not only which model performs best, but also whether different forms of visual information are associated with improved localization. In particular, comparing direct keypoint estimation with segmentation-based localization can clarify the value of explicitly modeling spatial landmark regions, while comparing single-frame segmentation with a temporal multi-frame approach can evaluate the added benefit of incorporating information across consecutive frames [4].
Rather than asking only which model performs best, this study therefore investigates a broader research question: what type of visual information is most useful for surgical tool-tip localization? To investigate this question, three representative approaches are compared on the same annotated frames from the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS), a publicly available surgical activity dataset containing synchronized endoscopic video and robot kinematic recordings of surgeons performing suturing, knot-tying, and needle-passing tasks on the da Vinci surgical system [12]. ViTPose represents heatmap-based keypoint estimation and directly predicts confidence maps for six surgical landmarks from individual frames using a transformer-based architecture [8,11]. DeepLabV3 [2] represents segmentation-based localization, predicting a labeled region for each landmark and recovering its coordinate as the centroid of that region. MFCNet extends the segmentation-based approach by incorporating information from consecutive frames, including aligned segmentation predictions and additional spatial information, to perform temporal localization [4,7].

This work's primary contribution is a unified experimental comparison of heatmap-based [11], segmentation-based [2], and temporal multi-frame [4] approaches for surgical tool-tip localization under a common dataset and evaluation protocol. Rather than comparing architectures trained on different datasets or measured with different metrics, we evaluate all three approaches using the same underlying landmark annotations [12] and localization criteria. Because the three approaches also differ in architecture, pretraining, and model capacity, performance differences cannot be attributed to representation choice alone; the comparison instead isolates the effects of dataset and evaluation protocol, which prior work confounded. The single-frame and multi-frame segmentation models provide the most controlled contrast, as they differ primarily in whether temporal context is available. This design more directly assesses how different localization representations perform in surgical video and offers practical evidence for

choosing among direct keypoint prediction, spatial-region modeling, and temporal context when localizing surgical instrument landmarks.

2. Methods

2.1 Dataset

All experiments were conducted using the JIGSAWS (JHU-ISI Gesture and Skill Assessment Working Set) surgical training dataset. As originally released, JIGSAWS provides stereo endoscopic video and synchronized robot kinematics from surgeons performing suturing, knot-tying, and needle-passing, but its manual annotations cover only surgical gestures and skill scores and do not include per-frame tool landmark labels [12]. Figure 1 shows sample photos from the JIGSAWS dataset. Ghanekar et al. released the landmark annotations used in this study alongside MFCNet, which annotated the JIGSAWS video frames with tool pose maps [4,7]. A pose map is an image of the same size as its corresponding video frame in which each surgical landmark is marked by a small circular region of a distinct color, and all remaining pixels are left as background. Each color identifies one landmark on one instrument, so the jaw base, left tool tip, and right tool tip of the left instrument are marked in different colors from the same three landmarks on the right instrument. The landmark coordinate is recovered by taking the centroid of each colored region. The annotations were created in MATLAB using the ImageLabeler tool [4]. Their annotated set consists of 1,350 training frames sampled at 5 frames per second from 6 randomly selected JIGSAWS videos, giving an average of 225 annotated frames per video, and 450 test frames drawn from 6 additional videos, for a total of 1,800 annotated frames at a resolution of 480 × 640 pixels [4]. Each frame contains two surgical instruments, one on the left and one on the right. Each frame contains six annotated keypoints corresponding to the jaw base, left tool tip, and right tool tip for both the left and right surgical instruments. These six keypoint coordinates served as the prediction target for all localization pipelines. This study used 12 videos, divided at the video level: 6 for training, 3 for validation, and 3 for testing.

Figure 1. Sample video frames from the JIGSAWS surgical training dataset showing robotic surgical instruments and the visual environment used for tool-tip localization.

2.2 Data Preparation

Although all three localization pipelines predicted the same six keypoints, each model required a different target representation during training. Therefore, the original JIGSAWS

annotations were converted into model-specific training targets while preserving the same underlying landmark locations. The ViTPose pipeline used Gaussian heatmap targets, while the DeepLabV3 and MFCNet pipelines used segmentation masks. Sections 3.3, 3.5, and 3.7 describe the conversion procedure for each pipeline.

Before training, all images were resized to the input resolution required by the corresponding model. Because these input resolutions differ across the three models, all predicted landmark coordinates were rescaled back to the original JIGSAWS frame resolution before any localization error was computed. All reported errors are therefore expressed in the same coordinate space and are directly comparable across models.

Training images for the segmentation-based models were augmented online using random horizontal (50% probability) and vertical (50% probability) flipping, random rotations within ±15°**,** and color jittering (50% probability) with small brightness, contrast, and saturation adjustments. Whenever a horizontal flip was applied, the landmark labels were remapped so that the left and right instrument landmarks remained correctly labeled after augmentation. Validation and testing images were resized but were not augmented.

2.3 ViTPose Training Process

For the ViTPose pipeline, the original JIGSAWS pose-map annotations were converted into COCO-style keypoint annotations [6]. Each pose map encoded the jaw base and two tooltips using colored landmark regions. The centroid of each colored region was extracted and used as the corresponding target keypoint coordinate. For each video frame, the left and right instrument annotations were combined into a single six-keypoint annotation containing the jaw base and two tooltip landmarks for both instruments. These COCO-format annotations were then used to train and evaluate the ViTPose model.

For each of the six keypoints, a Gaussian target heatmap was generated at a resolution of 64×48 pixels using a Gaussian kernel with $\sigma = 1.0$. The heatmap reached its maximum value at the annotated keypoint location and gradually decreased away from the center, providing a soft supervision signal for heatmap regression.

ViTPose is available in several variants that share the same architecture but differ in the size of the Vision Transformer backbone. ViTPose-B, the variant used in this study, is built on the ViT-Base backbone and is the smallest commonly used configuration, making it less computationally demanding to fine-tune than the larger ViTPose-L and ViTPose-H variants. The ViTPose-B model used a Vision Transformer backbone with a deconvolutional heatmap prediction head [11]. The final prediction layer was modified to output six heatmaps corresponding to the six surgical keypoints. The model was initialized with pretrained ViTPose weights, and the original prediction layer was replaced because it was incompatible with the six-keypoint task. Training was performed for 20 epochs using the AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 16. The model was supervised using a weighted heatmap mean squared error loss that emphasized pixels near the ground-truth keypoints.

$$\mathcal{L}_{\text{heatmap}} = \frac{1}{KHW} \sum_{k=1}^{K} \sum_{y=1}^{H} \sum_{x=1}^{W} w_{kxy} (\hat{H}_{kxy} - H_{kxy})^2 \quad (1)$$

Equation (1). Weighted heatmap means squared error loss used to train ViTPose. Here, K represents the number of keypoints, H and W represent the heatmap height and width, $\hat{H}_{kxy}$ is the predicted heatmap value, H_{kxy} is the ground-truth heatmap value, and W_{kxy} assigns greater weight to pixels near the ground-truth keypoint than to background pixels.
During validation, each predicted heatmap was converted into a coordinate by selecting the peak heatmap response.

2.4 ViTPose Testing Process

For final evaluation, the latest ViTPose checkpoint selected on the validation set was applied to the held-out test set. Predicted heatmaps were converted into keypoint coordinates by selecting the maximum heatmap response for each keypoint. Each predicted coordinate was compared with the corresponding ground-truth coordinate using Euclidean pixel distance,

$$d = \sqrt{((x_p - x_g)^2 + (y_p - y_g)^2)},$$

Equation 2. Euclidean pixel distance is used to calculate the localization error between the predicted and ground-truth landmark coordinates. Here, d represents the localization error in pixels, (xp,yp) represents the predicted landmark coordinates, and (xg,yg) represents the ground-truth landmark coordinates.

Metrics were computed both per keypoint and per frame. Percentage of Correct Keypoints@5 (PCK@5) and Percentage of Correct Keypoints@10 (PCK@10) were calculated as the percentage of visible keypoints whose predicted coordinates were within 5 and 10 pixels of the ground-truth location, respectively [9]. Annotated output frames were also generated for qualitative inspection.

2.5 DeepLabV3 Training Process

For the DeepLabV3 pipeline, the original JIGSAWS pose-map annotations were converted into segmentation masks. Each colored landmark region in the pose maps was assigned a unique class label, producing pixel-wise annotations consisting of one background class and four foreground classes corresponding to the right tip and jaw base of the right instrument and the left tip and jaw base of the left instrument.

The background class represented pixels that did not belong to a landmark and was not converted into a predicted keypoint. For each instrument, the tool-tip class could form two separate segmented regions, one around each of the two tool tips. A centroid was calculated for each of these regions, producing two tool-tip coordinates, while the jaw-base class produced one

centroid. This resulted in three predicted keypoints per instrument and six predicted keypoints total.

The DeepLabV3 model used a ResNet-101 backbone initialized with ImageNet pretrained weights [4]. The final prediction layer was modified to output five segmentation classes. Training used a combined loss function consisting of Negative Log-Likelihood (NLL) loss and Soft Jaccard loss, weighted at 0.7 and 0.3, respectively. NLL loss measured how accurately individual pixels were assigned to the correct class, while Soft Jaccard loss measured the overlap between the predicted and ground-truth segmentation regions. The combined loss was calculated as shown in Equation (3).

$$\mathcal{L}_{total} = 0.7\mathcal{L}_{NLL} + 0.3\mathcal{L}_{Jaccard} \quad (3)$$

Equation 3. Combined loss function used to train DeepLabV3. Here, L_{total} represents the total training loss, L_{NLL} represents the Negative Log-Likelihood loss, and $L_{Jaccard}$ represents the Soft Jaccard loss.

Training was performed for 20 epochs using a learning rate of 3×10^{-5} and a batch size of 4. These parameters were selected to maximize training stability and model performance within the available computational resources. Because the landmark regions occupied only a small portion of each image, class weighting was applied during training to place greater emphasis on the foreground regions than the background. The background class received a weight of 1, while each of the four foreground classes received a weight of 1000, increasing the contribution of errors on the much smaller landmark regions to the training loss.

During validation, the predicted segmentation masks were converted into landmark coordinates by computing the centroid of each segmented region. A separate centroid was calculated for each of the four foreground classes using the pixels predicted to belong to that class. Each predicted centroid was then compared with the corresponding ground-truth centroid using the Euclidean pixel distance defined in Equation (2).

2.6 DeepLabV3 Testing Process

During testing, the model weights from the final training checkpoint were loaded and used to generate a predicted segmentation mask for each image in the JIGSAWS test set.

The predicted segmentation masks were converted into landmark coordinates by computing the centroid of each segmented landmark region. These predicted coordinates were then compared with the corresponding ground-truth landmark coordinates to calculate localization error. Visualization images showing the predicted segmentation masks and landmark locations were also generated.

To directly compare DeepLabV3 with ViTPose, localization errors from all predicted landmarks were combined to compute the overall mean pixel error, PCK@5, and PCK@10.

2.7 MFCNet Training Process

The Multi-Frame Context Network (MFCNet) extends the single-frame DeepLabV3 model by incorporating temporal information from three consecutive video frames [4]. Each frame was first processed independently by the pretrained single-frame DeepLabV3-ResNet101 model to generate segmentation predictions. For each frame, DeepLabV3 produced five 240 × 320 segmentation maps, with each map containing the model's prediction for one of the five segmentation classes at every pixel. Across the three frames, this produced 15 segmentation maps.

Optical flow was then used to align neighboring-frame predictions with the target frame, while depth maps provided additional spatial information [4]. Optical flow contained two values at each pixel representing horizontal and vertical movement between frames. The depth maps were generated using Depth Anything V2 and stored as single-channel grayscale images, providing one relative depth value at each pixel. Each depth map was resized to the 240 × 320 model input resolution. The optical flow was used to spatially align the segmentation and depth maps from neighboring frames with the target frame before they were combined.

After alignment, the 15 segmentation maps and three depth maps were concatenated along the channel dimension, producing 18 input channels for the multi-frame network. The network consisted of four convolutional layers. The first layer used an 11 × 11 kernel to combine information over a larger spatial region, followed by two 3 × 3 convolutional layers that further processed it. The final 1 × 1 convolutional layer reduced the output to five channels corresponding to the five segmentation classes for the target frame.

As with DeepLabV3, these consisted of one background class and four foreground classes representing the tool-tip and jaw-base regions of the two instruments. The background class was not converted into a keypoint. Each tool-tip class could produce two separate regions, whose centroids represented the two tool tips, while each jaw-base class produced one centroid. The five-class segmentation output was therefore converted into six predicted keypoints total. The network was initialized using the trained single-frame DeepLabV3 checkpoint. During training, the single-frame DeepLabV3 model was kept fixed while only the multi-frame network was updated. This allowed the model to build upon the spatial features learned by DeepLabV3 while learning to combine temporal information across consecutive frames. Training was performed for 20 epochs, with the number of epochs limited by the available computational resources, using a learning rate of 1×10^{-4} and a batch size of 2 (limited by the memory requirements of processing three frames and the additional optical flow and depth inputs simultaneously). Training was conducted for a fixed 20 epochs without early stopping, while validation performance was monitored throughout training. Table 1 below lists the training parameters used for each model.

2.8 MFCNet Testing Process

During testing, the trained MFCNet model generated a predicted segmentation mask for each target frame using three consecutive video frames. Optical flow estimated pixel movement between neighboring frames and was used to align information from adjacent frames before

combining segmentation predictions. Depth maps were incorporated together with the aligned segmentation predictions to improve localization in challenging frames.

The predicted segmentation masks were converted into landmark coordinates by computing the centroid of each segmented landmark region. These predicted coordinates were compared with the corresponding ground-truth landmark coordinates to calculate centroid prediction error. Visualization images were generated showing the predicted segmentation masks together with the predicted and ground-truth landmark locations.

To directly compare MFCNet with the single-frame DeepLabV3 and ViTPose models, we combined centroid errors from all predicted landmarks to compute the overall mean pixel error, PCK@5, and PCK@10.

Table 1 summarizes the training parameters used for each computer vision approach.

Model	Input Size	Epochs	Batch Size	Learning Rate
ViTPose-B	256×192	20	16	5×10^{-5}
DeepLabV3	480×640	20	4	3×10^{-5}
MFCNet	240×320	20	2	1×10^{-4}

3. Results

This study compared three computer vision approaches for surgical tool-tip localization: keypoint estimation using ViTPose, segmentation-based localization using DeepLabV3, and temporal multi-frame localization using MFCNet. Performance was evaluated using mean pixel error, PCK@5, and PCK@10 [9].

Table 2 summarizes the localization performance of all three approaches.

Model	**Mean Pixel Error (pixels)**	**PCK@5 (%)**	**PCK@10 (%)**
ViTPose	11.99	45.28	76.04
DeepLabV3	4.67	71.02	90.38
MFCNet	**3.97**	**74.97**	**94.91**

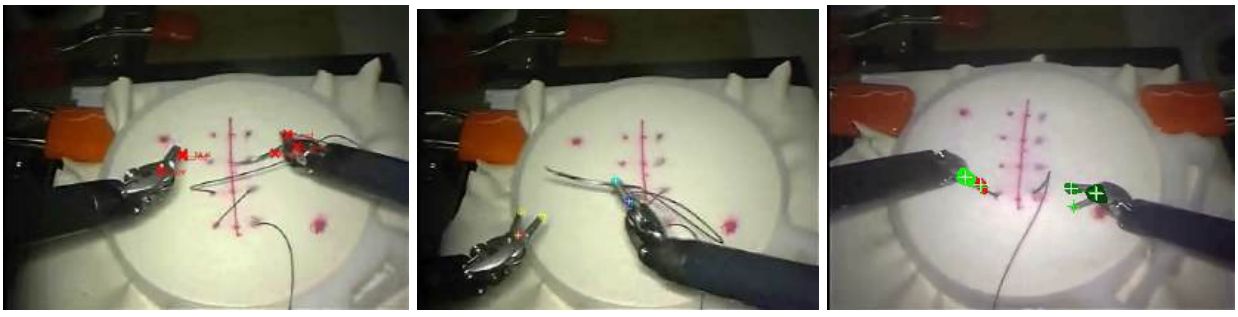


Figure 2. Representative localization results from a single sample frame in the JIGSAWS test set for the three evaluated models: ViTPose(left), DeepLabV3(middle), and MFCNet(right).

As shown in Table 2, MFCNet achieved the lowest mean pixel error (**3.97 pixels)**, indicating the most accurate localization among the three approaches. MFCNet also achieved the highest Percentage of Correct Keypoints, with about **75%** of predicted landmarks within 5 pixels of the ground-truth location and nearly **95%** within 10 pixels.

DeepLabV3 produced the second-best performance across all evaluation metrics, achieving a mean pixel error of **4.67 pixels**, a PCK@5 of **71.02%**, and a PCK@10 of **90.38%**. Although it did not match MFCNet's accuracy, it consistently outperformed ViTPose. ViTPose achieved the lowest localization accuracy, with a mean pixel error of **11.99 pixels**, a PCK@5 of **45.28%**, and a PCK@10 of **76.04%**. Compared with the segmentation-based approaches, ViTPose produced substantially larger localization errors and fewer predictions within both evaluation thresholds.

4. Discussion

The purpose of this study was to investigate which type of visual information is most useful for surgical tool-tip localization. The results suggest that progressively richer visual information improves localization accuracy. Using only keypoint information (ViTPose) produced the lowest performance, incorporating spatial region information through segmentation (DeepLabV3) substantially improved localization, and combining spatial and temporal information (MFCNet) achieved the highest localization accuracy.

ViTPose represents a keypoint estimation approach that directly predicts landmark coordinates from a single image. Although this approach has performed well on human pose estimation tasks, it was less effective for surgical tool localization. One possible explanation is that ViTPose was originally designed for landmarks such as elbows, wrists, and knees [8,11], which are typically separated by much larger distances than surgical tool landmarks. In contrast, the jaw base and tool tips are often only a few pixels apart. As a result, even small prediction errors can confuse neighboring landmarks, leading to larger localization errors.

DeepLabV3 achieved higher localization accuracy because it represented the problem differently. Rather than predicting landmark coordinates directly, it first segmented the image regions surrounding each landmark before computing the landmark locations from those regions. This allowed the model to incorporate spatial context from neighboring pixels, making the predictions more robust to small localization errors and image noise. These findings suggest that segmentation-based localization provides a more effective representation than direct keypoint estimation when surgical landmarks occupy only a small portion of the image.

MFCNet further improved localization performance by incorporating temporal information from consecutive video frames. Instead of relying on a single image, the model combined spatial features with motion information and depth to estimate landmark locations. When a landmark became blurred or partially occluded in one frame, information from neighboring frames helped maintain accurate localization. This ability to use temporal context likely helped MFCNet achieve the lowest mean pixel error and highest PCK scores. These

findings support the fact that temporal information is particularly valuable for surgical video, where instrument motion is naturally continuous across consecutive frames [4,9].

Although MFCNet achieved the highest localization accuracy in this study, several limitations warrant consideration. All experiments used the JIGSAWS surgical training dataset, which contains a limited number of surgical tasks performed in a controlled training environment. Performance may differ on more diverse clinical datasets with additional procedures, instruments, and imaging conditions. In addition, we evaluated only three localization approaches. Other recent computer vision methods, such as video transformers or diffusion-based localization models, may provide further improvements. Also, differences in training parameters, including learning rate, batch size, input resolution, and model pretraining, may have contributed to some variation in model performance.

5. Conclusions

This study addressed the lack of a direct comparison between different computer vision approaches for surgical tool-tip localization. Previous work showed that both heatmap regression and segmentation can be used for surgical landmark localization, but these approaches were evaluated on different datasets, landmark definitions, and metrics, making direct performance comparisons difficult. To address this gap, ViTPose, DeepLabV3, and MFCNet were trained and evaluated using the same JIGSAWS dataset, landmark targets, and evaluation metrics. The results showed a consistent pattern across all three metrics. ViTPose had the highest mean pixel error (11.99 pixels), DeepLabV3 improved substantially to 4.67 pixels, and MFCNet achieved the lowest error (3.97 pixels). PCK@10 followed the same pattern, increasing from 76.04% for ViTPose to 90.38% for DeepLabV3 and 94.91% for MFCNet. These results suggest that representing landmarks as segmented regions may be more effective than directly predicting keypoints when surgical landmarks are small and closely spaced. The additional improvement from DeepLabV3 to MFCNet suggests that incorporating information from consecutive frames can further improve localization accuracy. The comparison between DeepLabV3 and MFCNet provides the clearest evidence for the benefit of temporal information because both approaches use the same DeepLabV3 segmentation model, with MFCNet additionally using information from neighboring frames.

Future work could extend this comparison by evaluating these localization approaches on additional robotic surgery datasets and investigating more advanced temporal localization methods. Future studies could also examine how improvements in surgical tool localization influence downstream applications such as surgical skill assessment, workflow analysis, autonomous robotic assistance, and robotic surgical navigation.

Works Cited

[1] Bouget, David, et al. "Vision-based and marker-less surgical tool detection and tracking: a review of the literature." Medical Image Analysis 35 (2017): 633-654.

[2] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

[3] Du, Xiaofei, et al. "Articulated multi-instrument 2-D pose estimation using fully convolutional networks." IEEE Transactions on Medical Imaging 37.5 (2018): 1276-1287.

[4] Ghanekar, Bhargav, et al. "Video-based surgical tool-tip and keypoint tracking using multi-frame context-driven deep learning models." 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). IEEE, 2025.

[5] Laina, Iro, et al. "Concurrent segmentation and localization for tracking of surgical instruments." International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, 2017.

[6] Lin, Tsung-Yi, et al. "Microsoft COCO: Common objects in context." European conference on computer vision. Cham: Springer International Publishing, 2014.

[7] Shadowfax11. “GitHub - Shadowfax11/Mfcnet-Tracker: Official Code Repository for Ghanekar et Al. 2025 “Video-Based Surgical Tool-Tip and Keypoint Tracking Using Multi-Frame Context-Driven Deep Learning Models” [Accepted to IEEE ISBI 2025].” GitHub, 2025, github.com/shadowfax11/mfcnet-tracker/tree/main. Accessed 5 June 2026.

[8] ViTAE-Transformer. “GitHub - ViTAE-Transformer/ViTPose: The Official Repo for [NeurIPS’22] “ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation” and [TPAMI’23] “ViTPose++: Vision Transformer for Generic Body Pose Estimation.”” GitHub, 2025, github.com/ViTAE-Transformer/ViTPose.

[9] Wang, Jinbao, et al. "Deep 3D human pose estimation: A review." Computer Vision and Image Understanding 210 (2021): 103225.

[10] Wu, Zijian, et al. "Surgpose: a dataset for articulated robotic surgical tool pose estimation and tracking." 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025.

[11] Xu, Yufei, et al. "ViTPOSE: Simple vision transformer baselines for human pose estimation." Advances in Neural Information Processing Systems 35 (2022): 38571-38584.

[12] Yixin Gao, S. Swaroop Vedula, Carol E. Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C. Lin, Lingling Tao, Luca Zappella, Benjam´ın B´ejar, David D. Yuh, Chi Chiung Grace Chen, Ren´e Vidal, Sanjeev Khudanpur and Gregory D. Hager, The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling, In Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop, 2014.

A Cross-Program Benchmark of Electric Propulsion Design By Christian Sent

Abstract

Optimizing the efficiency of electric propulsion (EP) can help maximize payload fraction, shorten mission transfer times, and cut down on required propellant. Since the 1960s, nuclear and solar electric propulsion (NEP and SEP) programs have projected specific mass and specific impulse (Isp) values intended to meet mission targets, yet few high-power systems have flown. This paper asks a question answerable from public data alone: given a program's self-reported specific mass and thruster efficiency, how closely does its chosen or flown Isp align with the mass-optimal Isp implied by classical electric propulsion theory, and what explains any deviation? This paper presents evidence that Isp choices diverge from theoretical mass-optimality in explainable, non-random ways. To find these results, specific mass, thruster efficiency, and mission-duration data from major SEP missions and NEP design studies are compiled and evaluated against Ernst Stuhlinger's power-limited optimum exhaust velocity relation. This study is significant because it lets current high-power proposals be benchmarked against a mass optimality standard. Additionally, this paper provides localized data collection of a wide array of missions, and identifies specific gaps in public data records. Comparison against the historical trends within this dataset help provide a more realistic outlook on EP as an industry.

Keywords Electric propulsion, Stuhlinger optimization, specific mass, specific impulse

Introduction

Electric propulsion (EP) is an incredibly efficient technology that uses electricity to accelerate a propellant and generate thrust. Instead of burning fuel like conventional chemical rockets, EP ionizes a gas into charged particles and then uses electric or magnetic fields to shoot these ions out at high velocities.[1] Due to the acceleration mechanism being electrical rather than chemical, EP thrusters can reach specific impulse (Isp) values in the realm of 1200-1500 seconds.[1] In contrast, chemical propulsion has a much lower range that is commonly between 300-450 s.[2] The higher Isp of EP reduces the mass of propellant needed to achieve a given velocity change, which can allow for higher payloads or smaller spacecraft. Unfortunately, while efficient, EP systems have very low thrust magnitudes.[1] Therefore, they are often required to run continuously and compound speed over time. This can substantially extend mission transfer times.

There are two primary classifications of EP: solar electric propulsion (SEP) and nuclear electric propulsion (NEP). SEP draws power directly from solar panel arrays, and NEP converts heat from a fission reactor into electricity through a power conversion cycle.[3] Of course, each type of technology is best suited towards different mission tasks, and SEP has a much more extensive history of real-world use. In trajectory optimization, SEP array output varies inversely

to the square of the distance from the sun, while an NEP reactor is treated as a constant source of power.[4]

Within both power architectures, there are several different thruster types that convert electrical power into thrust. Gridded ion thrusters (GIT) accelerate ions through charged grids with high voltage differences.[1] Hall effect thrusters (HET) ionize propellant within a channel using a radial magnetic field,[1] and are therefore not subject to the grid-spacing thrust limitations of GITs. Finally, magnetoplasmadynamic (MPD) thrusters rely on the Lorentz force generated by current flowing through ionized propellant and offer the highest thrust and power density amongst the above options.[5] This makes MPD thrusters popular amongst high-power NEP concepts, as GITs and HETs are more often used in lower power applications.

A quantity that can govern how well an EP system performs is known as specific mass, defined as the mass of a subsystem divided by the electrical power it processes or delivers.[6] This metric is expressed in kilograms per kilowatt-electric (kg/kWe). Specific mass can be applied both to individual subsystems, and to the EP system as a whole. Interestingly, specific mass can function as a common unit across dissimilar hardware: a solar panel and a fission reactor can be compared in a similar manner once their output is expressed per kilowatt of delivered power. Specific Impulse (Isp) is another useful metric that measures the ratio of thrust to propellant weight flow rate.[2] In effect, Isp values demonstrate how efficiently a propulsion system uses propellant.

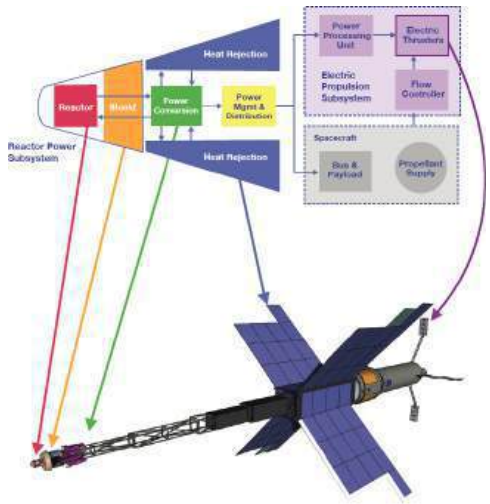


Figure 1: Nuclear electric propulsion subsystems and conceptual design.[3]

As Figure 1 shows, an NEP spacecraft is commonly described in terms of smaller subsystems: the reactor, a radiation shield, a power conversion subsystem that transforms reactor heat into electricity, waste heat radiators, a power management and distribution (PMAD) subsystem, and the EP subsystem itself [3]. This EP subsystem includes the power processing unit (PPU), propellant management system, and thrusters. An example of an NEP spacecraft is shown above in Figure 1.

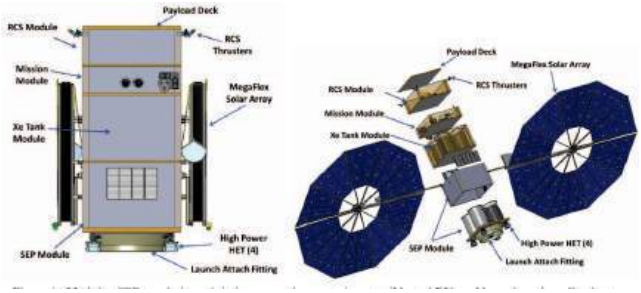


Figure 2: Solar electric propulsion system example design.[7]

A SEP spacecraft follows a similar structure, but photovoltaic arrays replace the reactor, shield, and power conversion subsystem.[8] An example structural layout of a SEP spacecraft can be seen in Figure 2. A SEP PPU also requires an added DC-to-AC inversion stage that the NEP PPU does not, since the array output is low-voltage DC rather than the high-voltage AC produced by NEP's power conversion subsystem.[8] Given the variations of subsystems between EP designs, a more general categorization framework has been historically used in design studies: payload, propulsion system, propellant, and propellant tankage.[4]

Ultimately, this paper asks how accurately EP programs' chosen or flown Isp aligns with the mass optimal Isp implied by their own reported specific mass and thruster efficiency. This question holds importance because no high power NEP system has ever flown, so the traditional approach of comparing predicted to achieved performance cannot be applied to roughly half of this paper's chosen EP programs. By using a theoretical optimality standard, though, comparisons can be applied to any program that has published its design parameters. This paper reveals that Isp diverges from theoretical mass-optimality in explainable, non-random ways. Across every program with publicly accessible data, actual or design Isp falls below the Stuhlinger-implied optimum.

Section I of this paper discusses how EP is used in the modern industry, and what technology is foundational. Section II presents the methods used to calculate each program's theoretical mass optimal Isp. Section III analyzes key data points from relevant SEP missions and studies. Section IV analyzes the same metrics from relevant NEP design studies. Section V compiles the information into two normalized comparison tables spanning both technologies. Section VI applies Stuhlinger's optimum exhaust velocity relation to a subset of these programs, benchmarking each against its own theoretical mass optimal Isp.

Section I - The Modern Industry

Electric propulsion has become the default choice for large parts of the satellite industry. A primary driver of this shift has been GEO communications satellites, where operators use EP for station-keeping and trajectory adjustments instead of chemical thrusters. For example, Boeing's 702SP platform, which first launched in 2015, was the first fully electric commercial satellite bus, relying on xenon ion propulsion system thrusters for all maneuvering.[9] By removing

chemical propellant, launch mass was substantially reduced, which lowered launch costs and freed up mass for revenue-generating payload.[9]

EP has also become a core technology for LEO megaconstellations. SpaceX's Starlink constellation, with over 10,000 satellites in orbit, equips each satellite with a krypton or argon-fueled Hall effect thruster.[10] Unlike GEO comsats, which fly a small number of high-value, long-life spacecraft, megaconstellations depend on EP systems that are cheap, compact, and mass-produced. They favor simpler and lower-power Hall thrusters over the higher-efficiency gridded ion designs used in some GEO applications.[1-10] This has pushed EP manufacturing toward high-volume production.

Beyond Earth orbit, EP is the enabling technology behind many deep-space science missions. NASA's Deep Space 1 and Dawn missions demonstrated that ion propulsion could deliver effective total velocity changes, with Dawn's ion engines providing roughly 11 km/s of cumulative Δv over its mission to Vesta and Ceres.[11-12]

At the high-power end of the industry, NASA's Gateway Power and Propulsion Element (PPE) represents the current state of the art in SEP, and a useful bridge toward the higher-power regimes this paper is ultimately concerned with. PPE will generate 60 kWe of electrical power and use multiple Hall thrusters, including 12 kW class units. This makes it the most powerful SEP spacecraft ever flown once operational.[13] It intends to support long-duration lunar infrastructure rather than singular flybys or orbiter missions, thus illustrating the industry's growing interest in EP systems that must sustain high power output.[13]

Taken together, these programs show that SEP has an extensive and growing flight record across very different mission classes. NEP, by contrast, has no equivalent flight heritage. Every true NEP system discussed in this paper exists only as a design study, ground test, or canceled program.

Section II - Methods

This paper's central comparison focuses on the results from Ernst Stuhlinger's power limited optimization of electric propulsion trajectories.[5-6] This method looks at a specific question: for a propulsion system whose powerplant mass scales with the electrical power is processes, what exhaust velocity maximizes the payload fraction delivered over a fixed thrust duration? The answer is reliant on three core inputs: specific mass α, thruster efficiency η, and thrust duration t. Given these three inputs, the exhaust velocity is:

$$V_{ex,opt} = \sqrt{(2\eta t/\alpha)}$$

and the corresponding optimum specific impulse is:

$$Isp_{opt} = v_{ex,opt}/g_0$$

where $g_0 = 9.80665$ m/s^2 is the standard gravitational acceleration.[6] Table 3 in Section VI applies these equations to all programs in the dataset where α, η, and t are all publicly reported. These

values are then compared against each program's actual or design Isp. A ratio near 1.0 indicates a program operated close to its own theoretical mass optimality. In contrast, a ratio well below 1.0 suggests that a program's chosen Isp fell short of what the Stuhlinger relation would have recommended for that program's specific mass, efficiency, and mission duration.

Results and Discussion

Section III - SEP Programs

1. Deep Space 1 (October 1998)

Deep Space 1 was the first NASA mission to use ion propulsion as a primary propulsion system. This mission flew a single NSTAR ion thruster (a 28.6 cm ring-cusp design) developed by NASA Glenn, JPL, and Hughes/Boeing.[1-3] At its full-power point of 2325 W, NSTAR produced 92.7 mN of thrust at a specific impulse of 3127 s and 61.8% total efficiency. At its minimum-power point of 518 W it produced 20.7 mN at 1979 s and 42.0% efficiency.[1] The thruster's mass was 8.2 kg and its power processing unit (PPU) was 14.77 kg, with a separate digital control and interface unit (DCIU) at 2.51 kg. This gave a combined thruster-plus-PPU specific mass of 10.0 kg/kWe at the 2.3 kWe operating point.[1-3] Its SCARLET solar array included two wings that weighed 55.4 kg. Producing 2.5 kW at 1 AU, SCARLET had a specific mass of 22.2 kg/kWe. Therefore, combined thruster, PPU, and array specific mass was approximately 32 kg/kWe.[1] This combined value becomes the baseline against which Section VI evaluates whether NSTAR's 3127 s operating point matched the mass optimal exhaust velocity implied by a 32 kg/kWe power system.

2. Dawn (September 2007)

Dawn was the first NASA mission to rely on ion propulsion for its entire post-launch Δv, using the same NSTAR thruster design as Deep Space 1 in a triple-redundant configuration: three thrusters and two PPUs, with only one thruster active at a time.[11] Operating at the same 2.3 kWe full-power point (92.7 mN, 3127 s, 61.8% efficiency), each thruster weighed 8.2 kg and each PPU was 14.77 kg.[3-11] This gives an identical specific mass of 10.0 kg/kWe. The full triple-redundant ion propulsion system included two digital control units, three thruster gimbals, and a xenon feed system.[11] It had a dry mass of 129 kg excluding propellant, which is equivalent to approximately 56 kg/kWe when divided by the single-thruster operating power; this higher figure reflects the mass of carrying redundant hardware, not a loss of thruster efficiency.[11] Moreover, Dawn's solar array, two wings totaling 126 kg, produced 10.3 kW at 1 AU, and fell to 1.3 kW at 3 AU end-of-life.[11] This gives an array specific mass of 12.2 kg/kWe and a combined specific mass of approximately 22 kg/kWe.[11] The spacecraft carried 425 kg of xenon, having a wet mass of approximately 1219 kg.[11] Since Dawn reused the NSTAR's operating point over a mission roughly three times longer, the two missions provided a test for whether an Isp chosen

for one mission remains mass-optimal when the same hardware is deployed on a longer duration mission.

3. Boeing 702SP (March 2015)

The Boeing 702SP was the first commercial satellite line to eliminate chemical propulsion entirely, relying on four 25-cm XIPS ion thrusters and two PPUs per satellite.[9-14] In its high-power mode (4.3 kW), XIPS-25 delivers 165 mN of thrust at a specific impulse of 3550 s and 68.8% total efficiency; in its low-power station-keeping mode (2.0 kW) it delivers 79 mN at 3400 s and 67% efficiency.[14] The thruster weighs 13.7 kg and its PPU is 21.3 kg, giving a thruster-PPU specific mass of 8.1 kg/kWe at high power and 17.5 kg/kWe at low power.[14] Unfortunately, solar array mass data is not publicly available, so array and combined specific mass cannot be reported for this platform. This gap highlights a repeat limitation in the public record. Proprietary data prevents a full specific mass benchmark, which restricts 702SP to only a thruster level comparison.

4. BepiColombo (October 2018)

BepiColombo's Mercury Transfer Module will separate in September 2026 ahead of Mercury orbit insertion planned for November 2026.[15] The module carries four QinetiQ T6 ion thrusters, a 5 kW-class Kaufman-type design, where two operate simultaneously with a single PPU in the baseline architecture.[16] Each thruster produces 75-145 mN of thrust at a specific impulse reaching approximately 4120 s at full power in test data.[16] This spacecraft has up to 290 mN of combined thrust from two simultaneously firing thrusters, and its total efficiency is not publicly reported.[16] The full baseline electric propulsion system, comprising four thrusters, one PPU, four flow control units, and harnessing, has a mass of 131.2 kg, with architecture variants ranging from 65 to 139 kg depending on redundancy level.[16] Additionally, the transfer module's solar array has a reported mass of 290 kg, but its power rating conflicts by more than a factor of six between two sources, so neither array nor combined specific mass is reported.

5. Psyche (October 2023)

Psyche is flying Hall thrusters beyond cislunar space, using four SPT-140 thrusters and two PPUs to reach the metal asteroid of the same name.[17] The thrusters operate across a power range of 0.9-4.5 kW at a specific impulse of 1800 s.[17] This produces a beginning-of-life thrust prediction of approximately 260 mN at 4.5 kW.[17] Thruster mass is 8.5 kg which gives 1.9 kg/kWe, but unfortunately PPU mass is not found in the accessible literature so thruster-plus-PPU specific mass cannot be fully reported.[17] Psyche's twin five-panel solar arrays total 75 square meters and produce 20 kW at 1 AU, but array mass, and therefore array and combined specific mass, was not published.[17] Interestingly, this mission carries a xenon load of 1030 kg, the largest of any flown mission to date.[17]

6. Gateway PPE

The Power and Propulsion Element (PPE) combines three 12 kW AEPS Hall thrusters (Aerojet Rocketdyne, HERMeS-derived, magnetically shielded, 600 V) with four 6 kW Busek BHT-6000 Hall thrusters across ten power processing units on a Maxar 1300-series bus [13]. AEPS operates at a specific impulse of 2800 s and 67% total efficiency.[3] Its thruster is 47.5 kg and its PPU is 50.0 kg, giving a thruster-plus-PPU specific mass of 7.8 kg/kWe.[13] The BHT-6000 thruster weighs 12.5 kg, at 2.08 kg/kWe [9]. PPE's Redwire ROSA solar array has a published specific power of 100-120 W/kg, which is equivalent to a specific mass of 8.3-10.0 kg/kWe.[13] When combined with the AEPS thruster and PPU mass, this gives a system specific mass of approximately 16-18 kg/kWe. The spacecraft also carries 2770 kg of xenon at launch in dual 825-liter tanks, is refuelable in orbit, and has a total impulse capability exceeding 150 MN-s.[13] At 16-18 kg/kWe, PPE's combined specific mass sits below Deep Space 1's and Dawn's figures, which suggests that improvement was made via a lighter power system per kilowatt delivered.

Section IV - NEP Programs

1. SNAP 10-A (April 1965)

SNAP-10A is the only nuclear reactor the United States has launched into space [3]. Its 30 kWt reactor produced approximately 0.5 kWe via silicon-germanium thermoelectric conversion and operated for 43 days before a voltage regulator failed [3]. Despite this, a parallel ground-test ran for more than 10,000 hours.[3] The unshielded reactor massed approximately 290 kg, and the full flight system was under 430 kg, for a whole-system specific mass of approximately 1900 kg/kWe [3]. SNAP-10A was part of the broader SNAP program (1958-1972), which spanned electrical outputs from 0.5 kWe (SNAP-10A) to 350 kWe (SNAP-50).[3]

2. SP-100 (1983 to 1994)

SP-100 was a joint NASA, Department of Defense, and Department of Energy program that aimed to develop a 100 kWe-class reactor with a growth path to approximately 1 MWe using dynamic power conversion.[3] Unfortunately, the program was cancelled before any flight hardware was created, but the highly enriched uranium-fueled reactor had a design outlet temperature of approximately 1350 K.[3] A thermoelectric converter tested in 1993 achieved 1.5 kWe at a 1300 K reactor exit temperature and 4.2% efficiency.[3] The program's specific mass design target was 20-30 kg/kWe for the power system, but its Reference Flight System came in at 5675 kg for 100 kWe, excluding propellant and feed system, for a specific mass of 56.75 kg/kWe.[18] The overall program weight goal was approximately 4500 kg, or about 45 kg/kWe.[19]

3. Topaz-II

TOPAZ-II, a Soviet-designed thermionic reactor, was purchased by the United States in 1992 for ground testing only.[3] The reactor produced 5.0-5.47 kWe from a total system mass of approximately 1000 kg (one tonne), for a specific mass of approximately 183-200 kg/kWe.[3-20]

The system stood 3.9 m tall and had a 1.4 m base diameter, with its core comprising 37 single-cell thermionic fuel elements in a zirconium hydride moderator.[20]

4. Project Prometheus (2003-2005)

Project Prometheus intended to create a Jupiter Icy Moons Orbiter (JIMO), and it reached component-level testing before budget reprioritization led to its cancellation.[3-21] Source documents vary on the exact reactor design point, citing figures between 180 and 200 kWe.[21-22] The reactor module weighed 6182 kg, for a specific mass of 30.9 kg/kWe at 200 kWe.[3] Combined with a heat rejection subsystem specific mass of 10.1 kg/kWe, the reactor-plus-heat-rejection figure is approximately 41 kg/kWe.[3] Total spacecraft launch mass is estimated at 36,375-37,000 kg across two independent sources, with a design xenon tank capacity of 18,000 kg.[21] The program had a 2 kWe Brayton power conversion unit which powered a 2.3 kWe ion thruster for 800 hours.[3] Interestingly, in a direct-drive power demonstration a 1.6 kW Brayton system rectified 55 V AC to 1100 V DC at 91% efficiency.[3] A later study found that the baseline reactor and design envelope could satisfy five of six candidate follow-on destinations, a Saturn tour, a Neptune tour, a Kuiper Belt Object rendezvous, a multi-asteroid sample return, and a comet sample return, with only an interstellar precursor mission excluded.[23]

5. Compass

The COMPASS team published a 2021 paper study for a hybrid nuclear-electric/chemical propulsion vehicle intended to reduce a crewed Mars round trip to two years for 2039.[24] No hardware was built and reactor or system mass was not reported.[24] The design targeted approximately 1.9 MWe of power, and set a specific mass requirement of 20 kg/kWe for the entire nuclear electric propulsion system, stated as a design goal rather than an achieved or tested result.[24] The reactor outlet temperature was set at approximately 1200 K, and the piloted vehicle was constrained to five Space Launch System launches with an 8.4 m fairing; radiator packaging within a single fairing serving as the limiting design driver.[24]

6. Kilopower (2018)

Kilopower demonstrated a small, scalable, Stirling-converted reactor technology across 1-10 kWe.[25] Its KRUSTY ground-test experiment ran for 28 hours at full power, validating the concept, though the technology has never flown or been used for propulsion.[25] NASA's stated design range for specific power is 2.5-6.5 W/kg; a 1 kWe system has been estimated at approximately 400 kg (400 kg/kWe), and a 10 kWe upsized system with a highly-enriched-uranium core at approximately 1300 kg (130 kg/kWe).[25] This highlights how specific mass improves as power scales up.

Section V - Comparative Program Data

Table 1: SEP Missions

Mission	Status	Launch date	Thruster type	Power (kWe)	Specific impulse (s)	Thruster + PPU specific mass (kg/kWe)	Solar array specific mass (kg/kWe)	Combined specific mass (kg/kWe)
Deep Space 1	Flown[1-3]	Oct 1998[1-3]	NSTAR[1-3]	2.3[1-3]	3127[1-3]	10.0[1-3]	22.2[1-3]	~32[1-3]
Dawn	Flown[3-11]	Sept 2007[3-11]	NSTAR x3[3-11]	2.3 (1 op.)[3-11]	3127[3-11]	10.0 / ~56*[3-11]	12.2[3-11]	~22[3-11]
Boeing 702SP	Flown[9-14]	Mar 2015[9-14]	XIPS-25[9-14]	2.0–4.3[9-14]	3400–3550[9-14]	8.1–17.5[9-14]	Not reported	Not reported
BepiColombo	Flown[15-16]	Oct 2018[15-16]	T6 x4[15-16]	5.0[15-16]	>4000[15-16]	Not reported**	Not reported	Not reported
Psyche	Flown[17]	Oct 2023[17]	SPT-140 x4[17]	0.9–4.5[17]	1800[17]	Not reported**	Not reported	Not reported
Gateway PPE	Planned[3-13]	2028[3-13]	AEPS + BHT-6000[3-13]	12.5 / 6.0[3-13]	2800[3-13]	7.8 (AEPS)[3-13]	8.3–10.0[3-13]	~16–18[3-13]

** Dawn: 10.0 kg/kWe is the thruster+PPU figure (identical hardware to Deep Space 1) & ~56 kg/kWe is the triple-redundant IPS mass divided by single-thruster operating power. ** BepiColombo and Psyche figures are thruster-only*

Table 2: NEP Missions

Program	Status	Program dates	Power (kWe)	Specific mass (kg/kWe)
SNAP-10A	Flown[3]	Flown 1965[3]	0.5[3]	~1900[3]
SP-100	Cancelled, never flew[3]	1983–1994[3]	100 (design)[18-19]	Target 20–30; actual 45–57[18-19]
TOPAZ-II	Ground-tested only[3-20]	Purchased 1992[3-20]	5.0–5.47[3-20]	~183–200[3-20]
Prometheus (JIMO)	Cancelled, never flew[3]	2003–2005[3]	180–200[21-22-23]	Reactor module 30.9; +heat rejection ~41[21-22-23]
Compass (NEP-Chem)	Paper study, not built[24]	2006 / report 2021[24]	~1,900,000[24]	Target 20[24]
Kilopower	Ground-tested only[25]	2015; tested 2017–18[25]	1–10[25]	130–400[25]

Table 1 and Table 2 allow for a centralized, legible way to compare specific mass across two different EP technologies, and over six decades of programs. Read together, the tables make the paper's central data problem visible: combined specific mass, the figure that matters most for cross-program comparison, is missing for a number of SEP programs. Specific mass alone, however, doesn't fully convey how effective an EP system is. The next section applies a theoretical optimum to a subset of these programs to provide more comprehensive data analysis.

Section VI - Stuhlinger Optimum-Isp Benchmarking

Ernst Stuhlinger's power limited optimization gives the exhaust velocity that maximizes payload fraction for a system with specific mass α and thruster efficiency η operating

continuously for duration t.[6] Crucially, using this methodology lets a flown SEP mission and a cancelled NEP design study be benchmarked on the same footing.

This analysis is limited to four programs in this paper that satisfy the necessary data requirements for combined, full-system specific mass, thruster efficiency, and continuous-thrust duration. These programs are Deep Space 1, Dawn, Project Prometheus (JIMO), and Compass. Note that where a source reported a range rather than a single value, both endpoints of the range are shown rather than an assumed midpoint.

Table 3: Stuhlinger Isp

Program	α (kg/kWe)	η	Continuous thrust duration	Actual/design Isp (s)	Stuhlinger-optimal Isp (s)	Actual:Optimal
Deep Space 1	32.0 (combined)[3-26]	0.618[3-26]	>16,000 hr (total in-space operation)[3-26]	3,127[3-26]	4,810[3-26]	0.65[3-26]
Dawn	22.2 (combined)[11-12]	0.618[11-12]	50,000 hr (~5.7 yr, full mission)[11-12]	3,127[11-12]	10,208[11-12]	0.31[11-12]
Project Prometheus (JIMO)	30.0 (NRA full-system target)[26]	0.76 (measured, 24.4 kW row)[26]	7 yr (lower bound of stated range)[26]	8,270[26]	10,784[26]	0.77[26]
Project Prometheus (JIMO)	30.0 (NRA full-system target)[26]	0.76 (measured, 24.4 kW row)[26]	14 yr (upper bound of stated range)[26]	8,270[26]	15,251[26]	0.54[26]
Compass	19.4 (Table I mass ÷ 1.8 MWe)[27]	0.594[27]	259 days effective (288-day transit × 90% duty cycle)[27]	2,600[27]	3,773[27]	0.69[27]
Compass	18.4 (Table I mass ÷ 1.9 MWe, title figure)[27]	0.594[27]	259 days effective (288-day transit × 90% duty cycle)[27]	2,600[27]	3,877[27]	0.67[27]

In Table 3, every EP program falls short of its own theoretical optimum. Interestingly though, the gap reveals how far each mission asked its propulsion system to operate outside the designated Isp range it was chosen for. Deep Space 1's NSTAR thruster, evaluated against the approximate 16000 hours of flight time,[26] lands at a ratio of 0.65. This is reasonably close, since the mission length was in the same range as what the NSTAR had been developed and life-tested against. Dawn, on the other hand, took the same hardware at an identical operating point,[11] and had it operate in a mission with a duration more than three times longer.[12] As a result, the ratio went down to 0.31, which is a sharp deviation. Additionally, Dawn's xenon load exceeded its combined powerplant mass by a factor of 1.67,[11-12] which is well past the roughly 1:1 split that Stuhlinger's optimum predicts.[6] In contrast, Deep Space 1 landed within a percent of that ratio.[1-3]

Even more interestingly, however, is that Project Prometheus and Compass also don't land near to their optimums. Prometheus' HiPEP's measured Isp of 8,270 s sits at 0.54 to 0.77 of its own optimum depending on which end of JIMO's stated 7-to-14-year duration requirement

applies.[26] This spread shows how unresolved some of the mission's core parameters were when the program was cancelled. Furthermore, Compass, which was designed a generation later around an explicit whole-system mass target, lands at 0.69.[27] This highlights how programs without hardware reuse restraints like Dawn still hit roughly the same wall.

According to the Compass paper, radiator packaging inside a single SLS fairing set the ceiling on its Isp choice.[27] This was not a mass constraint; it was a volume constraint. HiPEP existed specifically because NSTAR-derived hardware was unable to meet life requirements, and its higher-Isp replacement was not flight-qualified hardware when Prometheus was cancelled.[26] This reveals a technology readiness constraint. Ultimately, in many of these cases, the constraint was not solely mass. Stuhlinger's formula treats specific mass as fixed regardless of the exhaust velocity chosen.[6] Unfortunately, most pathways to higher Isp in this dataset carry costs that the formula doesn't account for.

Conclusion

This paper set out to analyze how EP programs' Isp aligned with the mass-optimal Isp implied by their reported specific mass and thruster efficiency, and to explain any deviations. The evidence compiled across Sections III through VI supports the paper's central thesis: every program in the case study operated or was designed to operate below its Stuhlinger optimal Isp. In each case, shortfalls are traceable back to specific, documented causes rather than uniform patterns of underperformance in repeated areas of concern.

This framework is useful to engineers today because it doesn't require flight data to apply. Any program can be benchmarked against its own mass optimality standard as long as its specific mass, thruster efficiency, and intended mission duration are known. Current high-power EP proposals could use this methodology as a sanity check before committing to new, costly hardware.

Moreover, the clearest gap that this paper identifies is that the public record for combined specific mass is unavailable for multiple missions, such as BepiColombo and Psyche. Closing this gap would let future work extend the Stuhlinger benchmark across a wider array of missions or design studies. Also, Stuhlinger's relation treats specific mass as fixed regardless of the exhaust velocity chosen. However, the source material in Section VI indicates that higher Isp hardware carries real mass, volume, and qualification costs that the formula doesn't capture. A version of this formula that lets α vary with Isp would likely provide more precise results.

Acknowledgments

I would like to acknowledge Dr. Devin Carroll and Kieran Tait from Indigo Research for their insightful feedback throughout the research process. Additionally, I would like to thank my family for their constant support and encouragement. I attest that the ideas, graphics, and writing in this paper are entirely my own.

Works Cited

1. Goebel, D. M. "Flight Ion and Hall Thrusters." Fundamentals of Electric Propulsion: Ion and Hall Thrusters, JPL/Caltech Descanso Series, Jet Propulsion Laboratory, ch. 9.
2. Sutton, G. P., and O. Biblarz. Rocket Propulsion Elements. 9th ed., John Wiley & Sons, 2016.
3. National Academies of Sciences, Engineering, and Medicine. Space Nuclear Propulsion for Human Mars Exploration. The National Academies Press, 2021, doi.org/10.17226/25977.
4. Oleson, S. R. An Analytical Optimization Method for Electric Propulsion Orbit Transfer Vehicles. Sverdrup Technology, Inc./NASA Lewis Research Center, May 1993, NASA CR-191129.
5. Jahn, R. G. Physics of Electric Propulsion. McGraw-Hill, 1968.
6. Stuhlinger, E. Ion Propulsion for Space Flight. McGraw-Hill, 1964.
7. Deininger, W. D., et al. "Modular Solar Electric Propulsion (SEP) Tug Concept." 2016 IEEE Aerospace Conference, 2016, pp. 1-12.
8. Frisbee, R. H., and I. G. Mikellides. "The Nuclear-Electric Pulsed Inductive Thruster (NuPIT): Mission Analysis for Prometheus." 41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Tucson, AZ, July 2005, AIAA Paper 2005-3892.
9. eoPortal Satellite Missions Database. "All-Electric." European Space Agency.
10. Oh, D. Y., and P. Neumann. "A Year of Firsts for Electric Propulsion." Aerospace America, AIAA, 1 Dec. 2023.
11. Brophy, J. R., et al. "The Dawn Ion Propulsion System – Getting to Launch." 30th International Electric Propulsion Conference, Florence, Italy, 2007, IEPC-2007-083.
12. German Aerospace Center (DLR). "The Ion Propulsion System." DLR Mission Dawn, www.dlr.de/en/research-and-transfer/projects-and-missions/dawn/the-ion-propulsionsystem. Accessed 15 Aug. 2026.
13. Herman, D., et al. "Development and Qualification Status of the Electric Propulsion Systems for the NASA PPE Mission and Gateway Program." 37th International Electric Propulsion Conference, 2022, IEPC-2022-465.
14. Goebel, D. M., et al. "Evaluation of 25-cm XIPS Thruster Life for Deep Space Mission Applications." 31st International Electric Propulsion Conference, Ann Arbor, MI, 2009, IEPC-2009-152.
15. European Space Agency. "BepiColombo Factsheet." ESA Science & Technology.
16. Hutchins, M., et al. "QinetiQ's T6 and T5 Ion Thruster Electric Propulsion System Architectures and Performances." Joint Conference of the 30th ISTS, 34th IEPC and 6th NSAT, Kobe-Hyogo, Japan, 2015, IEPC-2015-131/ISTS-2015-b-131.
17. Snyder, J. S., et al. "Electric Propulsion for the Psyche Mission." 36th International Electric Propulsion Conference, Vienna, Austria, 2019, IEPC-2019-244.
18. "Spacecraft and Mission Design for the SP-100 Flight Experiment." IEPC-1988-013, 1988.
19. Office of Scientific and Technical Information (OSTI). SP-100: The U.S. Space Nuclear Reactor Power Program. U.S. Department of Energy.

20. Office of Scientific and Technical Information (OSTI). TOPAZ II System Description. U.S. Department of Energy.
21. NASA Jet Propulsion Laboratory. Prometheus Project Final Report. Oct. 2005, JPL 982-R120461.
22. Oleson, S. R. Electric Propulsion Technology Development for the Jupiter Icy Moons Orbiter Project. NASA Glenn Research Center, 2004, NASA/TM-2004-213290.
23. Noca, M. A., and K. J. Hack. "Results of the JIMO Follow-on Destinations Parametric Studies." Space Technology and Applications International Forum (STAIF), Albuquerque, NM, 2005.
24. Oleson, S. R., et al. Compass Final Report: Nuclear Electric Propulsion (NEP)-Chemical Vehicle 1.2. NASA Glenn Research Center, 2021, NASA/TM-20210017131.
25. NASA. Kilopower Reactor Development and the Path to Higher Power Missions. 2017, NASA/TM-2017-219467.
26. Foster, J. E., et al. "The High Power Electric Propulsion (HiPEP) Ion Thruster." NASA Glenn Research Center, 2004, NASA/TM-2004-213194, AIAA-2004-3812.
27. Oleson, S., et al. 1.9 MWe Nuclear Electric Propulsion-Chemical Propulsion Piloted Mars Opposition Vehicle. NASA Glenn Research Center, 2022, NETS-2022-38628.
28. Oleson, S., et al. "A Kuiper Belt Object Orbiter Enabled by 10 kW Kilopower Electric Propulsion." AIAA Propulsion and Energy Forum, Indianapolis, IN, 2019, AIAA 2019-3963.
29. Jet Propulsion Laboratory. Kilopower – Nuclear Electric Propulsion for Outer Solar System Exploration (NEP Benefits Study). 2020, JPL D-103385.
30. Krebs, G. D. "JIMO (Prometheus 1)." Gunter's Space Page, space.skyrocket.de/doc_sdat/jimo.htm. Accessed 15 Aug. 2026.
31. Sengupta, A., et al. "An Overview of the Results from the 30,000 Hour Life Test of the Deep Space 1 Flight Spare Ion Engine." 40th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Fort Lauderdale, FL, 2004, AIAA Paper 2004-3608.

Evaluating ASPM as a Genetic Modifier of Congenital Zika-Associated Microcephaly
By Aanya Sood

Abstract

Prenatal Zika virus (ZIKV) infection does not affect every fetus the same; only some exposed fetuses develop Zika syndrome, including severe microcephaly, suggesting that genetic differences may influence susceptibility. The purpose of this review is to specifically examine whether variants in ASPM, the gene most commonly associated with inherited primary microcephaly, could alter how prenatal ZIKV infection affects the developing brain. ASPM helps neural progenitor cells divide correctly and produce the cells necessary for brain growth. Any harmful variants in the ASPM gene can lead to autosomal recessive primary microcephaly. Overall, the findings indicate that ZIKV disrupts the growth and survival of developing brain cells and alters pathways associated with microcephaly. Past studies have shown harmful health effects of ZIKV. However, no study has directly compared ASPM variants in fetuses exposed to ZIKV with microcephaly versus those without. Although not confirmed, current evidence supports a possible gene-environment interaction in which mild or regulatory ASPM variants may make neural progenitor cells more susceptible to damage from ZIKV infection. Identifying this gap in research could contribute to significant progress in this field by guiding future genetic sequencing, explaining why fetal outcomes differ after ZIKV exposure, and helping researchers distinguish inherited microcephaly from Zika-associated microcephaly.

Keywords *ASPM*; Zika virus; congenital Zika syndrome; microcephaly; p53 stress pathway; genetic susceptibility; genetic sequencing; neural progenitor cells

Introduction

Prenatal exposure to the same virus can produce very different outcomes. While some fetuses exposed to Zika virus (ZIKV) develop congenital Zika syndrome, including severe microcephaly, others are born without major congenital abnormalities. This variation suggests that ZIKV exposure alone may not determine whether microcephaly develops and that fetal genetics could influence susceptibility (Marques et al., 2025).

ZIKV can cross the placenta and infect cells that are essential for fetal brain development. ZIKV targets radial glial cells, which are neural progenitor cells responsible for producing many of the neurons in the cerebral cortex. Neural progenitor cells are crucial for brain development. Consequently, the infection reduces the neural progenitor population, preventing cortical development, and decreases the expression of several genes associated with microcephaly (Wu et al., 2016). The gene most commonly associated with microcephaly is *ASPM*, which helps neural progenitor cells divide symmetrically, allowing sufficient cells to be produced for normal growth (Fish et al., 2006). Even without ZIKV exposure, harmful *ASPM* variants can also lead to primary microcephaly. Thus, inherited *ASPM*-related microcephaly and

Zika-associated microcephaly have different causes, but both can interfere with the growth and division of neural progenitor cells.

Various human studies depict that genetic background may influence the effects of prenatal ZIKV infection. A separate study by Caires-Júnior et al. (2018) studied three pairs of dizygotic, or fraternal, twins in which only one twin developed congenital Zika syndrome. Neural progenitor cells from the affected twins showed greater viral replication and more severe reductions in cell growth than cells from their unaffected siblings. Because the twins shared the same pregnancy but differed genetically, the results suggest that genetic differences may influence cellular susceptibility to ZIKV. Another case-control study compared 40 children with congenital Zika syndrome with 48 ZIKV-exposed children without congenital abnormalities. Variants in the immune-related genes *NOS2* and *TNF* were associated with the risk or severity of congenital Zika outcomes (Gomes et al., 2021). Although these studies did not investigate *ASPM*, they demonstrate that host genetic variants may contribute to differences in fetal outcomes.

Despite this evidence, no study has directly compared *ASPM* variants in ZIKV-exposed fetuses with microcephaly against those in similarly exposed fetuses without congenital abnormalities. A recent scoping review identified several possible maternal and fetal genetic modifiers of prenatal ZIKV-related brain injury but did not establish *ASPM* directly as the susceptible gene (Marques et al., 2025). This matters because ZIKV infection and reduced *ASPM* function affect similar cells and developmental processes, yet their possible interaction remains unclear. Identifying a genetic modifier could help explain why fetal outcomes differ after ZIKV exposure, distinguish inherited microcephaly from Zika-associated microcephaly, and guide future genetic studies. However, current evidence is insufficient to support the use of *ASPM* variants for clinical risk prediction.

This paper examines whether variants in the *ASPM* gene contribute to the reason why only some fetuses exposed to ZIKV develop microcephaly, while others don’t. It argues that *ASPM* is a biologically plausible genetic modifier because *ASPM* dysfunction and ZIKV infection both disrupt neural progenitor-cell division and brain growth. However, the evidence currently available is indirect and does not prove that *ASPM* variants increase susceptibility to congenital Zika-associated microcephaly.

The review compares evidence from human genetic studies, neural-cell experiments, cerebral organoids, and animal models. It first examines how *ASPM* supports neural progenitor-cell division and how ZIKV disrupts similar developmental processes. It then evaluates evidence that genetic background influences how ZIKV affects the fetus. It considers the indirect link between ZIKV and reduced *ASPM* expression and compares this proposed interaction with alternative factors, such as infection timing and placental damage, that could influence susceptibility.

Body

I - Normal *ASPM* Function and Consequences of Disruption

The *ASPM* gene is critical for cell division, especially in the developing brain. The cells that divide are known as neural progenitor cells. These cells divide a limited number of times before developing into neurons and other brain cells. These cells eventually form the cerebral cortex, the brain's outer layer, which is involved in functions such as movement, memory, language, and reasoning (i.e., higher-level thinking). Any decrease in the number of successful progenitor cell divisions can be significant and lead to apoptosis (programmed cell death). To examine *ASPM's* role in this process, a specific study done by Higgins et al. (2010) reduced *ASPM* activity in cultured human cells. The paper originally established that this specific gene is required for normal spindle orientation and for the efficient completion of cytokinesis during cell division. The spindles pull chromosomes apart, and cytokinesis is the final step in which the original cell physically divides into two daughter cells. One cannot occur without the other, which can disrupt cell division. When *ASPM* activity is reduced, the researchers observed abnormalities in spindle organization, orientation, and cytokinesis. The reduction led to frequent failures to complete these processes properly. Their experiments also indicated that the C-terminus — the final section of the *ASPM* protein — is required for *ASPM* to reach the correct location and perform the proper functions. This finding helps explain why mutations that produce shortened *ASPM* proteins may be harmful. While the protein may still be produced, it may not reach the spindle. However, the study used general cultured human cells rather than neural progenitor cells inside the developing brain. It therefore provides direct evidence of *ASPM's* cellular function, but cannot show how those defects affect brain size. To expand on this finding, a 2016 study found that *ASPM* interacts with CITK, another protein involved in cell division, to regulate astral microtubules (Gai et al., 2016). These are specific microtubules that stem from the mitotic spindle and help position it correctly during cell division. In neural progenitor cells, spindle orientation can influence whether they continue dividing or begin differentiating into specialized brain cells. This is what results in an early stop to brain development.

Animal studies connect *ASPM's* cellular function to reduced brain growth and the loss of developing neural cells. One specific study created mice with mutations that resulted in shortened *ASPM* proteins due to a loss of genetic instructions. The mutant mice developed mild microcephaly, linking abnormal *ASPM* to reduced brain size in living organisms (Pulers et al. 2010). A similar study in mice demonstrates that when researchers removed *ASPM* from the developing mouse brain, they observed increased DNA damage and apoptosis in neural progenitor cells. While some apoptosis is normal, excessive apoptosis can reduce the number of cells available to form the cortex (Tonosaki et al., 2023). Overall, the *ASPM* gene supports brain growth by maintaining both accurate division and survival of neural progenitor cells. It helps determine the correct number of cells in a developing cerebral cortex.

II - How ZIKV Damages the Developing Brain

ZIKV can reduce fetal brain growth by interfering with the reproduction and survival of neural progenitor cells. Garcez et al. (2016) exposed human neural stem cells, neurospheres, and

cerebral organoids to ZIKV. Neurospheres and cerebral organoids both originate from neural stem cells and serve as laboratory models of brain tissue development. The infected neurospheres showed abnormal growth, cell detachment, and a risk of reduced survival. More importantly, over the course of 11 days, ZIKV-exposed organoids grew approximately 40% less than uninfected cells.

A similar pattern emerged in another system: Li et al. (2016) studied prenatal ZIKV infection in mouse embryos and found that the infection concentrated in brain regions containing neural progenitor cells, where it was associated with increased apoptosis, disrupted cell division, and thinning of the cerebral cortex. The study further found reduced expression of several genes associated with microcephaly, including *ASPM.* Together, these studies show that ZIKV negatively affects the neural progenitor-cell population required for normal cortical growth.

Further evidence explains how ZIKV disrupts neural progenitor cell division. A 2016 study examined human neuroepithelial stem cells. These are stem cells that are crucial for maintaining and building up the nervous system. They later develop into radial glial cells, which produce neurons important for the cerebral cortex. Along with these stem cells, the study also observed fetal brain tissue maintained in a laboratory and tissue from one fetus with ZIKV-associated microcephaly. After conducting each observation, ZIKV primarily infected radial glial cells in the ventricular and subventricular zones, where progenitor cell division occurs. Following exposure to ZIKV, an activated protein involved in viral immunity and cell division, known as TBK1, moved away from the centrosomes. Centrosomes organize the spindle fibers that separate chromosomes during cell division. Without properly functioning centrosomes, cells may not divide correctly. The infected cells consequently showed abnormal numbers of centrosomes and disrupted mitosis, the process in which a cell divides its chromosomes between two cells (Onorati et al., 2016). A subsequent study by Gabriel et al. (2017) investigated the developmental effects of similar damage in cerebral organoids. The researchers found that ZIKV disrupted centrosome organization and altered the direction of radial glial cell division. Once again, the progenitor cells differentiated into brain cells too early, leaving fewer progenitors available to continue dividing. As a result, the developing brain tissue became thinner. Onorati et al. identify a possible cause of disrupted cell division, while Gabriel et al. show how this disruption can limit brain growth.

III - Genetic Background Influences ZIKV Outcomes

To determine whether genetic background contributes to different outcomes after prenatal ZIKV exposure, Caires-Júnior et al. (2018) studied twins whose outcomes differed despite sharing the same pregnancy. The researchers identified nine twin pairs: seven dizygotic, or fraternal, pairs and two monozygotic, or identical, pairs. The results indicated that six out of seven of the fraternal pairs were discordant, meaning one twin developed congenital ZIKV syndrome while the other did not, whereas both identical twin pairs were concordant because both twins were affected. This pattern suggested a genetic influence because identical twins share nearly all of their DNA, while fraternal twins share about half. However, the number of

twins was too small to establish this relationship on its own; the twin pattern raises the possibility of genetic susceptibility, but it cannot show whether genes directly alter how cells respond to ZIKV. The researchers therefore performed an experiment using cells from three discordant fraternal twin pairs.

Blood cells from each twin were reprogrammed into stem cells and then developed into neural progenitor cells. Before ZIKV exposure, cells from affected and unaffected twins grew at similar rates. However, cells from the affected twin showed significantly reduced growth compared to those from the unaffected twin. The distinction appearing only after infection is important because it supports a gene interaction: the decrease in growth indicated that the affected twins were unable to withstand ZIKV. Ultimately, genetic susceptibility could alter the extent to which fetal neural cells are damaged by infection.

Similarly, Borda et al. (2021) sequenced 40 unrelated children with confirmed prenatal ZIKV exposure: 29 with congenital Zika syndrome and 11 without any related prenatal complications. 10 out of the 29 affected children had at least one harmful variant in a gene linked to the features of congenital Zika syndrome. The researchers also found evidence that changes in two genes, *DISP3* and *IL12RB2*, can reduce the risk of severe effects. This matters because not every child reacted the same when exposed to ZIKV, and genetic differences can help explain why some children developed congenital Zika syndrome while others did not. Additionally, the study examined when the mothers were infected during pregnancy. Infection during the first trimester occurred in 65.5% of those affected with congenital Zika syndrome, but in only 27% of unaffected children. Because infection during this period may increase the risk of birth defects, the researchers included the timing of exposure in their analysis. Fundamentally, Borda et al. (2021) supported the idea that many factors may influence ZIKV outcomes and, more importantly, that one such factor may be genetic differences.

Overall, these studies suggest that a person's biological background may influence how ZIKV affects them. Caires-Júnior et al. (2018) reflected on how differences in biological background can lead to different responses to the same ZIKV infection. Moreover, Borda et al. (2021) found that changes in several genes may also affect the risk of severe outcomes. These findings suggest that *ASPM* could influence ZIKV-related microcephaly.

IV - Indirectly Evaluating the Proposed *ASPM* and ZIKV interaction

The existing evidence makes an *ASPM*-ZIKV interaction plausible because reduced *ASPM* function and ZIKV infection disrupted several of the same developmental processes. Studies examined in Section I showed that major reductions in *ASPM* function can interfere with spindle organization, progenitor cell division, and brain growth and development. Furthermore, studies examined in Section II similarly showed that ZIKV can disrupt the mitotic spindle, alter the direction of progenitor cell division, and cause progenitors to develop early on. These findings are significant because both *ASPM* dysfunction and ZIKV interaction can reduce the progenitor cell population crucial for the construction of the cerebral cortex. Karuppan et al. (2020) provided a more detailed example of how a genetic weakness can alter the effects of

prenatal ZIKV infection. The researchers exposed both groups of mice to ZIKV: mice with normal *BECN1* expression and mice with one working copy of the *BECN1 gene*. These mice produced about 60% less Beclin1 protein. This specific protein helps cells break down cellular components. This process, called autophagy, can affect viral replication, inflammation, and cell growth in response to ZIKV exposure.

About 25% of the ZIKV-exposed mice with reduced *BECN1* expression developed poor growth and undeveloped brains. These effects were most noticeable two to three weeks after birth. This reflects how the same ZIKV exposure caused different outcomes in the two genetic backgrounds. Reduced Beclin1 exacerbated the effects of the infection. This supports the idea that a person's genes can affect the severity of ZIKV-related brain abnormalities and that another developmental gene can also modify the response.

The researchers also measured four genes linked to microcephaly, including *ASPM*. Even without ZIKV, mice with reduced *Becn1* expression presented lower *ASPM* expression than mice with normal *Becn1* expression. The reduction was even more significant after ZIKV exposure, reporting an overall difference of about 2.5- to 3-fold. This suggests a possible link between ZIKV susceptibility and reduced *ASPM* activity.

However, it is important to note that a reduction in *ASPM* expression is not the same as inheriting an *ASPM* variant, since the DNA sequence does not change. Therefore, the study provides indirect support for the *ASPM* hypothesis, but it does not show that inherited *ASPM* variants are responsible for the outcomes. Nevertheless, the strongest conclusion is that *ASPM* is a possible but unproven modifier.

V. Alternative Explanations for Unequal ZIKV Outcomes

While *ASPM*-ZIKV interaction is plausible, differences in infection timing, placental health, and maternal immunity could also explain why exposed fetuses experience unequal outcomes. Coutinho et al. (2021) studied the timing of infection in 511 women with confirmed ZIKV infection, representing 513 exposed fetuses. 35.9% of fetuses exposed during the first trimester faced adverse outcomes, compared with 2.8% exposed during the second trimester and 1.9% exposed during the third trimester. These outcomes included pregnancy loss, microcephaly, and other brain abnormalities.

The researchers found that infection before 11 completed weeks of pregnancy was linked to a 14-fold higher risk of an adverse outcome compared to later in the pregnancy. This explains that fetuses with similar genetic backgrounds may have different outcomes simply because infection occurs at different stages of development. Early in pregnancy, brain-forming cells multiply rapidly, so disruption during this period may have a greater effect on brain growth.

Similarly, Miner et al. (2016) provided a possible explanation for how prenatal ZIKV infection can affect fetal development through the placenta. The study examined two mouse models and found that ZIKV infected trophoblasts and fetal blood-vessel cells in the placenta. Trophoblasts are placental cells that help form the barrier between the mother's and the fetus's

blood. The infection led to cell death, damage to placental blood vessels, and thinning of the placental layer that exchanges materials between the mother and fetus.

The placenta normally provides the fetus with oxygen and proper nutrients while helping protect it from infection. Damage to its cells and blood vessels can reduce blood flow, weaken the protective barrier, and limit fetal growth. Additionally, the study found that placental infection was linked to restricted growth inside the uterus and could lead to fetal loss. The researchers also detected ZIKV in fetal tissue, which can actually affect brain development directly by entering fetal brain tissue and indirectly by damaging the placenta. Therefore, reduced brain growth does not require an *ASPM* variant. Despite there being other alternative factors that could affect prenatal ZIKV outcomes, these two studies are especially important. Coutinho et al. (2021) show that early infection is strongly linked to adverse outcomes, while Miner et al. (2016) show how placental damage can impair fetal growth.

Discussion

Although current studies provide no evidence of a direct association between *ASPM* variants and congenital Zika-associated microcephaly, the findings make such a relationship biologically plausible. *ASPM* supports spindle organization and neural progenitor-cell division, whereas ZIKV disrupts centrosomes, mitotic spindles, and neural progenitor cells. Both factors can therefore reduce the number of progenitor cells required for the proper development of the cerebral cortex. This overlap suggests that a mild *ASPM* variant could leave these cells less able to withstand ZIKV-related damage.

The genetic-background studies further support this possibility. Caires-Júnior et al. (2018) found that cells from affected and unaffected twins grew similarly before infection, but cell growth in affected twins decreased after infection. This suggests that the genetic difference between the twins, specifically fraternal twins, increases vulnerability only when infection occurs. Borda et al. (2021) also found evidence that several developmental and immune-related genes may influence ZIKV outcomes. However, neither study identified *ASPM*, but together they emphasize that genetic background can influence susceptibility.

Nevertheless, Karuppan et al. (2020) provided the closest connection to *ASPM*. Approximately 25% of ZIKV-exposed offspring with reduced *BECN1* expression developed smaller, more underdeveloped brains, while several microcephaly genes, including *ASPM*, showed reduced expression. This corroborates the idea that weakening a host gene can worsen ZIKV-related effects and affect the *ASPM* gene. However, reduced *ASPM* expression is not the same as inheriting an *ASPM* variant. The study therefore provides indirect support for the hypothesis rather than evidence that *ASPM* is what specifically caused the abnormal outcome.

Because this evidence remains indirect, the possible contribution of *ASPM* is more likely to be considered alongside other factors. Coutinho et al. (2021) found that infection before 11 completed weeks of pregnancy was associated with a 14-fold higher risk of an adverse outcome, while another study done by Miner et al. (2016) showed that ZIKV can damage the placenta and restrict fetal growth. These findings provide more direct evidence for infection timing and

placental damage than currently exists for *ASPM*. However, together, these explanations suggest that the *ASPM* gene could act as a modifier in a multi-step process.

Moreover, the research being compared currently has animal and human models, and does not directly identify *ASPM* as a genetic modifier. However, future research should first compare a large group of ZIKV-exposed children with microcephaly to exposed children without abnormalities. A genome-wide association study (GWAS) could first identify common genetic differences connected to significant ZIKV outcomes. Researchers could then use genome sequencing to read almost all of the DNA of affected and unaffected children and their parents. Studying the parents would show whether a child inherited an *ASPM* variant from one or both of their parents. This is important because *ASPM*-related effects may be recessive, meaning a child may need harmful variants in both copies of the gene to develop microcephaly. Additionally, this method could identify rare variants and changes in regulatory DNA that control the strength of *ASPM* activation and may be missed by GWAS. Researchers can then test whether these variants are more common in ZIKV-exposed children with microcephaly than in ZIKV-exposed children without abnormalities such as microcephaly. The groups should be matched for ancestry and any other factors that contribute to ZIKV severity.

Researchers could create two groups of neural progenitor cells with the same genetic background, except that one group would carry an *ASPM* variant while the other would have normal *ASPM*. Each group would then be divided again so that half of the cells are exposed to ZIKV and half are not. A comparison can be made based on how quickly the cells divide, how many survive, and how much the groups grow. If the variant cells developed normally without ZIKV but showed greater damage after infection than the normal cells, this would support an interaction between the variant and the virus.

Conclusion

This review examined whether *ASPM* variants could help explain why only some fetuses exposed to ZIKV develop microcephaly. The evidence shows that *ASPM* dysfunction and ZIKV infection affect many of the same developmental processes. Human studies show that genetic background can influence responses to ZIKV, and animal research provides indirect evidence linking ZIKV susceptibility to reduced *ASPM* expression. However, current evidence doesn't examine *ASPM* variants and instead focuses on alternative factors, such as infection timing, that can influence ZIKV susceptibility. The available research therefore supports *ASPM* as a biologically plausible but unconfirmed modifier.

Distinguishing inherited microcephaly from Zika-associated microcephaly matters because similar outcomes may result from different causes. Harmful variants in either copy of *ASPM* may independently cause inherited primary microcephaly, whereas a milder variant might only increase vulnerability when combined with ZIKV. Understanding this difference is what allows us to further understand the topic of microcephaly. Clarifying the relationship would support more accurate diagnoses, improve the interpretation of sequencing results, and guide research into why similarly exposed fetuses experience different outcomes.

Work Cited

Borda, Victor, et al. "Whole-Exome Sequencing Reveals Insights into Genetic Susceptibility to Congenital Zika Syndrome." PLOS Neglected Tropical Diseases, edited by Helene Dutartre, vol. 15, no. 6, June 2021, p. e0009507. DOI.org (Crossref), https://doi.org/10.1371/journal.pntd.0009507.

Caires-Júnior, Luiz Carlos, et al. "Discordant Congenital Zika Syndrome Twins Show Differential in Vitro Viral Susceptibility of Neural Progenitor Cells." Nature Communications, vol. 9, no. 1, Feb. 2018, p. 475. DOI.org (Crossref), https://doi.org/10.1038/s41467-017-02790-9.

Coutinho, Cm, et al. "Early Maternal Zika Infection Predicts Severe Neonatal Neurological Damage: Results from the Prospective Natural History of Zika Virus Infection in Gestation Cohort Study." BJOG: An International Journal of Obstetrics & Gynaecology, vol. 128, no. 2, Jan. 2021, pp. 317–26. DOI.org (Crossref), https://doi.org/10.1111/1471-0528.16490.

Fish, Jennifer L., et al. "Aspm Specifically Maintains Symmetric Proliferative Divisions of Neuroepithelial Cells." Proceedings of the National Academy of Sciences, vol. 103, no. 27, July 2006, pp. 10438–43. DOI.org (Crossref), https://doi.org/10.1073/pnas.0604066103.

Gabriel, Elke, et al. "Recent Zika Virus Isolates Induce Premature Differentiation of Neural Progenitors in Human Brain Organoids." Cell Stem Cell, vol. 20, no. 3, Mar. 2017, pp. 397-406.e5. DOI.org (Crossref), https://doi.org/10.1016/j.stem.2016.12.005.

Gai, Marta, et al. "ASPM and CITK Regulate Spindle Orientation by Affecting the Dynamics of Astral Microtubules." The EMBO Reports, vol. 17, no. 10, Oct. 2016, pp. 1396–409. DOI.org (Crossref), https://doi.org/10.15252/embr.201541823.

Garcez, Patricia P., et al. "Zika Virus Impairs Growth in Human Neurospheres and Brain Organoids." Science, vol. 352, no. 6287, May 2016, pp. 816–18. DOI.org (Crossref), https://doi.org/10.1126/science.aaf6116.

Gomes, Julia A., et al. "Association between Genetic Variants in NOS2 and TNF Genes with Congenital Zika Syndrome and Severe Microcephaly." Viruses, vol. 13, no. 2, Feb. 2021, p. 325. DOI.org (Crossref), https://doi.org/10.3390/v13020325.

Higgins, Julie, et al. "Human ASPM Participates in Spindle Organisation, Spindle Orientation and Cytokinesis." BMC Cell Biology, vol. 11, no. 1, Dec. 2010, p. 85. DOI.org (Crossref), https://doi.org/10.1186/1471-2121-11-85.

Karuppan, Mohan Kumar Muthu, et al. "Reduced-Beclin1-Expressing Mice Infected with Zika-R103451 and Viral-Associated Pathology during Pregnancy." Viruses, vol. 12, no. 6, June 2020, p. 608. DOI.org (Crossref), https://doi.org/10.3390/v12060608.

Li, Cui, et al. "Zika Virus Disrupts Neural Progenitor Development and Leads to Microcephaly in Mice." Cell Stem Cell, vol. 19, no. 1, July 2016, pp. 120–26. DOI.org (Crossref), https://doi.org/10.1016/j.stem.2016.04.017.

Marques, Fernanda J. P., et al. “Genetic Modifiers of Prenatal Brain Injury After Zika Virus Infection: A Scoping Review.” Genetic and Genomic Medicine, 2 Jan. 2025. DOI.org (Crossref), https://doi.org/10.1101/2025.01.02.25319896.

Miner, Jonathan J., et al. “Zika Virus Infection during Pregnancy in Mice Causes Placental Damage and Fetal Demise.” Cell, vol. 165, no. 5, May 2016, pp. 1081–91. DOI.org (Crossref), https://doi.org/10.1016/j.cell.2016.05.008.

Onorati, Marco, et al. “Zika Virus Disrupts Phospho-TBK1 Localization and Mitosis in Human Neuroepithelial Stem Cells and Radial Glia.” Cell Reports, vol. 16, no. 10, Sept. 2016, pp. 2576–92. DOI.org (Crossref), https://doi.org/10.1016/j.celrep.2016.08.038.

Pulvers, Jeremy N., et al. “Mutations in Mouse Aspm (Abnormal Spindle-like Microcephaly Associated) Cause Not Only Microcephaly but Also Major Defects in the Germline.” Proceedings of the National Academy of Sciences, vol. 107, no. 38, Sept. 2010, pp. 16595–600. DOI.org (Crossref), https://doi.org/10.1073/pnas.1010494107.

Tonosaki, Madoka, et al. “Loss of Aspm Causes Increased Apoptosis of Developing Neural Cells during Mouse Cerebral Corticogenesis.” PLOS ONE, edited by Carlos Oliva, vol. 18, no. 11, Nov. 2023, p. e0294893. DOI.org (Crossref), https://doi.org/10.1371/journal.pone.0294893.

Wu, Kong-Yan, et al. “Vertical Transmission of Zika Virus Targeting the Radial Glial Cells Affects Cortex Development of Offspring Mice.” Cell Research, vol. 26, no. 6, June 2016, pp. 645–54. DOI.org (Crossref), https://doi.org/10.1038/cr.2016.58.

Astronomical Technologies and Cosmology in Ancient Greece, China, Egypt, and the Maya
By Maya Fitzgerald

Abstract

This paper examines how astronomical technologies shaped the ways ancient civilizations observed, organized, and interpreted the sky. Focusing on Ancient Greece, Ancient China, Ancient Egypt, and the Maya, this study uses a comparative qualitative framework to analyze three categories: key astronomical technology, practical function, and cultural significance. This structure allows the civilizations and their methods of transforming celestial observation into cultural meaning to be compared rather than ranked by perceived scientific advancement. The Greek examples examined here reflect a view that celestial motions, and by extension, the cosmos, were predictable and could be understood through mathematics. Chinese star catalogs and calendrical systems connected celestial observation to governance, social order, and agricultural timing. Egyptian solar and stellar alignments connected sacred architecture with cosmic and divine order, while Maya calendrical systems linked celestial cycles to ritual timing, agriculture, and aspects of political and religious authority. Together, these case studies show that civilizations created distinct forms of order by embedding celestial interpretation within cultural practices and institutions. The results of this study suggest that celestial phenomena were not only objects of observation, but also sources of identity, authority, spirituality, and cultural order.

Introduction

For millennia, humanity has gazed upon the sky. From using the positioning of celestial bodies for orientation, timing, and meaning, to creating stories from the patterns within the stars, the night sky has always had an impact on human culture. Narratives created by early civilizations based on what was learned from the night sky grew into vital aspects of a culture's identity. Astronomy played a central role in shaping systems of timekeeping, governance, and religious practice across multiple ancient civilizations as technologies and methods, such as the examples used in this paper, were developed to observe and interpret celestial phenomena. Through instruments, calendars, and architectural alignments, ancient astronomers transformed observations of the sky into systems of meaning that structured daily life and belief. Historians study the significance of these technologies of ancient civilizations to understand the purpose and effects of astronomy in different cultures.

Accounts of the development of Western astronomy often emphasize ancient Greek contributions. Scholarship often views the Greeks as significant cultural precursors to the understanding of modern Western astronomy. Their astronomical traditions included both mythological interpretations of constellations and increasingly systematic observations and mathematical interpretations of the sky (Ferreira et al. 2). Greek thinkers such as Aristotle and Ptolemy contributed significantly to astronomical theory, cosmology, and mathematical modeling (Ferreira et al. 1–2). However, astronomical knowledge also developed across many other regions and cultures, and Ferreira et al. emphasize this broader history by identifying the

"Sumerians, Babylonians, Hindus, [...] and several indigenous cultures" as contributors to early astronomical knowledge (Ferreira et al. 1–2). This wider context demonstrates that sophisticated astronomical practices were not confined to Greece. The present study focuses on Ancient China, Ancient Egypt, and the Maya alongside Ancient Greece to compare how distinct cultural contexts shaped the development of different astronomical technologies.

This paper approaches ancient astronomy through a technological lens to understand how astronomical technology shaped how civilizations observed, organized, and culturally interpreted the sky. In this paper, "technology" refers to tools and any other method used to measure, map, or align with celestial bodies, including star charts, mechanical devices, calendars, observatories, and architectural alignments. Using this definition of technology provides a more comprehensive understanding of what is being compared across cultures. This paper examines how different astronomical technologies mediated relationships between human societies and the sky, and how these relationships shaped cultural interpretation of celestial phenomena–through political, mathematical, and spiritual frameworks. This theoretical framing provides a basis for comparing how different cultures used technology through different forms of observational, mathematical, architectural, and calendrical practices to observe and interpret the sky.

This paper examines how astronomical technologies structured the ways ancient civilizations used and interpreted celestial phenomena. Using a comparative framework, it evaluates each civilization through three categories: key astronomical technology, practical function, and cultural significance. The analysis focuses on Ancient Greece, Ancient China, Ancient Egypt, and the Maya, comparing how each society used tools, observational methods, calendars, or architectural alignments to organize celestial knowledge. By placing these case studies in conversation with one another, this paper shows that astronomical technologies did more than support observation. They also shaped distinct worldviews by connecting the sky to mathematics, governance, sacred architecture, ritual practice, and cultural identity.

Methods

This paper uses comparative and qualitative research to examine how different ancient civilizations developed and used technologies to interpret the stars. The selected civilizations are compared through the same categories. Rather than conducting new observational or experimental measurements, this study draws on existing knowledge relating to the historical, archaeological, and cultural astronomy of the selected civilizations. The purpose of this approach is to understand how astronomical tools shaped different civilizations' interpretations of the sky, rather than to rank civilizations by perceived levels of advancement. This method emphasizes the importance of thorough interpretation, context, and the relationship between astronomical technology and meaning within each society (McCluskey 27–28).

The research for this paper draws on peer-reviewed academic journal articles, archaeological site studies, museum or institutional resources, academic databases, and modern scholarly interpretations of ancient astronomy. These sources provide information about the function and potential significance of astronomical technologies such as mechanical devices, star

catalogs, calendar systems, architectural alignments, and observational devices. Popular and institutional sources are used mainly for background context, while peer-reviewed and scholarly sources support the paper's main historical and archaeological claims. Secondary sources are also used to interpret how these technologies influenced various cultures. These sources help interpret the cultural meanings of tools when surviving evidence does not explicitly identify their original functions. The significance of these technologies can be seen across areas of cultural life, including religion, agriculture, politics, and storytelling. Sources were selected because of their direct discussion of astronomical technology, methods of observation, and cultural interpretation.

To compare these civilizations fairly and consistently, this study applies a structured analytical framework. This structure uses key astronomical technology, practical function, and cultural significance to analyze the four selected civilizations. Using identical categories for each case study helps ensure that the cases are analyzed consistently. This framework makes the comparison purposeful by focusing on the distinct contributions and interpretive systems developed by each civilization.

This study focuses on Ancient Greece, Ancient China, Ancient Egypt, and the Maya because the selected examples from each civilization illustrate distinct approaches to astronomical technology and cosmological interpretation that were revealed after preliminary research. Examples were selected for their strong documentation in scholarly literature and because they provide variation in technological form, practical function, geography, and cultural context. The Greek examples emphasize a mathematical, mechanical, and analytical approach to astronomy. The Chinese examples emphasize a cataloging, calendrical, and state-centered approach to astronomy. The Egyptian examples emphasize an architectural, solar, stellar, and sacred orientation approach to astronomical technology. The Maya examples emphasize a calendar-based, ritual, and agricultural approach to using astronomical technology to organize important events within calendrical systems such as ritual timing, agricultural planning, and ceremonial practice. Together, these case studies create a broader cross-cultural comparison of astronomical traditions.

While each of these civilizations developed extensive astronomical traditions, this paper focuses specifically on the relationships between technological tools and cultural interpretation. A limitation of that focus is that the analysis relies heavily on modern scholarly interpretations of surviving archaeological and textual evidence, which is incomplete, and some cultural meanings remain debated (Ruggles 368). This paper does not attempt to provide an exhaustive account of astronomical traditions within the selected cultures. The examples are illustrative of selected astronomical traditions from different historical periods and different regions and are not intended to represent each civilization as culturally or technologically static. This selective approach allows a focused comparison of how particular astronomical technologies functioned within specific cultural contexts.

Results

Ancient Greece

The Greek astronomical technologies examined here emphasize mathematical and philosophical approaches to celestial order. Greek astronomical practices included instruments such as the Antikythera Mechanism, a complex mechanical astronomical device (Moussas 133). This device used gears to calculate celestial cycles and predict astronomical events, information that could then support calendrical planning (Freeth et al. 12–3). The mathematical principles embodied in the mechanism reflect broader Greek philosophical ideas about harmony and order (Moussas 141). This evidence suggests a conception of the cosmos as predictable, ordered, and mathematically structured.

Claudius Ptolemy, a Greek scholar, developed mathematical models for describing and predicting celestial motions in his work, the *Almagest*. (Sidoli 17). Specifically, the *Almagest* used geometrical models and tables to calculate and predict the apparent motions of celestial bodies (Beck 7). The mathematical models and predictive systems in the *Almagest* can be understood as technologies of interpretation because they organized celestial motion into a usable system. Ptolemy treated mathematics as a branch of theoretical philosophy best suited to explaining celestial motion (Beck 6). The mapping of the stars, specifically represented by Ptolemy's work, reflects the larger significance, within the Greek astronomical tradition examined here, to search for patterns and meaning within the movements of celestial bodies (Hannah 49). The selected Greek astronomical technologies reflect a conception of the cosmos as orderly, predictable, and mathematically intelligible.

Ancient China

In the court traditions examined here, astronomy was closely connected to governance and calendrical regulation. Chinese court astronomers observed stellar positions as far back as 104 BCE (Ahn 1). Marked reference stars were used by court astronomers to calculate the positions of other celestial bodies in order to create calendar systems (Ahn 16). Astronomy and astrology were deeply connected with state affairs; court astronomers believed celestial events had significant societal meanings, thus connecting their society to the night sky (Sun 103). In the *Star Manual of Master Shi*, constellations are grouped with depictions of an imperial court, including objects such as an emperor's chariot and a celestial marketplace, projecting elements of imperial and social organization onto the celestial sphere (Sokol 343). This suggests that Ancient Chinese astronomical systems did more than record stars; they also interpreted the cosmos to reflect social order and governance (Sun 98-99).

The technologies examined here supported overlapping goals involving governance, calendrical regulation, and celestial observation. Court astronomers also developed complex observational instruments, including armillary spheres (Lin and Hsiao 1). The armillary sphere could be used to calculate the positions of celestial bodies. Some versions, notably the Taiping armillary sphere built by Zhang Sixun during the Northern Song Dynasty, were designed with indicators for telling time (Hwang et al. 218). Agriculture made calendrical and seasonal regulation practically important in Chinese society (Shi 2031–32). Celestial phenomena could

also be interpreted as politically and socially meaningful signs connected to Heaven (Sun 103). Chinese astronomical technologies connected the sky to state affairs (Sun 103), seasonal regulation (Shi 2032), and the organization of society (Sun 104).

Ancient Egypt

In Ancient Egypt, solar and stellar alignment played a large role in the design of monumental architecture. Temples were oriented within the local landscape of the earth and the sky, and included patterns relating to the stations of the sun, stars, and cardinal directions (Belmonte et al., "On the Orientation" 65). The Amun temple at Karnak provides a notable example of intentional orientation to the sky; it is believed to have been oriented toward the winter-solstice sunrise (Belmonte and González-García 5). Karnak is located in Thebes, a city that served as an important religious capital associated with the deity Amun-Re, who was known to represent "the hidden power of the solar divinity" (Belmonte and González-García 3). The specific orientation of this temple connected the architecture to a larger solar symbolism within the religious context of Karnak (Belmonte and González-García 7). Scholars have interpreted the orientation of the Amun temple to reflect the outcome of using astronomy to create temples that embedded solar symbolism within religious architecture (Belmonte and González-García 7).

Spence argues that northern or circumpolar stellar observations may have been used to establish the alignment and orientation of certain Egyptian pyramids (Spence 321). Certain stars and constellations known as the *Imperishable Stars* served a significant function in Ancient Egyptian culture relating to the processes of death and the afterlife as they were "frequently mentioned as celestial destinies for the soul of the king after death" (Belmonte et al., "Astronomy, Landscape and Symbolism" 245). *Maat* was a concept in Egyptian religious, ethical, and cosmological thought, representing truth, balance, and order (Ferguson 2–4). *Maat* provides a cultural framework for understanding the importance of regularity and order in Egyptian religious and social thought (Ferguson 74–6). Taken together, these sources allow the architectural alignments to be considered within a broader Egyptian cultural emphasis on cosmic regularity and order.

The Maya

The Maya developed extensively documented astronomical and calendrical systems in the ancient world, integrating celestial observation into religion, agriculture, and political authority. A cosmogram represents the order of the universe and time (Inomata et al. 11). Aguada Fénix in southeastern Mesoamerica has been interpreted by scholars to be designed to function as a cosmogram that embodied calendrical concepts and ideas of cosmic order. The site served as a location for ceremonies, feasts, and a materialization of the order of the universe (Inomata et al. 1). Inomata et al. suggest that "calendrical specialists likely lived at Aguada Fénix year-round to conduct astronomical observations from fixed locations" (Inomata et al. 12). Inomata et al. do not interpret these specialists as exercising coercive authority, but suggest that specialized knowledge may have provided social influence (Inomata et al. 12). The importance of

cosmograms and calendars for ritual and political influence suggests a possible form of social influence rooted in specialized ritual and astronomical knowledge.

Maya architectural orientations corresponded to astronomically and calendrically significant dates and intervals, linking built space with recurring celestial and calendrical cycles (Šprajc and Sánchez Nava 320). These calendrical systems facilitated the scheduling of agricultural and religious events (Šprajc 17). Milbrath argues that the 260-day Tzolkin system may have developed in relation to agricultural cycles, with the 260-day interval corresponding to the Maya agrarian year (Milbrath 94). The ability to identify recurring calendrical dates was especially significant for religious purposes, as ritual planning depended on the regularity of the cyclical calendar (Šprajc et al. 1). Time-keeping was a valuable tool for authority figures; Rice argues that Maya political power was partly "cosmo-political, based on esoteric knowledge strategically deployed to create illusions of control of time and the cosmos" (Rice 294). This reflects how specialized calendrical knowledge had both practical function and political or ritual significance in Maya culture. These examples suggest a conception of time as sacred, cyclical, and socially organized.

Discussion

When viewed together, the case studies show how each of the four selected civilizations used technology to understand the sky and create order, but each defined order differently. The Greek examples emphasize finding mathematical order and predictive reasoning within their mathematical models of celestial motion (Hannah 49). In contrast, the Maya examples emphasize the use of their understanding of solar observation and calendrical cycles to guide their rituals and other cultural events (Šprajc 17). Examples from Ancient China and Ancient Egypt both connected the night sky to earthly order, but they did so in different ways: Chinese astronomy supported political (Sun 103-5) and calendrical regulation (Shi 2032), while Egyptian astronomy aligned sacred architecture with cosmic and divine order to emphasize solar symbolism (Belmonte and González-García 7). These differences in astronomical practice show how the same sky can be interpreted in many ways depending on cultural priorities.

These comparisons illustrate that although astronomical technology largely aided observation of the sky in each of these civilizations, this was not the sole impact. Astronomical technologies did more than enable observation of the sky; they transformed celestial motion into systems of cultural meaning. In different contexts, similar goals such as prediction and timekeeping could acquire different cultural significance. The Antikythera Mechanism embodied a conception of celestial motion as predictable and ordered (Moussas 141). Architectural alignment embedded celestial observation within sacred space and religious practice (Belmonte and González-García 7). Calendrical systems were used for regulation and organization: in China, calendrical systems supported governance (Sun 103) and seasonal regulation (Shi 2032), while the Maya structured ritual, agricultural (Šprajc 17), and political (Rice 294) time through calendars. The practical functions made possible by astronomical technology allowed civilizations to interpret the night sky in different cultural, political, and religious ways.

Taking a step back to view the bigger picture, studying ancient astronomical technologies helps current scholars understand how ancient human societies created meaning from the sky. Over thousands of years, different civilizations have looked upon the same sky and developed systems for understanding the vastness of the night sky. There is no singular linear history of astronomy: different systems produce different interpretations of the sky, showing how astronomy and cosmology are shaped not just by observational practice, but by culture, geography, politics, religion, and social needs. The study of ancient astronomy is important for more than just technical understanding. To understand why ancient astronomy matters, it is necessary to understand how civilizations differently embedded celestial interpretation into systems of time, power, ritual, and cultural identity.

Conclusion

This paper has shown how astronomical technology shaped not only how ancient civilizations observed the sky, but also how they interpreted its meaning. Across the four cases – Ancient Greece, Ancient China, Ancient Egypt, and the Maya – astronomical technologies transformed celestial observation into culturally specific systems of knowledge and order. Each civilization used its own tools and methods to create meaning from the same sky; however, the meanings associated with those technologies differed according to cultural priorities.

In Greece, prediction was associated with mathematical modeling of celestial motion, while in China it supported governance, calendrical regulation, and seasonal organization. In both Egypt and the Maya world, celestial observation was important for sacred time and ritual practice, and astronomical technologies were closely tied to the relationship between human society and cosmic order. Together, these examples show that different cultural needs shaped astronomical practice as technologies functioned as a bridge between observation, practical use, and cultural interpretation.

These findings challenge the idea that the history of astronomy is linear and can be understood through a single cultural tradition. In reality, ancient astronomy developed in many places, with each culture shaping astronomical practice to serve local needs, beliefs, landscapes, and technologies. The night sky has remained a shared human experience for millennia, but each civilization interpreted it through its own values and views. Ultimately, studying ancient astronomy helps us to understand how humans create meaning from the natural world. These findings suggest that celestial phenomena were not only objects of observation but also sources of identity, authority, spirituality, and cultural order.

Works Cited

Ahn, Sang-Hyeon. "Revisiting the Epoch of the Earliest Chinese Star Catalog Titled 'Shi Shi Xing Jing.'" Publications of the Astronomical Society of Japan, vol. 72, no. 5, 2020, article 87, https://doi.org/10.1093/pasj/psaa080.

Beck, Roger. A Brief History of Ancient Astrology. Blackwell Pub, 2007.

Belmonte, Juan, and A. César González-García. "The Pillars of the Earth and the Sky: Capital Cities, Astronomy and Landscape." Journal of Skyscape Archaeology, vol. 1, no. 1, 2015, pp. 8–37, https://doi.org/10.1558/jsa.v1i1.26952.

Belmonte, Juan Antonio, et al. "Astronomy, Landscape and Symbolism: A Study of the Orientation of Ancient Egyptian Temples." In Search of Cosmic Order: Selected Essays on Egyptian Archaeoastronomy, edited by Juan Antonio Belmonte and Mosalam Shaltout, Supreme Council of Antiquities Press, 2009, pp. 212–84.

Belmonte, Juan Antonio, et al. "On the Orientation of Ancient Egyptian Temples: (5) Testing the Theory in Middle Egypt and Sudan." Journal for the History of Astronomy, vol. 41, no. 1, Feb. 2010, pp. 65–93, https://doi.org/10.1177/002182861004100103.

Ferguson, R. James. The Ancient Egyptian Concept of Maat: Reflections on Social Justice and Natural Order. Research Paper Series: Centre for East-West Cultural & Economic Studies, no. 15, Bond University, 2016, research.bond.edu.au/en/publications/the-ancient-egyptian-concept-of-maat-reflections-on-social-justic/.

Ferreira, Marcello, et al. "Greek Astromythology: Intersections Between Mythology History and Modern Astronomy Education." Frontiers in Education, vol. 10, Feb. 2025, https://doi.org/10.3389/feduc.2025.1431336.

Freeth, Tony, et al. "A Model of the Cosmos in the Ancient Greek Antikythera Mechanism." Scientific Reports, vol. 11, 2021, article 5821, https://doi.org/10.1038/s41598-021-84310-w.

Hannah, Robert. "The Roles of Observational Astronomy in Ancient Greece." Scientific Culture, vol. 1, no. 2, 2015, pp. 47–56, sci-cult.com/wp-content/uploads/2018/04/5-%CE%97%CE%91%CE%9D%CE%9D%CE%91%CE%97_SC_formatted_RH_new2-copy.pdf.

Hwang, Zheng-Hui, et al. "On the Study of Ancient Chinese Mechanical Astronomical Clocks." Journal of the Chinese Society of Mechanical Engineers, vol. 43, no. 3, 2022, pp. 217–28, journal.csme.org.tw/download/On-the-Study-of-Ancient-Chinese-Mechanical--Astronomical-Clocks.pdf.

Inomata, Takeshi, et al. "Landscape-Wide Cosmogram Built by the Early Community of Aguada Fénix in Southeastern Mesoamerica." Science Advances, vol. 11, no. 45, 2025, article eaea2037, https://doi.org/10.1126/sciadv.aea2037.

Lin, Jian-Liang, and Kuo-Hung Hsiao. "Comparison of the Armillary Sphere in Ancient China and Western World." Engineering Proceedings, vol. 89, no. 1, 2025, article 15, https://doi.org/10.3390/engproc2025089015.

McCluskey, Stephen C. “The Cultures of Archaeoastronomy and the History of Science.” Proceedings of the International Astronomical Union, vol. 7, no. S278, 2011, pp. 19–29, https://doi.org/10.1017/S1743921311012439.

Milbrath, Susan. “The Role of Solar Observations in Developing the Preclassic Maya Calendar.” Latin American Antiquity, vol. 28, no. 1, Mar. 2017, pp. 88–104, https://doi.org/10.1017/laq.2016.4.

Moussas, Xenophon. “Early Greek Astrophysics: The Foundations of Modern Science and Technology.” American Journal of Space Science, vol. 1, no. 2, 2013, pp. 129–44, https://doi.org/10.3844/ajssp.2013.129.144.

Rice, Prudence M. “Time, Power, and the Maya.” Latin American Antiquity, vol. 19, no. 3, 2008, pp. 275–98. JSTOR, https://doi.org/10.2307/25478231.

Ruggles, Clive L. N. “Nature and Analysis of Material Evidence Relevant to Archaeoastronomy.” Handbook of Archaeoastronomy and Ethnoastronomy, edited by Clive L. N. Ruggles, Springer, 2015, pp. 353–72.

Shi, Yunli. “Ancient Chinese Astronomy: An Overview.” Handbook of Archaeoastronomy and Ethnoastronomy, edited by Clive L. N. Ruggles, Springer, 2014, pp. 2031–42, https://doi.org/10.1007/978-1-4614-6141-8_210.

Sidoli, Nathan. “Mathematical Tables in Ptolemy’s Almagest.” Historia Mathematica, vol. 41, no. 1, Feb. 2014, pp. 13–37, https://doi.org/10.1016/j.hm.2013.10.004.

Sokol, Joshua. “Chinese Star Catalog Is the World’s Oldest, Astronomers Claim.” Science, 18 Apr. 2025, www.science.org/content/article/china-lays-claim-world-s-oldest-surviving-star-catalog.

Spence, Kate. “Ancient Egyptian Chronology and the Astronomical Orientation of Pyramids.” Nature, vol. 408, no. 6810, 16 Nov. 2000, pp. 320–24, https://doi.org/10.1038/35042510.

Šprajc, Ivan. “Astronomical Aspects of Group E-Type Complexes and Implications for Understanding Ancient Maya Architecture and Urban Planning.” PLOS ONE, vol. 16, no. 4, 2021, article e0250785, https://doi.org/10.1371/journal.pone.0250785.

Šprajc, Ivan, and Pedro Francisco Sánchez Nava. “Equinoxes in Mesoamerican Architectural Alignments: Prehispanic Reality or Modern Myth?” Ancient Cosmologies and Modern Prophets: Proceedings of the 20th Conference of the European Society for Astronomy in Culture, edited by Ivan Šprajc and Peter Pehani, Slovene Anthropological Society, 2013, pp. 319–37.

Šprajc, Ivan, et al. “Origins of Mesoamerican Astronomy and Calendar: Evidence from the Olmec and Maya Regions.” Science Advances, vol. 9, no. 1, 6 Jan. 2023, article eabq7675, https://doi.org/10.1126/sciadv.abq7675.

Sun, Xiaochun. “Connecting Heaven and Man: The Role of Astronomy in Ancient Chinese Society and Culture.” Proceedings of the International Astronomical Union, vol. 5, no. S260, Jan. 2009, pp. 98–106, https://doi.org/10.1017/S1743921311002195.

Self-Reconfigurable Modular Robots for Ocean Exploration: Applications of Adaptive Robotic Systems By Xavier Song

Abstract

Ocean exploration is considered a challenging field of research, as deep underwater environments are often perilous and unpredictable. Current robotic systems such as AUVs and ROVs have enabled significant advances in ocean exploration yet there are still many limitations in exploring its changing environment. Other studies have examined the mechanisms underlying self-reconfiguring modular robots as well as the various factors in ocean exploration. Robotic systems such as the M-TRAN provide insight into modular robotic applications and answer questions of feasibility. Self-reconfigurable modular robots have the potential to improve ocean exploration by introducing adaptable systems capable of changing their structures to meet different environmental challenges. However, limitations in hardware complexity, communication, and reliability must be addressed before these systems can be widely implemented. An understanding of how modular robotics can be applied to ocean exploration may contribute to the development of more adaptable and efficient marine technologies. This research connects advancements in robotics with the growing need for accessing and discovering the depths of Earth's oceans.

Introduction

Human beings naturally have a desire for knowledge and exploration. Key examples are Earth's oceans, with their waters having been explored to a very small degree (NOAA, "How Many Species"). Oceans contain many valuable resources and ecosystems that remain largely undiscovered (NOAA, "How Many Species"). Ocean exploration has proven to be a substantial challenge as extreme pressure, and large depths limit human exploration efforts as well as deeper oceans featuring vast unpredictable environments (NOAA Ocean Exploration, "What Is the 'Deep' Ocean?"). Together, these conditions make the ocean depths largely unsuitable for humans, illustrating how technology is necessary for further exploration. Current technological implementations such as AUVs (Autonomous Underwater Vehicles) and ROVs (Remotely Operated Vehicles) allow scientists to collect data and explore regions humans cannot easily reach (NOAA Ocean Exploration, "Autonomous Underwater Vehicles"; "Remotely Operated Vehicles"). Remotely operated data collection allows for longer ocean expeditions and a larger focus on data rather than human life-support requirements. However, these robots have a major flaw being their fixed designs. Conventional technologies cannot physically adapt to changing, hazardous environments, making way for alternate explorative technologies.

To achieve this enhancement of ocean exploration efforts modular robots are proposed. Modular robots are robots defined by a series of connecting modules featuring individual parts working together (Yim et al.). Self-reconfiguration is the ability to change its shape or configuration in order to adapt to various tasks (Yim et al.). The applications of this can be seen in ocean exploration. Oceans have many narrow and tight spaces as well as oddly shaped curved

spaces which are difficult to navigate regularly (Watson et al.). However, a self-reconfiguring robot could shape itself to adapt to these spaces and successfully navigate its environment (Seo et al.). This paper will examine scholarly literature in modular robotics and ocean exploration applications to answer the question of how self-reconfiguration in modular robots can be applied to current ocean exploration efforts.

Background

Self-Reconfigurable Modular Robots

A modular robot can be defined as a system of modules that work together to perform tasks. The modules are smaller interchangeable units that allow transfer of mechanical forces as well as electrical power and communication (Yim et al.). These modular systems feature a primary structural actuated unit, and often additional specialized units, tailored to perform a task (Yim et al.). These robots can be classified as Self-Reconfigurable if they are able to manipulate and reconfigure their own shape and orientation (Yim et al.). Their primary purpose is to adapt to their environments in order to resolve given challenges. This adaptability is assumed to be highly advantageous in comparison to other robotics systems (Seo et al.). For example, Yim et al. propose an application in space exploration where a modular system could repair itself and alter its shape without requiring separate robotic systems (Yim et al.).

Self-reconfigurable modularity features three main benefits and advantages over conventional robotic implementations, versatility, robustness, and cost (Seo et al.). In the context of versatility modular robotics shows true promise, self-reconfigurable systems potentially have a greater degree of adaptability than standard systems (Seo et al.). The ability to reconfigure allows the robotics system to disassemble and reassemble with high efficiency. Furthermore, this ability allows for various new morphologies that are better suited for their applications (Yim et al.). Such forms of modularity also feature benefits in robustness as robotic parts are interchangeable (Seo et al.). This interchangeability facilitates autonomous part replacement and eventual self-repair (Yim et al.). In addition to robustness, the cost of self-reconfigurable robotic systems is advantageous as producing copies of one or a relative few of robotic modules at scale is generally more cost-effective than manufacturing several specialized robotics platforms (Yim et al.).

M-TRAN: A Foundation of Self-Reconfigurable Modular

One of the earliest and most influential implementations of these principles is M-TRAN (Modular Transformer), a distributed self-reconfigurable system composed of homogeneous robotic modules–developed during the early 2000s by researchers at the Tokyo Institute of Technology and the National Institute of Advanced Industrial Science and Technology (AIST) (Murata et al.). The system was developed to demonstrate autonomous self-reconfiguration and versatile locomotion during the early development of modular robotics research (Murata et al.). The system consists of a series of double-cube modules (see Figure 1) with a hybrid lattice/chain

architecture (Murata et al.). The significance of M-TRAN lies not only in its modular construction, but also in the mechanisms that allow it to coordinate and reconfigure (Murata et al.).

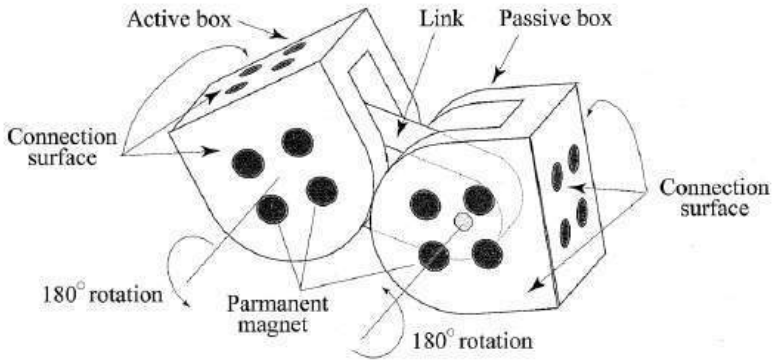


Fig. 1: M-TRAN modular robotic system (Murata et al.).

As shown in Figure 2, the system reconfigures by repeating elementary operations such as detaching from neighboring surfaces, rotating the semi-cylindrical segment of the module and reconnecting the surface to another neighbor (Murata et al.). The system achieves this through the interaction of non-linear springs and magnets within modules, based on the IBMU principle (Internally Balanced Magnetic Unit) (Murata et al.). The force of repulsion exerted by the springs is slightly weaker than the attraction of the magnets, allowing the two polar surfaces to connect (Murata et al.). An SMA (Shape Memory Alloy) coil can then generate a force greater than the net holding force to release connection, providing a reliable attachment and detachment mechanism (Murata et al.). Architecture isn't important by itself—it's important because it enables certain capabilities. The result of an experiment conducted on M-TRAN revealed its ability to both walk and crawl, demonstrating that a single platform can perform multiple locomotion tasks (Murata et al.). M-TRAN’s demonstration of modular adaptability and versatility suggests applications in unstructured operational environments. One possible application is autonomous navigation of deep waters. Through these experiments–demonstrating a robotic platform's ability to alter locomotion through morphology–M-TRAN offers a mechanical foundation for a proposed robotic application in marine environments.

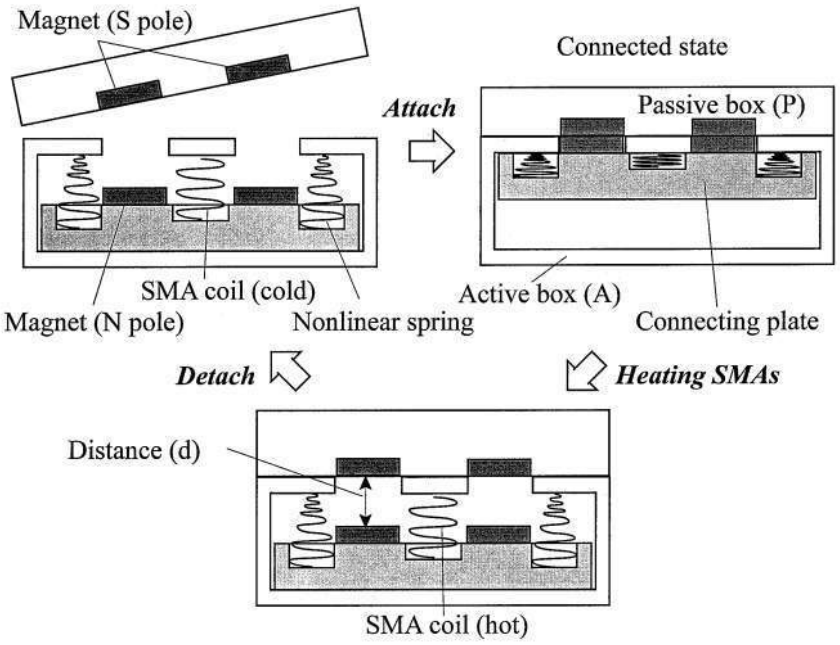


Fig. 2: M-TRAN module and reconfiguration mechanism (Murata et al.).

Ocean Exploration: Beneath The Surface

Environmental Constraints of Deep-Ocean Exploration

Deep-ocean vehicle design is governed by several physical variables. Hydrostatic pressure increases by approximately one atmosphere for every 10 m of depth, requiring pressure-resistant or accommodating mechanisms and components. In addition to increasing pressure, the deep ocean averages roughly 4 °C with waters deeper than approximately 1,000 m lacking any exposure to sunlight (NOAA Ocean Exploration, "What Is the 'Deep' Ocean?"). Salinity, dissolved oxygen, pressure, temperature, and flow rate also influence corrosion behavior in deep-sea materials (Cao et al.). In addition to these water-column conditions, irregular topography and confined passages can restrict vehicle access, reducing the effectiveness of localization methods designed for open waters (Watson et al.). Conventional ROVs and AUVs address these conditions through pressure-resistant structures, acoustic navigation, and mission-specific fixed geometries; however, these solutions do not allow the vehicle to physically change its geometry after deployment (Capocci et al.; Wynn et al.).

Despite covering approximately 70% of Earth's surface, more than 80% of the ocean remains unmapped, unobserved, and unexplored (NOAA, "How Many Species"). As a result, modern ocean exploration has become increasingly dependent upon robotic systems (NOAA Ocean Exploration, "Autonomous Underwater Vehicles"; "Remotely Operated Vehicles"). Currently, UUVs (Unmanned Underwater Vehicles) are widely employed in underwater operations (Watson et al.). These systems can be classified as either ROVs or AUVs, depending on their method of operation (Watson et al.).

ROVs: Remotely Operated Vehicles

Remotely Operated Vehicles (ROVs) are robotic systems controlled via a tethered connection to a surface operator (Capocci et al.). In an aquatic context, their primary purpose is to reduce risk to human operators when navigating harsh underwater conditions (Capocci et al.). The systems are often used to perform tasks involving surveillance and patrolling (He et al.). The system operates using an umbilical tether which carries power, communications, and telemetry between the ROV and the control center (Capocci et al.). The pilot then controls the vehicle in real time using the live camera feeds and sensor data returned through the tether (Capocci et al.). ROVs demonstrate their advantages in ocean environments as they feature a constant power supply, high-bandwidth communication, and real-time human decision making (Capocci et al.). Despite these advantages, ROVs remain constrained by a fixed tether limiting operational range and risking tether entanglement, as well as the need for a control station, which increases operational complexity and cost (Capocci et al.). ROVs therefore exchange unrestricted range and autonomy for reliable power, communication, and direct human control, making their tether both a strength and a physical constraint.

AUVs: Autonomous Underwater Vehicles

Autonomous underwater vehicles (AUVs) are unmanned self-propelled vehicles that function without a tether or additional surface connection (Wynn et al.). These systems are typically deployed from a surface vessel, able to operate independently from the vessel for extended periods (from a few hours to multiple days) (Wynn et al.). They often feature a torpedo-like shape, which is advantageous for underwater operation due to its streamlined geometry (Wynn et al.). The vehicles are controlled through a preprogrammed set of instructions (Wynn et al.). However, this method of operation can limit adaptability in certain complex environments (Wynn et al.). The primary applications of AUVs are found in marine geoscience missions, where they are used in seabed mapping and subbottom profiling (Wynn et al.). This suitability results from their ability to maintain a direct linear trajectory when traveling underwater, as such applications typically require a constant altitude (Wynn et al.). Furthermore, due to the AUV's autonomous nature, deployment vessels can focus on other tasks, allowing for a substantial increase in the rate of data collection (Wynn et al.).

Despite these advantages, AUVs remain subject to several operational limitations. One such limitation is their operational speed range, typically reaching 1.5 – 2.0 m/s (Wynn et al.). This is due to tidal current velocities approaching and exceeding such velocities, potentially causing navigational drift, which can significantly impact data quality (Wynn et al.). In addition to speed, AUVs are vulnerable to acoustic signal disruption, as the lack of a tether results in dependence on acoustic signals (Wynn et al.). There are other operational hazards in human-dominated areas as well, with fishing nets, shipping vessels, and military activity increasing risk of collision or net entanglement (Wynn et al.). These limitations demonstrate that AUV performance is highly dependent on predictable environmental conditions, reducing their ability to adapt to uncharted environments and challenges (Wynn et al.). This combination makes

AUVs highly effective for structured survey paths while leaving a potential advantage for systems capable of changing geometry when terrain or access requirements change.

Due to challenges such as terrain inconsistency, limited adaptability, and communication constraints, there is an opportunity for an adaptive exploration technology to be proposed. Although it features some of these challenges, a self-reconfigurable modular AUV has viable applications. Due to its high degree of adaptability and geometric design, it may prove to be better suited to ocean exploration than conventional UUVs.

Results and Discussion

For the purpose of clarity, Deep Scaler is proposed as a universal system identifier. The various applications of Deep Scaler will be reviewed under application-specific variants. These three will be classified as Deep Scaler Vitality (DSV), Deep Scaler Cartographer (DSC), and Deep Scaler Weaver (DSW). They will function as flora and fauna profilers, seafloor mappers, and navigators of narrow spaces, respectively.

Proposed Deep Scaler Architecture and Design

The base Deep Scaler model (Figure 3) will feature a similar design to M-TRAN (Murata et al.). It will be composed of a variable number of modules each equipped with a marine thruster and a lithium-ion battery. Each module will be capable of independent locomotion and operation. To achieve self-reconfiguration, Deep Scaler will implement a mechanism derived from that of M-TRAN (Murata et al.). For attachment and detachment, Deep Scaler will utilize the IBMU principle in tandem with SMA coil actuation (Murata et al.). Additionally, to enable reconfiguration, the system will feature an embedded geared servomotor system (Murata et al.). Recent modular underwater prototypes provide limited evidence that reconfigurable architectures can function in water. Lan et al. demonstrated a modular underwater robot with multiple configurations and pressure-resistant operation to 30 m, supporting the feasibility of underwater modularity while leaving deep-sea operation unresolved (Lan et al.).

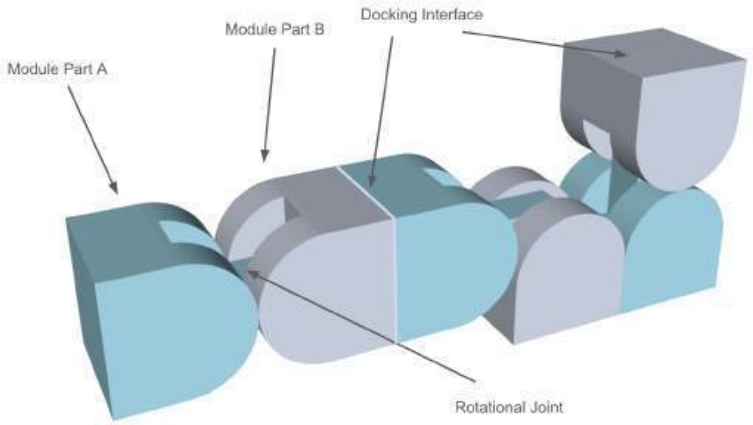


Fig. 3: Proposed base architecture of the Deep Scaler modular robotic system.

Deep Scaler Vitality

DSV (Figure 4) is proposed as a specialized variant for profiling marine fauna and flora in deep-sea environments. Current UUV-based ecological surveys commonly rely on high-frequency digital cameras, able to identify individual organisms on the seafloor (Morris et al.). These systems can also survey biological habitats across large regions, reaching depths of up to 4850 m (Morris et al.). However, the previously discussed challenges of AUVs present important design considerations for DSV. One potential advantage of a modular design is adaptability. Due to its self-reconfiguring ability, DSV could transform into a wide range of configurations, within the scope of its modules. For example, when surveying large ecological habitats, the system could adapt its morphology to avoid collision with surrounding fauna. Reduced disturbance to surveyed organisms would allow for a more natural and authentic perspective on their behaviors (Benoit-Bird et al.). Furthermore, various fauna have been known to exhibit behaviors of avoidance and attraction when exposed to underwater robotic platforms (Benoit-Bird et al.). An additional potential advantage of the system is its maneuverability. Due to each module featuring an individual thruster, the system could reorient its thrusters, potentially enabling omnidirectional movement. Such movement would allow for precise navigation and unobtrusive entry into ecosystems, possibly reducing disturbance to life. Functionally, the camera system would collect high-frequency image sequences at greater than 1 Hz that could be georeferenced and assembled into continuous survey images. These methods are commonly demonstrated in abyssal AUV photography (Morris et al.). A successful DSV survey could therefore be measured through the percentage of imagery suitable for organism identification and the consistency of resulting density estimates. For comparative testing, the same survey region could be imaged by DSV and a conventional fixed AUV, with image usability, organism-detection consistency, collision events, and observable avoidance responses compared between platforms.

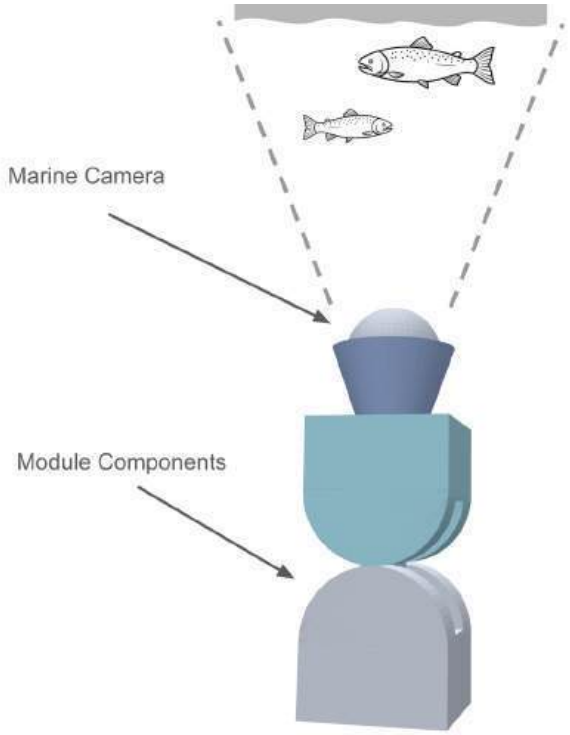


Fig. 4: Proposed Deep Scaler Vitality (DSV) configuration for biological surveying with a marine camera.

However, the reconfiguration process does not come without challenges. Difficulties of water turbidity as well as light dependence must be considered. For one, DSV is highly dependent on visual photography for data gathering, requiring a clear and clean view of an organism. Phytoplankton and Zooplankton as well as other microscopic organisms scatter and absorb light, obscuring visuals in their aqueous environments (Davies et al.). In addition, DSV is planned to operate in deeper ocean environments such as the abyssal zone, which receives zero sunlight, resulting in a completely dark environment (NOAA Ocean Exploration, "What Is the 'Deep' Ocean?"). This may prove to be harmful to the system as most marine photography devices require light to capture imagery (Morris et al.). In addition to it traveling a minute length, light may also disturb wildlife (Benoit-Bird et al.). Most creatures dwelling in these oceanic zones are acclimated to a pitch black surrounding, resulting in disturbance from light (Benoit-Bird et al.). In order to avoid harming native organisms, DSV would need to avoid using illuminated imagery.

Deep Scaler Cartographer

DSC (Figure 5) is proposed as a specialized variant for high-resolution seafloor mapping. Most of the current UUV-based seafloor mapping systems rely on echo sounding, firing hundreds of sound waves that they then receive to deduce surrounding textures (Wynn et al.). Such mechanisms will be implemented into DSC, allowing for terrain mapping in environments

with little to no light exposure (Wynn et al.). In addition, separate modules will feature dedicated sensing instruments. For the current application, three modules will be designated to data collection. The first module will feature an echo sounder, used for gathering acoustic data in aphotic environments (Wynn et al.). The second module will feature a LiDAR sensor for mid-range ocean mapping and shallower oceanic levels (Heffner et al.). It will gather data by firing rapid laser pulses at the floor and measuring the return time of the laser pulse (Heffner et al.). Finally, the third module will feature a standard marine camera, used for gathering image data of shallower and clearer environments such as coral reefs. Additionally, DSC may be advantageous in the quality of its photography. Due to its potential ability to approach organisms with minimal distance, DSC could achieve high-resolution imagery of sea life, as a closer view results in a greater spatial resolution (Wynn et al.). Sensor performance could be assessed by comparing the reconstructed terrain against known reference geometry and measuring positional drift across the survey. Underwater laser-scanner experiments have demonstrated point-cloud SLAM capable of limiting navigation drift while producing consistent high-accuracy 3D maps, providing a practical benchmark for DSC's LiDAR implementation (Palomer et al.). A DSC trial could compare its reconstructed map against a known reference surface using mean terrain error as well as percentage of area successfully reconstructed and accumulated localization drift. Heffner et al. reported approximately a 1.95 mm range precision and a 6.01 mm mean relative range accuracy for an underwater time-of-flight laser scanner in controlled conditions, providing a quantitative laboratory benchmark for the implemented LiDAR system (Heffner et al.).

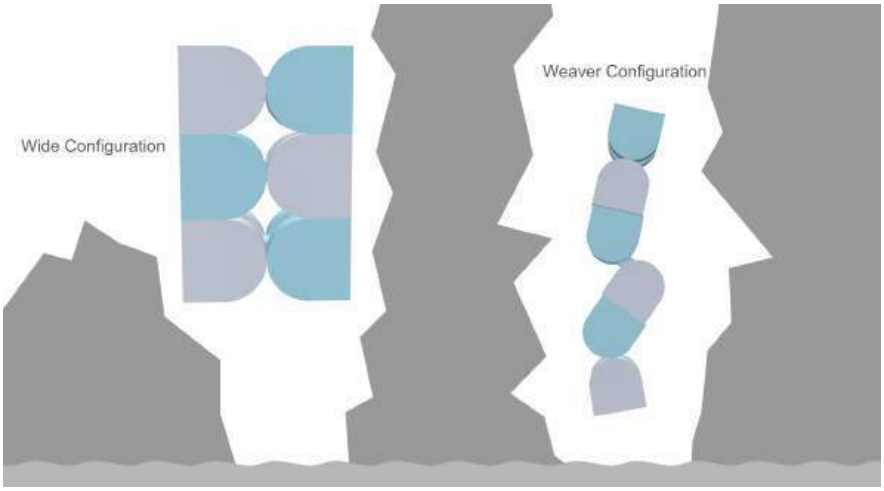


Fig. 5: Proposed Deep Scaler Cartographer (DSC) configuration showing LiDAR, echo-sounder, and camera sensing for seafloor mapping.

However, such advantages do not come without various limitations. For one, floor mapping is highly dependent on positional accuracy and knowing where measurements were collected (Wynn et al.). When unwanted displacements are caused by currents or system instability, mapping accuracy is at risk (Wynn et al.). In addition, configuration changes while surveying risk interference with sensor positioning and motion. As a result, DSC may need to

maintain one configuration, reducing the usefulness of continuous reconfiguration. Moreover, conventional AUVs are already extremely capable marine mappers in open seafloors (Wynn et al.). DSC's terrain adaptability would likely provide the greatest benefit in irregular or variable terrain, rather than ordinary seafloor surveys.

Deep Scaler Weaver

DSW (Figure 6) is proposed as a specialized variant designed for navigating caves, sedimentary tunnels, bottlenecks, and various other geological spaces. Standard UUVs with rigid vehicle dimensions are restricted in their access to physical underwater passages (Watson et al.). DSW aims to address this issue by adapting its configuration to the various geological passages it encounters. An advantage of adaptability is variability in shape and size. Software cannot alter a rigid vehicle's geometry and ability to fit through tight spaces, however changing morphology potentially can. As such, DSW's modular abilities would allow it to reconfigure into narrower, elongated configurations when facing restricted passages, enabling its traversal. In addition, maneuverability suggests multiple benefits as well. Due to each of its modules featuring a thruster, DSW can distribute its propulsion. An adaptable position of propulsion allows for navigation of irregular passages without depending on a forward orientation. Such maneuverability enables easier traversal through sharp turns, irregular tunnel geometry, and various other obstacles present in the ocean. To support navigation when direct surface communication is unavailable, DSW could use paired acoustic sonars to map surrounding horizontal and vertical surfaces while scan-matching SLAM estimates its position (Mallios et al.). Success could be measured by whether DSW traverses and returns from a confined passage without collision while maintaining bounded localization error throughout the mission. Performance should additionally record minimum traversable passage cross-section, collision count, mission completion rate, and final localization error relative to a fixed-morphology UUV.

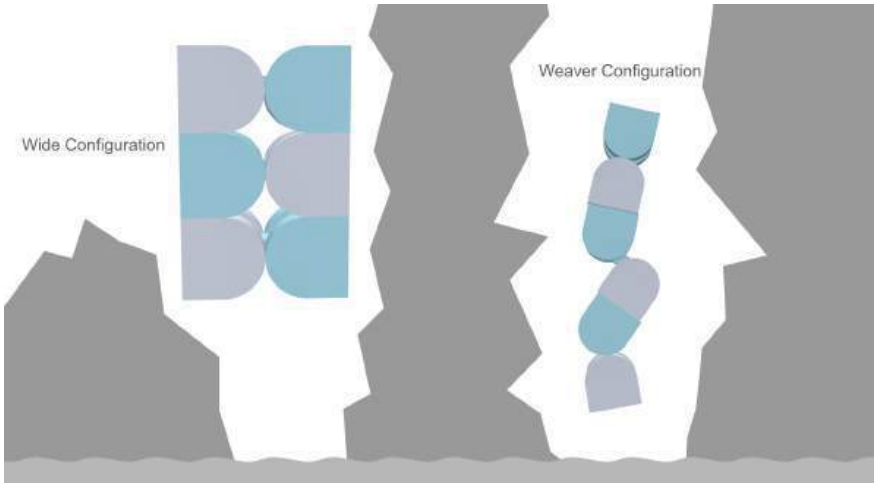


Fig. 6: Proposed Deep Scaler Weaver (DSW) configuration for navigation through confined underwater passages.

However, such advantages do not come without various limitations. For one, confined underwater environments interfere with acoustic communication (Dikarev et al.). Without communication to the surface, DSW is at risk of losing localization in geological space inaccessible to humans (Mallios et al.). Such risk would pose a significant risk as an autonomous system may not have a method of returning to its origin. An additional possible limitation is reconfiguration clearance, as the system would require sufficient clearance to transform into the necessary figure. A passage may be large enough for the system's initial entry but too restrictive to safely reconfigure within the space. Confined-space exploration may be one of Deep Scaler's strongest applications as variable morphology addresses physical access constraints. However, severe navigation and recovery risks require further analysis of its realistic applications.

Overall Advantages and Limitations of Deep Scaler

Overall Deep Scaler presents many advantages in its architecture, manufacturing, and adaptability. In terms of architecture, the same fundamental module architecture is employed in DSV, DSC, and DSW. Varying module count and arrangement allows for specialization without the need for an individual unrelated robotic architecture for each mission (Seo et al.). As a result of this consistency, manufacturing proves to be relatively cost-efficient (Seo et al.). Production of standardized modules could be simpler than numerous specialized components and systems (Seo et al.). Replacement and maintenance may also prove to be easier as damaged modules could be replaced individually rather than replacing the system entirely (Seo et al.). Additionally, in terms of adaptability, Deep Scaler features morphology responsive to environmental requirements. Such variable morphology suggests more applications than conventional fixed-morphology platforms (Seo et al.). Overall Deep Scaler boasts many possible advantages in comparison to modern explorative implementations.

Overall Deep Scaler also presents many limitations in energy consumption, mechanical complexity, hydrodynamic efficiency, control complexity, and deep-water durability. Starting with energy consumption, it has been established that every module contains propulsion, power, and actuated reconfiguration components, leaving little space for power storage. Locomotion and repeated reconfiguration could constrain endurance, expending substantial energy during long and deep missions. Deep Scaler's multiple docking mechanisms, servomotors, thrusters, batteries and moving interfaces create a system with high mechanical complexity (Lan et al.). Such complexity poses higher risks of component or connection failure (Lan et al.). Extreme depths prevent immediate repair of the system resulting in such failures being detrimental to operation. In addition to simplicity, hydrodynamic efficiency may be sacrificed. Modular figures may not achieve standard streamlined geometry that conventional torpedo-shaped AUVs employ (Wynn et al.). This would suggest adaptability gained at the expense of efficient sustained travel. As well as mechanical complexity, control complexity may also be an issue (Lan et al.). Deep Scaler would be required to coordinate propulsion as well as movement across multiple independently capable modules while its topology changes (Lan et al.). Each new figure potentially raises new movement and control requirements for the system (Lan et al.). This is a recognized issue in

modular underwater robotics research (Lan et al.). Finally, deep-water durability must be achieved. In order for operation, every module must remain functional under extreme conditions. Modular connection points and moving mechanisms may prove to be particularly vulnerable to ocean environments (Lan et al.). Overall, Deep Scaler requires many challenges to be addressed before proper implementation.

Future Directions

Deep Scaler currently exists as a proposed architecture and requires physical validation before further application. To properly assess the viability of the proposed system a prototype must be developed. The prototype should demonstrate underwater attachment and detachment as well as reconfiguration and coordinated propulsion. Additionally, the prototype should begin in a controlled shallow environment before progressing toward greater depths and environmental complexity. Individual prototypes of DSV, DSC, and DSW should also be constructed and tested in their intended operating conditions. In terms of future development, Deep Scaler should be optimized for energy consumption, module geometry, propulsion, control, and connection reliability. Physical experimental results could determine where self-reconfiguration genuinely outperforms conventional UUVs, proving its real advantages and limitations. More broadly, underwater self-reconfigurable modular robotics requires pressure-resistant moving joints and connectors, energy-efficient distributed control, reliable docking under current disturbances, and localization methods that remain functional in confined environments (Lan et al.; Mallios et al.; Dikarev et al.). Progress in these areas would benefit not only Deep Scaler but future modular underwater platforms. Comparative experiments should use fixed-morphology UUVs as controls so that improvements in access, mapping quality, biological observation, or maneuverability can be separated from the added cost and complexity of reconfiguration.

Conclusion

This paper indicates that conventional underwater robots are highly capable yet remain constrained by their fixed morphologies and environment-specific limitations. As such, deep and geometrically complex environments create opportunities for new technologies such as physical adaptability. Deep Scaler applies self-reconfigurable modular robotics to underwater exploration through its variants: DSV, DSC, and DSW. Each specialized model requires demonstration of prototyping, and testing before establishing superiority over conventional UUVs. Such research in self-reconfigurable modular platforms has potential significance in current efforts to explore the Earth's deepest waters.

Works Cited

Benoit-Bird, Kelly J., et al. “Schrödinger’s Fish: Examining the Robotic Observer Effect on Pelagic Animals.” Limnology and Oceanography: Methods, vol. 21, no. 9, 2023, pp. 563–580. https://doi.org/10.1002/lom3.10565.

Cao, Pan, et al. “Research Progress on Corrosion and Protection in Deep-Sea Environment.” Journal of Chinese Society for Corrosion and Protection, vol. 35, no. 1, 2015, pp. 12–20. https://doi.org/10.11902/1005.4537.2013.253.

Capocci, Romano, et al. “Inspection-Class Remotely Operated Vehicles—A Review.” Journal of Marine Science and Engineering, vol. 5, no. 1, 2017, article 13. https://doi.org/10.3390/jmse5010013.

Davies, Emlyn J., et al. “The Hidden Influence of Large Particles on Ocean Colour.” Scientific Reports, vol. 11, no. 1, 2021, article 3999. https://doi.org/10.1038/s41598-021-83610-5.

Dikarev, Aleksandr, et al. “Combined Multiuser Acoustic Communication and Localisation System for μAUVs Operating in Confined Underwater Environments.” IFAC-PapersOnLine, vol. 48, no. 2, 2015, pp. 161–166. https://doi.org/10.1016/j.ifacol.2015.06.026.

He, Ying, et al. “A Review of Different Designs and Control Models of Remotely Operated Underwater Vehicle.” Measurement and Control, vol. 53, nos. 9–10, 2020, pp. 1561–1570. https://doi.org/10.1177/0020294020952483.

Heffner, Ellen, et al. “Laboratory-Based Assessment of the Underwater Laser Scanner ULi.” The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLVIII-2/W10-2025, 2025, pp. 123–129. https://doi.org/10.5194/isprs-archives-XLVIII-2-W10-2025-123-2025.

Lan, Lincong, et al. “Design and Validation of a Modular Reconfigurable Underwater Robot.” Ocean Engineering, vol. 347, 2026, article 124040. https://doi.org/10.1016/j.oceaneng.2025.124040.

Mallios, Angelos, et al. “Toward Autonomous Exploration in Confined Underwater Environments.” Journal of Field Robotics, vol. 33, no. 7, 2016, pp. 994–1012. https://doi.org/10.1002/rob.21640.

Morris, Kirsty J., et al. “A New Method for Ecological Surveying of the Abyss Using Autonomous Underwater Vehicle Photography.” Limnology and Oceanography: Methods, vol. 12, no. 11, 2014, pp. 795–809. https://doi.org/10.4319/lom.2014.12.795.

Murata, Satoshi, et al. “M-TRAN: Self-Reconfigurable Modular Robotic System.” IEEE/ASME Transactions on Mechatronics, vol. 7, no. 4, 2002, pp. 431–441. https://doi.org/10.1109/TMECH.2002.806220.

NOAA. “How Many Species Live in the Ocean?” NOAA National Ocean Service, https://oceanservice.noaa.gov/facts/ocean-species.html. Accessed 6 Sept. 2026.

NOAA Ocean Exploration. “Autonomous Underwater Vehicles.” National Oceanic and Atmospheric Administration, https://oceanexplorer.noaa.gov/technology/subs-auvs/. Accessed 6 Sept. 2026.

NOAA Ocean Exploration. “Remotely Operated Vehicles (ROVs).” National Oceanic and Atmospheric Administration, https://oceanexplorer.noaa.gov/technology/subs-rovs/. Accessed 6 Sept. 2026.

NOAA Ocean Exploration. “What Is the ‘Deep’ Ocean?” National Oceanic and Atmospheric Administration, https://oceanexplorer.noaa.gov/ocean-fact/deep-ocean/. Accessed 6 Sept. 2026.

Palomer, Albert, et al. “Inspection of an Underwater Structure Using Point-Cloud SLAM with an AUV and a Laser Scanner.” Journal of Field Robotics, vol. 36, no. 8, 2019, pp. 1333–1344. https://doi.org/10.1002/rob.21907.

Seo, Jungwon, et al. “Modular Reconfigurable Robotics.” Annual Review of Control, Robotics, and Autonomous Systems, vol. 2, 2019, pp. 63–88. https://doi.org/10.1146/annurev-control-053018-023834.

Watson, Simon, et al. “Localisation of Unmanned Underwater Vehicles (UUVs) in Complex and Confined Environments: A Review.” Sensors, vol. 20, no. 21, 2020, article 6203. https://doi.org/10.3390/s20216203.

Wynn, Russell B., et al. “Autonomous Underwater Vehicles (AUVs): Their Past, Present and Future Contributions to the Advancement of Marine Geoscience.” Marine Geology, vol. 352, 2014, pp. 451–468. https://doi.org/10.1016/j.margeo.2014.03.012.

Yim, Mark, et al. “Modular Self-Reconfigurable Robot Systems [Grand Challenges of Robotics].” IEEE Robotics & Automation Magazine, vol. 14, no. 1, 2007, pp. 43–52. https://doi.org/10.1109/MRA.2007.339623

Competing on Values - A Multiple-Case Qualitative Analysis of Fashion Retail Platform Strategic Positioning By Ariana Riggi

ABSTRACT

The rapid expansion of fashion resale has transformed secondhand clothing from a niche retail category into a dynamic, mainstream growth segment of the global fashion industry. Existing research has extensively examined the consumer motivations underlying this growth, identifying sustainability, affordability and accessibility, identity and self-expression, entrepreneurship, community, and trust as the key drivers of participation. Less attention, however, has been devoted to how resellers strategically respond to these multidimensional consumer values. This study addresses that gap, seeking to uncover how fashion resale platforms differentiate themselves to attract their target users, utilizing a multi-case qualitative analysis of six prominent fashion resale platforms: Depop, Poshmark, ThredUp, The RealReal, Vinted, and Vestiaire Collective. Each company was evaluated across the six consumer-value dimensions using a standardized qualitative scoring model. The findings demonstrate that resale platforms compete by emphasizing different combinations of benefits. Depop prioritizes identity, entrepreneurship, and community; Poshmark similarly prioritizes both identity and entrepreneurial participation; ThredUp centers sustainability and affordability while increasingly incorporating identity; The RealReal competes through trust and accessibility; Vinted emphasizes sustainability, affordability, and entrepreneurship; and Vestiaire Collective promotes fashion identity and sustainability. These findings inform a novel 2x2 framework that maps resale firms according to the consumer value they provide, ranging from functional to symbolic/aspirational, and their consumer relationship model, ranging from transaction-oriented to participation-oriented, resulting in four fashion resale positioning archetypes. The study demonstrates that competitive advantage in the maturing resale industry increasingly depends on constructing and communicating these differentiated value propositions that are aligned with evolving consumer values, with implications for both resale and retail in the fashion industry, and the circular economy and digitally mediated marketplaces more broadly.

Keywords fashion resale, secondhand fashion, online resale platforms, consumer motivations, strategic positioning, Generation Z, circular economy

1. INTRODUCTION

Historically, second-hand clothing often existed on the margins of mainstream culture. Thrift shops, like Goodwill and Salvation Army, were widely associated with customers shopping out of economic necessity. At the same time, beginning in the late 20th century, much of mainstream fashion came to be dominated by fast fashion, churning cheaply produced goods to meet season trends as retailers like H&M, Forever 21, and Zara transformed the fashion industry by dramatically shortening production cycles and lowering prices. New collections were continuously introduced, creating a sense of novelty and urgency that encouraged consumers to

buy clothing more frequently. Garments came to be seen as disposable, made possible by the low cost of fast fashion which allowed consumers to participate in a cycle of constant consumption without significant financial consequence. According to the World Economic Forum, while people bought 69% more clothing between 2000 and 2015, they kept them for only half as long (Mc-Fall-Johnsen). In that 15-year period, clothing utilization decreased by 36%, causing nearly $500 billion of value to be lost as more than half of fast fashion was disposed of in under a year ("New Textiles Economy").

Over the past three decades, the global fashion industry has undergone a profound shift as the secondary apparel market has evolved from a niche alternative to one of the fastest growing segments of retail. The emergence of the internet revolutionized retail, then eBay, founded in 1995, normalized and scaled the exchange of used goods by reducing the effort required to buy and sell pre-owned clothing. Mobile commerce, personalized recommendations, AI-enabled search, integrated payments, authentication technologies, and social commerce features have accelerated the industry's growth by reducing the friction associated with secondhand transactions. At the same time, Generation Z has come of age, causing resale to further emerge as a significant force reshaping the fashion industry as these younger consumers increasingly make values-driven fashion purchasing decisions. These cultural and technological shifts have led to the expansion of the circular economy and businesses to rethink traditional retail models.
The global secondhand market, currently estimated at $210-220 billion, is growing at 10% annually, 3X faster than the firsthand market (Moizant et al). As competition within the secondary fashion market intensifies, platforms must do more than simply facilitate transactions. They have to differentiate themselves with distinct value propositions that resonate with today's consumers. Examining how fashion resale platforms position themselves to appeal to evolving consumer preferences provides insight into not only the future of fashion retail but can also offer broader implications for the circular economy as businesses across industries aim to respond to the values-driven consumer preferences of younger generations.

2. LITERATURE REVIEW

The rapid growth of the secondary fashion market has prompted a substantial body of recent research examining why consumers choose to purchase and sell pre-owned clothing. Research shows that motivation for secondhand consumption is multidimensional, including a combination of economic, ethical, psychological, and social factors. Pinto and Royo-Vela's 2026 review integrates prior research on sustainability concerns, economic value, uniqueness and identity motives, and socio-cultural influences, concluding that pro-environmental values frequently coexist with utilitarian and symbolic motives rather than operating independently. Their proposed model looks at conditions that influence whether favorable attitudes translate into actual purchasing behavior, identifying three tiers of drivers. These include ecological concern, mindful consumption, and green behavior intention as direct drivers (Pinto and Royo-Vela 6) while recognizing green lifestyle, price sensitivity, personal values as moderating relationships (Pinto and Royo-Vela 7), and generational cohort as control relationships (Pinto and Royo-Vela

10). This multidimensional interpretation is echoed by Blaz et al's systematic review of digital resale research, which organizes consumer motivations into four broad categories: cost-driven, operational convenience, identity-driven, and sustainability-driven motivations (Blaz et al 500). Likewise, Szűcs and Földi characterize the literature as demonstrating "a mix of practical, hedonic, and ethical motivations," with affordability and value for money, unique style and treasure hunting thrill, environmental and ethical considerations, and social influence and norms repeatedly appearing across studies and contexts (Szűcs and Kata Földi 31). Collectively, this literature establishes that the growth of resale cannot be attributed to any singular motivation but rather, consumers are attracted by different combinations of value that can vary by generation, market segment, and platform.

Sustainability is nevertheless among the most extensively studied motivations for secondhand consumption. Blaz et al identifies digital platforms as enablers of environmentalism because online marketplaces allow consumers to translate environmental values into behavior by making reused goods easier to find and circulating them beyond localized thrift networks (Blaz et al 503). Pinto and Royo-Vela argue that sustainability-related intentions are moderated by other factors that may precede environmental concern depending on the setting and context (Pinto and Royo-Vela 7). Their review also notes that younger cohorts generally display stronger engagement with ecological issues because of the social and educational environments in which they developed. In fact, 65% of Generation Zs say that the environment is a major source of anxiety and that environmental concerns influence their consumer behavior ("2025 Gen Z and Millennial Survey"), revealing that sustainability is not an abstract concern, but an emotionally salient issue that informs their everyday buying decisions.

A second and frequently more immediate motivation is economic value and affordability. Szűcs and Földi describe lower prices as a primary attraction across demographic groups and cite studies in which financial advantages outranked environmental concerns as a reason for purchasing secondhand clothing. For younger consumers in particular, resale provides access to fashion at prices below traditional retail, while perceived economic benefits can also strengthen perceived value, trust, and willingness to recommend secondhand purchasing to others (Szűcs and Kata Földi 30). Economic realities have contributed to the rise of resale among Gen Zs. Today's young people entered the consumer market during a period marked by economic uncertainty, including the COVID-19 pandemic, rising inflation, and concerns about student debt. Cost of living is a top concern for Gen Z as 48% say they do not feel financially secure ("2025 Gen Z and Millennial Survey"). These economic conditions have made affordability a practical consideration when purchasing clothing.

Blaz et al broadens this economic dimension beyond low prices, identifying fair pricing, bargaining power, bargain hunting, and the opportunity to generate extra income as distinct cost-related motivations (Blaz et al 500). Recent industry research strongly reinforces these academic findings. A 2025 BCG and Vestiaire Collective survey of 7,800 resale-engaged consumers found that 78% strongly agreed that value for money was an important reason for purchasing secondhand premium, designer, or luxury fashion, and 47% identified access to

otherwise unaffordable brands as a key motivation. The same study also found economic motivations on the selling side as 44% sold in order to fund additional secondhand purchases and 41% to generate money more generally (Moizant et al). ThredUp's 2025 industry research similarly places "better deals" and access to higher-end brands among the leading reasons consumers shop secondhand and finds that younger generations increasingly factor an item's future resale value into initial purchase decisions ("ThredUp Resale Report"). These findings indicate that resale changes the economics of fashion consumption in two directions whereby buyers gain access to lower prices for otherwise unattainable products, while sellers can recover value from existing wardrobes.

The ability to earn money from resale also connects economic motivations to entrepreneurship, particularly as digital platforms allow consumers to participate on both sides of the market. Blaz et al observe that online resale has shifted transactional responsibilities toward consumers, who increasingly function not only as purchasers but also as sellers, curators, and other active market actors (Blaz et al 501). This differs meaningfully from conventional retail, in which consumers generally occupy only the demand side of the transaction. Digital resale instead permits clothing to function as a partially recoverable asset where consumers may purchase an item, use it, and subsequently resell it. Rather than viewing clothing as a permanent possession, young shoppers treat garments as temporary assets, creating a fluid approach to garment ownership that enables resale marketplaces to act as financial tools. Generation Z is more entrepreneurially minded than its predecessors and resale platforms have given rise to "closet entrepreneurs." ThredUp's 2025 research found that 39% of younger consumers resold apparel in 2024 and that younger consumers attempted to resell an average of 52% of their unwanted clothing ("ThredUp Resale Report"). As such, the economic appeal of resale cannot be understood entirely through buyer savings but also as a result of the income-generating opportunities they create by lowering the barriers to informal entrepreneurial activity.

Research also emphasizes that identity, uniqueness, and social influence shape demand for secondhand goods since secondhand shopping gives consumers access to vintage, scarce, discontinued, or one-of-a-kind goods that may be unavailable in conventional stores, allowing clothing to function as a means of differentiation from mass-produced fashion. Szűcs and Földi identify uniqueness, originality, the "treasure hunt" experience, enjoyment, and creative self-expression as recurring motivations and also report evidence that younger cohorts derive particularly strong hedonic value from secondhand shopping (Szűcs and Kata Földi 31). Blaz et al extends this analysis specifically to online platforms, arguing that identity-driven motivations are connected to how individuals understand themselves and how they wish to be perceived by others (Blaz et al 502). Digital environments intensify this process by facilitating curated self-presentation and visibility while rare or vintage products can become forms of cultural capital. A McKinsey study found that 58% of customers associate brands being resold with trendiness ("The State of Fashion 2026"). Additionally, social media creates network effects because reselling clothing has digitally native Generation Zs wear "new to them" outfits that they have not yet posted on social media, allowing them to refresh their online identities without

contributing to the churn of fashion that is antithetical to their values. Research synthesized by Szűcs and Földi finds that perceived approval from peers can significantly increase purchase intentions and that the normalization of thrifting within a social group reduces the stigma historically associated with used clothing. Social media and influencers may further accelerate this transition by publicly displaying thrift hauls and vintage fashion, thereby changing secondhand clothing from a potential marker of financial necessity into a visible marker of taste (Szűcs and Kata Földi 32). This is further demonstrated as 14.6 million Instagram posts were tagged #thrifting as of November 2025 (Chew). Therefore, community and social validation are mechanisms capable of changing the meaning and desirability of resale.

At the same time, the literature makes clear that motivations operate alongside important trust and risk considerations, particularly online and in luxury resale. Consumers cannot physically inspect items before purchasing them, creating uncertainty surrounding condition, quality, seller credibility, authenticity, and hygiene. Blaz et al identify product-related uncertainty, seller credibility, psychological barriers, and platform constraints as four major categories of resistance to digital secondhand purchasing (Blaz et al 508). Transparency, reviews, guarantees, and detailed listings can reduce perceived uncertainty, while authentication is particularly consequential for higher-value luxury products. Thanasi-Boçe and Al-Issa's review identifies that counterfeits, undisclosed defects, hygiene, and data security are among the barriers to luxury resale participation and describes authentication risk as a central inhibitor. Their analysis finds that luxury consumers pursue symbolic benefits such as self-expression, uniqueness, status, and exclusivity alongside rational benefits such as affordability and sustainability (Thanasi-Boce and Al-Issa 12-15).

The existing literature provides a strong foundation for identifying the six consumer-value themes used in the empirical analysis of this study: sustainability, affordability and accessibility, identity and self-expression, entrepreneurship, community, and trust. However, despite the depth of recent research on why consumers participate in resale, considerably less work examines how online resale firms translate those motivations into differentiated strategic positioning. This represents the central contribution of this study.

3. METHODS

This study utilizes multi-case qualitative research to examine how six fashion resale platforms strategically position themselves within the secondary fashion market. The companies analyzed are: Depop, Poshmark, ThredUp, The RealReal, Vinted, and Vestiaire Collective. They represent some of today's most prominent resale competitors who have different business models, target consumers, and branding objectives. Data was collected from publicly available sources, including corporate websites, annual and impact reports, marketing campaigns, social media content, executive interviews, and industry reports in order to assess company positioning. Each company was evaluated across six themes identified from the existing literature about the dimensions most important to today's Generation Z consumers including: 1) sustainability, 2) affordability and accessibility, 3) identity and self-expression, 4) entrepreneurship, 5)

community, and 6) trust. For each theme, a set of sub-questions was used to assess whether and to what extent the theme appeared in the company's messaging, platform features, and value proposition. Evidence was gathered and documented for every sub-question, after which each sub-question received a qualitative score ranging from 0 to 2, where 0 indicated that the dimension was absent, 1 indicated that it was present but not significant, and 2 indicated that it was consistently emphasized and central to the company's positioning. To determine the overall score for a theme, the sub-question scores within the theme were averaged. The scoring system was intended to create a consistent basis to compare companies while allowing for qualitative interpretation. The ultimate goal of the study is to clarify the competitive landscape by mapping the strategic positioning of these leading firms and to develop a new framework through which the contemporary fashion resale industry can be understood.

4. RESULTS

Depop

Founded in 2011 in London, Depop has evolved from a peer-to-peer fashion marketplace into one of the defining platforms of Generation Z's resale economy. Acquired by Etsy in 2021, the platform distinguishes itself by presenting fashion as a cultural experience and positioning resale as an extension of personal identity, entrepreneurial opportunity, and community involvement, with sustainability serving as an important underlying corporate philosophy rather than the primary focus of consumer-facing marketing.

Across the six themes examined for this study, Depop received a high average category score for identity and self-expression (2/2), entrepreneurship (2/2), and community (1.75/2), indicating that these themes consistently permeate the company's mission, platform features, and public communications. By contrast, moderate scores were assigned to sustainability (1.33/2), trust (0.75/2) and affordability (1/2), reflecting that although these are acknowledged, they are rarely used as primary marketing messages. (See table A below)

Identity is one of Depop's strongest strategic priorities and the company encourages consumers to "discover" and cultivate a "unique personal style." Its recent campaign slogan, "Where taste recognizes taste" ("Depop—Where Taste Recognizes Taste"), further reinforces the idea that fashion is a form of cultural expression and shared identity, with Depop's marketplace serving as a destination where consumers can discover distinctive pieces that showcase their individuality. This approach aligns with Generation Z's desire to express personal identity, particularly given the visibility and pressure created by social media to constantly curate one's online image.

Entrepreneurship is another defining component of Depop's positioning. Unlike other platforms that frame resale as a way to declutter, Depop encourages users to view selling as an entrepreneurial opportunity, which is further reinforced by charging no seller fees and providing seller-oriented features including revenue tracking, product view analytics, downloadable sales reports, and transaction histories. This emphasis demonstrates Depop's support for users seeking

to pursue resale as an ongoing business, casting them not simply as suppliers of inventory but as co-creators whose personal success contributes to the vitality of the platform as a whole. The company's latest ad campaign stresses resale as a fruitful side-hustle where users can earn extra income, featuring the background track Kelis' song "Millionaire," in which the lyrics promote the collective empowerment of entrepreneurship stating, "I ain't rich, 'til he is rich, and she is rich, then we is rich" ("Easy to Sell").

Community also emerged as a core theme. Depop invests heavily in its Student Campus Ambassador programs ("Depop on Campus"), as well as pop-up events, festivals such as Governors Ball ("Depop Partners with Gov Ball"), virtual workshops, and celebrity partnerships with artists including Zara Larsson ("Zara Larsson rings Her Signature Style") and Kacey Musgraves. Prioritization of community as a strategic differentiator is seen on Depop's Instagram posts which highlights events, collaborations, and creator stories even over direct product promotion.

Depop substantiates its commitment to circular fashion through detailed impact reporting which boasts that "over 100 million items that already existed have been given a second life" ("Depop Impact") since the company's founding and estimates that over three in five purchases displace the purchase of new clothing. It also "invests in climate solutions like electric trucks and carbon offsets for everyday delivery" ("Depop's Climate Impact Plan"). Corporate communications underscore the importance of circularity, describing the company's mission as encouraging consumers to extend the life of clothing by discovering "what already exists." Depop indicates in its mission statement that its goal is "to make secondhand as exciting and rewarding as buying new" ("Depop About"), explaining why messages about environmentalism are largely absent from the company's consumer advertising and suggesting that sustainability functions primarily as the philosophical foundation that legitimizes the brand rather than the emotional appeal through which it attracts consumers.

Taken together, the results of this study reveal that Depop has deliberately positioned itself as a cultural ecosystem where fashion, entrepreneurship, creativity, and community intersect.

Poshmark

Since its inception in 2011, Poshmark has sought to combine peer-to-peer commerce with elements of social networking, allowing users to build digital closets, follow one another, and engage in community-driven shopping. The company consistently positions resale as both an entrepreneurial opportunity and a vehicle for personal style, with a community orientation serving as a secondary supporting feature.

Across the six themes examined in this study, Poshmark received the highest possible score for identity (2/2). Entrepreneurship, while only receiving a moderate score (1/2) for that theme, must be qualitatively assessed as a strategic imperative for the company due to its consequence in customer-facing marketing efforts, representing the one instance of deviation from the study's defined scoring approach. Both identity and entrepreneurship are prominently

reflected throughout Poshmark's brand messaging. Community received a moderately high score (1.5/2), as the platform incorporates numerous social features and ambassador initiatives but does not consistently emphasize community as a central marketing message. Affordability (1/2), sustainability (1/2), and trust (1/2) each received moderate scores, indicating that these themes are present within the platform but function as tertiary characteristics. (See Table B below)
Identity is one of Poshmark's defining strategic priorities. The company's recent "Vibe, Check" campaign ("Vibe, Check: Every Era") ("Vibe, Check") encourages users to discover distinctive pieces that reflect their personalities and style. As such, the company reframes resale as a means of cultural participation and a mechanism for self-expression.

Entrepreneurship represents the second pillar of Poshmark's positioning. Numerous advertising campaigns explicitly market the platform as "the perfect side hustle" ("Perfect Side Hustle") and encourage sellers to "start cashing in" ("Start Cashing In"). The platform enables sellers to track likes, communicate with interested buyers, send targeted discounts, and automate negotiations within predetermined price ranges. However, the platform offers relatively limited business analytics compared to dedicated e-commerce marketplaces, suggesting that Poshmark primarily targets casual entrepreneurs and side-hustlers. Notably, given this audience, Poshmark maintains seller fees of $2.95 for sales under $15 and 20% for higher-priced transactions, indicating that the company competes on convenience, marketplace reach, and ease of participation, not low transaction costs.

Community occupies a somewhat different position within Poshmark's value proposition. The platform enables users to follow one another, like listings, comment on products, and participate in ambassador programs ("How to Become a Posh Ambassador") and live events. However, this study's investigation revealed that many of these interactions are closely tied to active transactions, meant to facilitate purchases rather than broader social engagement.
The remaining themes are present but occupy less prominent positions within Poshmark's positioning. Sustainability received a moderate score because, although the company acknowledges the environmental benefits of resale through initiatives such as its "Posh for the Planet" campaign ("Posh for the Planet"), sustainability is not deeply integrated into the company's communications. Similarly, the idea of affordability is incorporated through discounted merchandise, negotiated pricing, and access to luxury products that may otherwise be unattainable, yet the platform does not market itself as a bargain retailer.

Finally, trust functions primarily as marketplace infrastructure whereby Poshmark provides authentication for luxury items valued above $500, buyer and seller protections through Posh Protect ("Posh Protect"), escrow-style payment processing, and seller ratings ("New Feature Alert"). However, since authentication applies only to selected luxury purchases ("Posh Authenticate"), lower seller ratings remain hidden from prospective buyers, and trust-related features receive relatively little attention within brand messaging and instead, these security features are treated almost as operational requirements.

Poshmark's messaging emphasis on personal expression and seller empowerment reflects the company's focus on identity and entrepreneurship as a way to attract consumers who want to simultaneously participate in both the buying and selling sides of the market.

ThredUp

Founded in 2009 in San Francisco, ThredUp has grown into one of the world's largest online resale platforms, operating primarily through a consignment model in which users receive a "Clean Out Bag" to send clothing directly to the company for processing, pricing, photography, and fulfillment. Unlike peer-to-peer marketplaces that facilitate direct transactions between buyers and sellers, ThredUp functions more as an online retailer, controlling much of the resale experience itself. This operating model allows the company to emphasize convenience while positioning secondhand shopping as both an environmentally responsible and financially sensible alternative to purchasing new apparel. As such, ThredUp's competitive positioning centers primarily on sustainability and affordability, while its recent branding has increasingly incorporated elements of identity.

Based on analysis of the six themes in this study, ThredUp received the highest possible score for sustainability (2/2) across every evaluation criterion, reflecting the prominence of circular fashion throughout its corporate communications, environmental reporting, and consumer advertising. Affordability also received the highest score (2/2) because cost savings consistently emerged as one of the company's primary consumer benefits. Additionally, identity received a strong score (2/2), reflecting the company's recent strategic shift toward emphasizing style discovery and evolving personal fashion. Low scores were assigned to entrepreneurship (0.33/2) and community (0.25/2). The trust also received a moderate assessment (0.75) due to the presence of buyer protections but relatively little emphasis on it as part of branding. (See Table C below)

Sustainability emerged as ThredUp's defining strategic priority. The company's mission "to inspire the world to think secondhand first" ("2024 Impact Report") places environmental responsibility at the center of its brand identity. Sustainability is consistently reinforced throughout the company's website, annual reports, and its highly influential annual Resale Report ("ThredUp Resale Report"), which quantifies the environmental benefits of extending garments' lifecycles ("ThredUp About") and positions fashion resale as a key component of the circular economy. Unlike several competitors that discuss sustainability primarily within corporate responsibility materials, ThredUp makes environmental stewardship a key component of consumer messaging ("Your spring closet spruce up") ("it doesn't need to be Earth Month"). Importantly, ThredUp markets resale as a way for consumers to purchase quality clothing at substantially lower prices than traditional retail, making affordability a key driver for users. As such, ThredUp targets value-conscious shoppers seeking pragmatic financial savings rather than aspirational luxury consumers seeking status. By linking affordability with sustainability, the company presents secondhand shopping not as a compromise but as a smarter purchasing decision that benefits both consumers and the environment.

Although sustainability and affordability remain the company's core positioning, ThredUp's more recent branding increasingly emphasizes identity and self-expression. Newer marketing campaigns highlight the enjoyment of thrifting ("y'all know must much I love a good vintage find"), discovering unique pieces, and assembling outfits that reflect an individual's evolving style. Rather than portraying secondhand shopping solely as an ethical or economical choice, these campaigns present resale also as an enjoyable experience centered on creativity and personal expression, encouraging customers to "thrift their next chapter" ("Thrift Your Next Chapter"). This strategic shift toward complementing functional value propositions with more emotional and identity-based appeal reveals that ThredUp has recognized the multidimensional nature of consumer preferences among younger shoppers.

Entrepreneurship occupies a considerably smaller role within ThredUp's positioning. While the company encourages consumers to "turn extra clothes into extra cash" ("Sell your clothes with ThredUp") this messaging focuses primarily on decluttering wardrobes rather than building businesses. The community similarly receives limited emphasis. Unlike the peer-to-peer marketplaces, ThredUp does not allow users to follow one another, exchange messages, or engage through comments and likes. Although the company occasionally partners with influencers and brands, such as Athleta and Madewell, and operates a modest ambassador program, these initiatives are not core to its business strategy. Finally, while the company provides generous buyer protections through flexible return policies, and luxury branded merchandise undergoes additional authentication through ThredUp LUXE, authentication and seller ratings receive little marketing attention since consumer trust in ThredUp is expected because of its retailer-managed business model.

ThredUp presents sustainability, affordability, and identity as mutually reinforcing, illustrating how resale platforms can simultaneously address a variety of needs, thereby enabling resale to appeal to a broader cross-section of consumers.

The RealReal

Founded in 2011, The RealReal pioneered authenticated online luxury consignment by creating a full-service marketplace for pre-owned designer fashion, jewelry, watches, art, and home goods. The RealReal operates as a consignment intermediary, responsible for authenticating, photographing, pricing, marketing, and shipping every item on behalf of consignors. This model relies on a high degree of curation which differentiates the company from much of the resale industry and reflects its mission to "empower consignors and buyers to extend the life cycle of luxury goods in a way that honors luxury brands." As such, The RealReal's competitive positioning rests primarily on the ideas of trust through authentication and accessible luxury, with other factors such as sustainability and identity serving as secondary reinforcement.

Looking across the six themes reviewed in this study, The RealReal received the highest possible scores for trust (2/2), sustainability (2/2), affordability and attainability (2/2), and identity (2/2). By contrast, both entrepreneurship and community received the lowest possible

scores (0/2). (See Table D below) With trust and luxury as the cornerstones of its market positioning, The RealReal has carved out a unique space in the industry, competing by reducing uncertainty in luxury resale while also making premium brands more attainable.

Unlike other resale platforms where trust functions primarily as marketplace infrastructure, authenticity is the central promise around which The RealReal has built its brand, promising that The RealReal "Always Keeps it Real" ("Authentication"). Even its name communicates this commitment, ensuring buyers that they are purchasing real luxury items. From an operational perspective, every item is authenticated by teams of luxury experts using proprietary technologies, including XRF machines for precious metals, the company's patented TRR Shield scanning technology, and TRR Vision microscopic imaging systems. Buyer protections further reinforce this positioning through guaranteed authenticity, return policies for many product categories, seller protections against lost or damaged merchandise, and a consignment model that anonymizes buyers and sellers, reducing fraud risk. The RealReal's new "The Last Real Man" campaign emphasizes their commitment to authentication especially in the age of AI ("Last Real Man").

The RealReal's second defining theme is affordability, although affordability is framed not as low cost but as access to premium brands that might otherwise remain financially out of reach. The company describes modern luxury as increasingly "accessible" and "inclusive." Advertising promotes the opportunity to purchase designer brands at significantly reduced prices while maintaining the exclusivity and prestige traditionally associated with luxury fashion ("Jackie the Talking Gucci Bag"), suggesting that The RealReal seeks aspirational shoppers seeking luxury at attainable price points, not consumers who are simply bargain hunting. Also, The RealReal's 2025 Resale Report ("RealReal 2025 Resale Report") suggests that investment in luxury items is financially prudent as accessories should be viewed as assets whose value compounds over time.

Identity and self-expression likewise received the highest possible score, though the platform frames identity differently than platforms like Depop or Poshmark. Whereas those companies emphasize originality, creativity, and individual style, The RealReal associates identity with luxury, refinement, and timeless fashion. Advertising campaigns such as "A Good Gift" ("Good Gift") and broader branding encourage consumers to discover distinctive designer pieces that reflect taste and sophistication. Additionally, The RealReal personalizes luxury by promising consumers that luxury is theirs to define ("Spring 2023 Campaign").

Sustainability is deeply embedded within the company's corporate identity as the founder describes the platform as a response to fashion's wasteful, disposable consumption. The RealReal quantifies its environmental impact through extensive statistics, including the number of products kept in circulation and estimates regarding how consigning displaces newly manufactured purchases ("RealReal About") ("RealReal Social Impact"). It integrates environmental messaging in both corporate communications and consumer-facing marketing in a manner that complements the company's primary emphasis on trust and luxury in order to strengthen the legitimacy of resale while also reinforcing the enduring value of luxury goods.

Entrepreneurship received the lowest scores because The RealReal does not encourage users to build resale businesses or pursue selling as a side hustle. Instead, consigning is consistently framed as a means of refreshing one's wardrobe or financing future luxury purchases rather than generating independent income as seen in their campaign that encourages consigners to "sell to shop, shop to sell" ("It's Resale"). Moreover, because The RealReal manages pricing, authentication, photography, shipping, and customer interactions, consignors receive neither the business tools nor the direct customer engagement found on peer-to-peer marketplaces. Similarly, community received the lowest score since the platform lacks social features such as following users, commenting, messaging, or public seller profiles, while affiliate programs and occasional pop-up events play only a minor role within its broader business model, as The RealReal instead favors delivering a highly curated, retailer-like shopping experience.

Rather than fostering a participatory marketplace where users simultaneously consume and sell fashion, The RealReal functions as a luxury intermediary that reduces risk, preserves brand prestige, and broadens access to high-end goods. Its positioning, relative to its mass-market peer-to-peer resale platform competitors, illustrates an alternative approach within the circular fashion industry.

Vinted

Founded in 2008 in Lithuania, Vinted has grown into a large peer-to-peer fashion resale platform by positioning resale as an everyday alternative to traditional retail and a practical way to save money and earn income while reducing environmental impact. As such, Vinted's competitive positioning rests most strongly on sustainability and affordability, with other themes playing limited roles.

Based on this study's methodology, Vinted received the highest possible rating for sustainability (2/2) and affordability (2/2) because environmental responsibility is consistently emphasized across its corporate website, impact reports, and consumer advertising, alongside extensive communication about consumer savings. Entrepreneurship also emerged as a major strategic priority, receiving a high score (1.67/2) as the platform encourages users to earn money through resale and promotes fee-free selling. The community received a moderate score (0.75/2) because most interaction is transactional rather than social. Trust receives a moderate score (1/2), reflecting the presence of buyer protections, public seller ratings, and optional authentication. Finally, identity received a low score (0/2). (See Table E below)

Vinted has made sustainability one of its clearest strategic priorities, maintaining a dedicated sustainability section on its website ("Vinted Equation") and publishing extensive environmental impact reporting. Its 2026 Impact Report ("Vinted Impact Report") states that Vinted members avoided approximately 1,607 kilotons of carbon-dioxide-equivalent emissions in 2025 and that 76% of purchases made through the platform displaced the purchase of a new item. The company also reports efforts to increase low-carbon delivery vehicles, expand renewable-energy use across its offices, warehouses, and data centers, reduce emissions in accordance with the Science Based Targets initiative, and achieve net-zero emissions by 2050.

Additionally, Vinted incorporates environmental themes into marketing campaigns, challenging overconsumption in its "Too Many?" ad ("Too Many?") and encouraging recirculation in its "New Again" spot ("New Again").

At the same time, Vinted links sustainability with economic value. The company explicitly states that "sustainability shouldn't be more expensive" ("Vinted Impact Report" 5) and quantifies the financial benefits users receive through its marketplace. According to its reports, members saved €21.6 billion on fashion in 2025 ("Vinted Impact Report" 6), paid an average of 72% less than original retail prices, and would have spent approximately 2.6 times more had they purchased those items ("Vinted Impact Report" 7). Additionally, 59% of members reported that their overall fashion spending had decreased since joining the platform. Its target market therefore appears to be budget-conscious consumers ("Vinted Impact Report" 7). Messaging about economic value is equally prominent on the selling side of the marketplace. Vinted's impact reporting states that sellers collectively earned €10.8 billion in 2025 ("Vinted Impact Report" 18) and portrays the platform as a financial resource during major life transitions, including having children, moving homes, or preparing for a new school year. The report further notes that many members use their Vinted earnings or savings to offset essential household expenses such as food and bills. This messaging differs from other platforms' emphasis on entrepreneurship as personal ambition, instead focusing on the unlocking economic value from clothing sellers already own. The platform supports this entrepreneurism by eliminating seller fees, a feature prominently highlighted across its advertising campaigns ("Naomi"). It also allows sellers to notify interested buyers with private offers, create bundle discounts, and reserve items for individual purchasers. However, beyond transaction histories and basic notifications, Vinted provides limited seller analytics or business management tools. Therefore, the target seller is an individual seeking supplemental income not someone interested in developing a robust selling business.

Identity in terms of fashion discovery and self-expression as defining benefits is not featured in the way it is for competitor platforms. Also, community is not a component of Vinted's branding as user interactions are largely functional and the company focuses on facilitating efficient exchanges between buyers and sellers. Trust similarly functions as marketplace infrastructure rather than a source of competitive differentiation, as Vinted offers many transactional safeguards including optional paid authentication for eligible designer items, secure payment processing that holds funds until buyers confirm receipt, encrypted transactions, fraud detection systems, and publicly visible seller ratings with written feedback ("We take your safety seriously"), but does not emphasize these in advertising materials.

Vinted's focus on practical circularity creates an internal cycle of selling and buying as seen in its "More Space" ad ("More Space"). As such, Vinted occupies a distinctive position within the resale market, not attempting to convey cultural or entrepreneurial aspiration, or luxury orientation, but normalizing resale as a financially prudent and environmentally responsible everyday habit.

Vestiaire Collective

Vestiaire Collective was founded in Paris in 2009. It focuses almost exclusively on designer apparel, handbags, jewelry, and accessories, presenting resale as a sophisticated and environmentally responsible way to participate in high-end fashion. As such, Vestiaire Collective's competitive positioning centers most strongly on identity and self-expression, closely followed by sustainability, with the platform presented as a fashion destination.
In this study's assessment, Vestiaire Collective received the highest possible score for identity and self-expression (2/2), reflecting the overwhelming emphasis on fashion trends, personal style, and curated luxury across the company's website and social media presence. Sustainability also received the maximum score (2/2) because circular fashion is consistently embedded throughout the company's corporate and consumer messaging. Affordability received a moderate score (1/2), as the company acknowledges discounted luxury prices but does not frame savings as a primary consumer benefit. Trust also received a moderate score (1/2), reflecting authentication services, buyer protections, and seller performance badges, although trust is less central to the company's branding than expected for a luxury marketplace, especially in comparison to The RealReal. Entrepreneurship received a limited score (0.33/2) because it also is not emphasized. (See Table F below)

The overwhelming majority of Vestiaire Collective's social media content focuses on outfit inspiration, seasonal fashion guides, celebrity wardrobes, designer trends, and curated styling advice, emphasizing discovering distinctive luxury pieces ("Baguette"), following emerging fashion trends, and cultivating personal style. As such, the company's identity-driven positioning suggests that it seeks to attract fashion enthusiasts.

Closely intertwined with this emphasis on style is Vestiaire Collective's commitment to sustainability. The company maintains a dedicated sustainability section within its website, publishes annual impact reports ("Shaping a Circular Future"), and communicates measurable environmental outcomes associated with purchasing pre-owned luxury goods, reporting that choosing a pre-owned item instead of buying new can reduce environmental impact by up to 90%, while 87% of purchases on the platform replace the purchase of a newly manufactured item. Vestiaire Collective explicitly critiques fast fashion, describing it as a driver of overconsumption, overproduction, and waste, making environmentalism core to its messaging ("What's rarer than a Himalayan Birkin?") ("Think First, Buy Second").

Affordability is positioned within Vestiaire Collective's value proposition in a nuanced manner. As a luxury marketplace, the platform provides access to premium designer brands at prices below traditional retail, promoting discounts of up to 70% alongside price-drop notifications and negotiated offers. However, affordability functions primarily as a mechanism for increasing access to luxury fashion and premium brands at attainable entry points, not bargain shopping.

For Vestiaire Collective, entrepreneurship functions primarily as a mechanism for refreshing wardrobes and recirculating luxury goods. Community is also not a key differentiator as users cannot follow one another, and interaction is largely confined to transaction-related

messaging and comments on individual listings. Although Vestiaire Collective maintains influencer programs, celebrity-curated collections, and periodic digital events such as Vestiaire Vibes ("Vestiaire Collective Launches 'Vestiaire Vibes'"), these marketing promotions function to enhance the company's image as a curated fashion marketplace not a social platform. Finally, despite its luxury nature, Vestiaire Collective received a moderate score on trust. It offers multiple authentication services, buyer protections, escrow-based payment processing, tracked shipping, and seller performance indicators including Trusted Seller and Expert Seller badges, however, authentication is not universal, direct-shipping purchases rely primarily on AI-assisted verification, and standard written seller reviews are absent. Ironically, the company's About page states that "trust has been at the heart of everything we do," yet trust occupies a less prominent place within consumer-facing advertising than style or sustainability, resulting in trust-oriented features being cast as essential, not differentiating, features of luxury resale.

By integrating luxury style, trend awareness, and sustainability into a unified value proposition, Vestiaire Collective distinguishes itself from other platforms by seeking to convince consumers that purchasing pre-owned fashion signals their sophisticated, fashion-forward taste.

5. DISCUSSION

Each fashion resale platform analyzed in this study has developed a distinct strategic position by emphasizing different combinations of themes of sustainability, affordability, identity and self-expression, entrepreneurship, community, and trust to appeal to the different consumer motivations that resonate most strongly with their target audiences. (See Table G below)
While each company presents a relatively unique value proposition, fashion resale platforms can be understood through a summarized 2x2 framework of strategic positioning where each axis is a continuum. (See Figure 1 below) The x-axis reflects the "Consumer Value" that the platform generates for its users, ranging from "Functional Value" to "Symbolic/Aspirational Value." Platforms positioned toward the functional end primarily appeal through tangible benefits such as affordability, sustainability, financial value, and reduced transaction risk, whereas those toward the symbolic end emphasize identity, style, prestige, uniqueness, and access to fashion as a form of self-expression. The y-axis reflects the platform's "Consumer Relationship Model," ranging from "Transaction-Oriented" to "Participation-Oriented". Transaction-oriented platforms stress their efficient facilitation of the purchase or sale of goods, while participation-oriented platforms encourage consumers to take a more active role as sellers and community members. Together, these dimensions create four distinct strategic archetypes: Quadrant I: "Resale as Lifestyle" (Symbolic/Aspirational + Participation-Oriented), Quadrant II: "Participatory Resale" (Functional + Participation-Oriented), Quadrant III: "Practical Resale" (Functional + Transaction-Oriented), and Quadrant IV: "Curated Aspiration" (Symbolic/Aspirational + Transaction-Oriented). (See Figure 2 below) This categorization reflects differences in strategic emphasis rather than mutually exclusive characteristics as every company offers some combination of functional, symbolic, transactional, and participatory value, but the framework

identifies which dimensions most strongly define a company's positioning. The six companies examined in this study distribute across these four quadrants.

Figure 1: Conceptual Framework for Fashion Resale Platform Strategic Positioning – Consumer Value (x-axis) v. Consumer Relationship Model (y-axis)

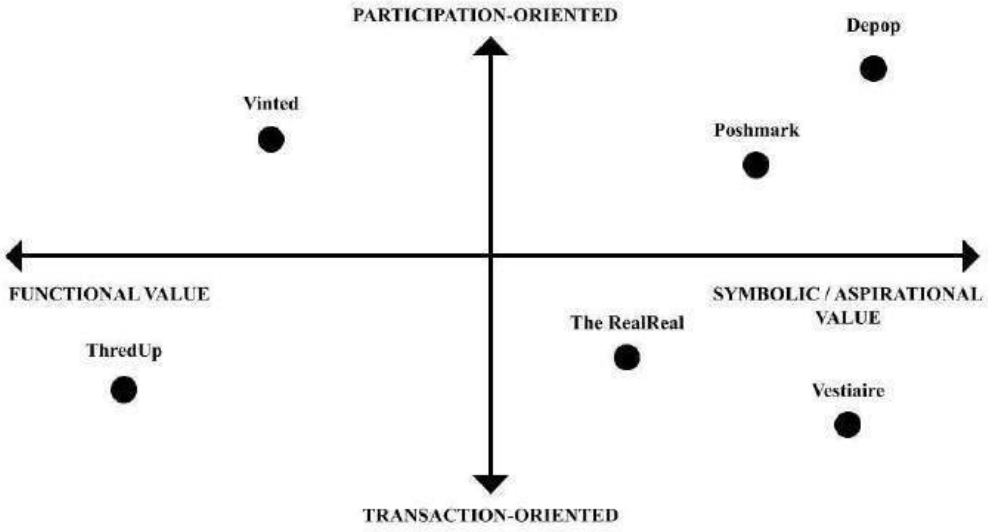
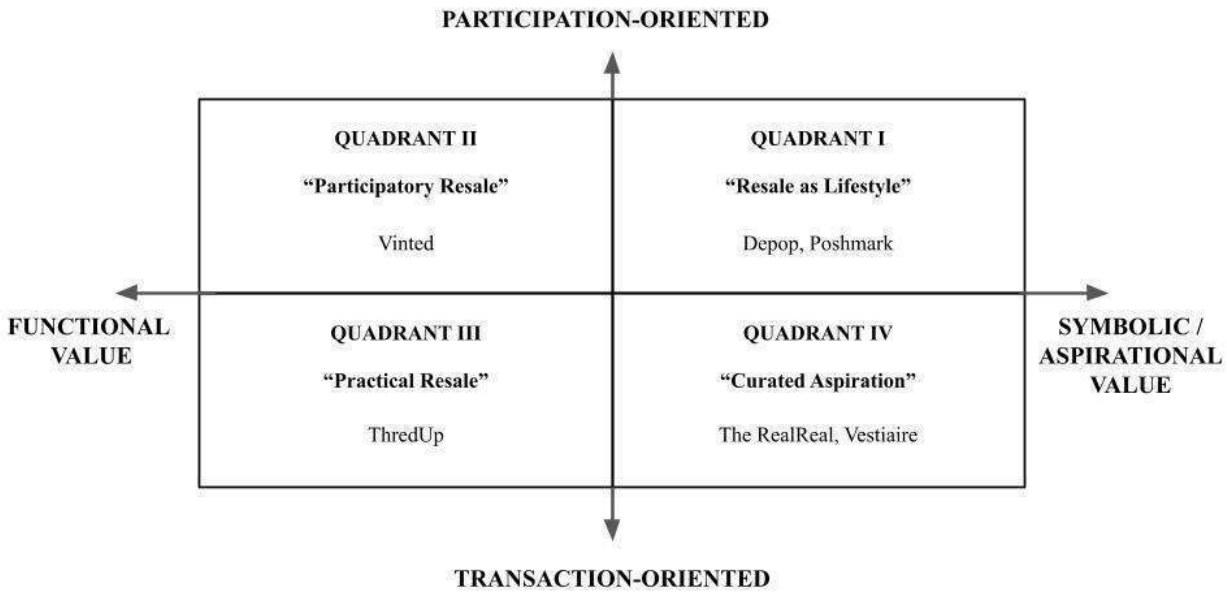


Figure 2: Strategic Archetypes for Fashion Resale Platform Strategic Positioning



Depop and Poshmark occupy the "Resale as Lifestyle" quadrant, where secondhand consumption serves as a means of identity expression and there is active marketplace participation among users. Depop's version of this archetype combines three dominant themes

including identity, entrepreneurship, and community fostered through a broader social culture surrounding resale. Poshmark also falls within this quadrant because identity and entrepreneurship are its two strongest themes, but it is positioned somewhat closer to the transaction-oriented midpoint than Depop because although users are encouraged to engage in the side-hustle economy, community is less central to Poshmark's positioning.
Vinted falls within the "Participatory Resale" quadrant because sustainability and affordability create a predominantly functional value proposition. However, its positioning also includes entrepreneurship as the company urges its customers to both purchase secondhand goods and also to monetize their existing wardrobes.

ThredUp occupies the "Practical Resale" quadrant because sustainability and affordability emerged as two of its strongest themes, making secondhand shopping economically and environmentally advantageous, with identity increasingly present as captured by its central location on the x-axis.

Finally, The RealReal and Vestiaire Collective appear in the "Curated Aspiration" quadrant, although for somewhat different reasons. The RealReal combines trust with affordability and accessibility to make authentic luxury attainable. Its functional mechanisms of authentication and price advantage enable the symbolic benefit of luxury consumption, depicted by its more central x-axis location relative to Vestiaire Collective. Vestiaire Collective sits further toward the symbolic end because identity and sustainability dominate its positioning, with fashion-mindset, personal style, and participation in luxury culture forming a larger part of their value proposition. Neither makes entrepreneurship or user community central enough to move into the participation-oriented upper half of the framework.

This framework demonstrates how online resale platforms differentiate themselves along two fundamental strategic questions: 1) what kind of value does resale represent to the consumer, and 2) what role does the consumer occupy within the resale experience? As a result, the six companies analyzed in this study have constructed substantially different competitive positions.

Synthesis

Taken together, the results reveal that the six companies do not compete by attempting to maximize every consumer value associated with resale but instead, strategically emphasize a set of coherently combined, prioritized consumer motivations while treating others as supporting benefits or simply marketplace requirements. The differences in emphasis across companies suggest that the presence of a consumer benefit alone does not constitute strategic positioning but rather it is the relative prominence of and meaning assigned to that benefit in consumer communications. These results reinforce the multidimensional nature of resale as identified in the literature while demonstrating how firms respond selectively to that multidimensionality. As such, there is no single winning resale model. The six firms examined in this study occupy different strategic spaces because resale satisfies heterogeneous consumer motivations, creating room for multiple viable competitive positions.

6. CONCLUSION

This study reviewed the competitive positioning of six notable fashion resale platforms and revealed how they seek to offer differentiated values that support strategic combinations of key customer values. Each company was analyzed and scored for the six most important values, called themes. The results demonstrate that although the six platforms operate within the same broader fashion resale market, they construct substantially different competitive positions by selectively emphasizing particular combinations of consumer motivations rather than attempting to maximize every source of value. This analysis was then synthesized in a new conceptual framework to understand how these platforms' positioning map across the industry, looking at "Customer Value" and "Customer Relationship Model". As a result, four distinct archetypes emerged, providing a lens through which to categorize platforms in this segment.

Depop and Poshmark emphasize identity and entrepreneurial participation, placing them within the "Resale as Lifestyle" archetype. Vinted combines functional benefits such as sustainability and affordability with active seller participation, representing the "Participatory Resale" archetype. ThredUp primarily combines sustainability and affordability within a more transaction-oriented "Practical Resale" archetype. Finally, The RealReal and Vestiaire Collective occupy the "Curated Aspiration" archetype, where resale provides access to the symbolic and aspirational benefits of fashion and luxury, while maintaining a more transaction-oriented relationship with consumers. Importantly, these archetypes represent differences in prioritization within a firms' marketing approach to create coherent value propositions for target customers. Limitations of the approach taken in this study are two-fold. First, it relies exclusively on publicly available materials. Second, despite the use of standardized criteria, ratings in the study are based on interpretive judgment. For simplicity, all sub-scores within a theme were treated as equally important, with no weightings assigned. In most situations, this streamlined approach yielded appropriate results, however, in one theme for one company, as indicated in the Results section for Poshmark, there was a need to prioritize the impact of customer-facing messaging above the other factors addressed by that theme's sub-questions, consequently modifying that platform's stated positioning. Nevertheless, the study's comparative design makes it possible to identify how major resale platforms differentiate themselves.

As the secondary fashion market continues to mature, competitive advantage will increasingly depend on effective communication of the differentiated value propositions a given platform offers. Traditional fashion brands and retailers are also incorporating resale into their own business models, signaling that secondhand commerce is becoming integrated into mainstream retail and is no longer a separate, peripheral market. This evolution suggests that today's consumers evaluate brands not only on intrinsic product characteristics but also on alignment with their multidimensional values. As a result, pressure can potentially be created on how firms across industries design, market, and price their products and engage with consumers. Additionally, growth of fashion resale platforms reflects a shift toward more participatory, digitally mediated marketplaces, where the distinction between consumers and sellers is

increasingly blurred, potentially giving rise to new sources of economic activity that will force businesses to redefine where and how they create consumer value.

Future research should investigate how established brands and retailers are evolving to utilize resale in their businesses, and what associated financial, operational, branding, and customer demographic impacts that is having. It could also examine whether the strategic archetypes identified in this study extend to brand-operated resale programs or whether integration of resale into traditional retail produces different forms of strategic positioning. Longitudinal research would be additionally useful for determining whether as resale programs mature over time, they function merely as a complementary component of core traditional retail or if continued expansion serves to reshape those firms' broader business models.

Works Cited

Antonio, Astrid. "With the Help of Gen-Z, Sustainable Fashion Is Becoming Big Business." South Jersey Climate News, 22 Oct. 2026, https://sjclimate.news/6684/features/with-the-help-of-college-students-thrifting-and-sustainable-fashion-are-going-digital/.

Blaz, Claudia Lorena Cárdenas, et al. "Second-Hand Goes Digital: A Systematic Literature Review of Consumer Motivations and Barriers in Online Resale Platforms." The International Review of Retail, Distribution and Consumer Research, 18 May 2026, pp. 493-517, https://doi.org/10.1080/09593969.2026.2674160.

Bolshaw, Elizabeth. "How Will Your Decisions Today Shape the Future for Generations to Come?" Ernst & Young, 16 Jan. 2025, www.ey.com/en_gl/innovation-realized/how-will-your-decisions-today-shape-the-future-for-generations-to-come.

Butler, Sarah. "AI and US Tariffs Expected to Fuel Fresh Surge in Secondhand Fashion Sales." The Guardian, 19 Mar. 2025, www.theguardian.com/business/2025/mar/19/ai-and-us-tariffs-expected-to-fuel-fresh-surge-in-secondhand-fashion-sales.

Cardoni, Vanessa. "Thrift vs. Fast Fashion: Why Gen Z Is Choosing Secondhand - Goodwill Industries International." Goodwill Industries International, 27 Feb. 2025, www.goodwill.org/blog/shop/thrift-vs-fast-fashion-why-gen-z-is-choosing-secondhand/.

Chew, Naomi. "Why Thrifting Is Surging among Gen Z." Hanabi, 2025, https://hanabi.asij.ac.jp/features/why-thrifting-is-surging-among-gen-z/.

"2025 Gen Z and Millennial Survey." Deloitte, 2025, www.deloitte.com/global/en/issues/work/genz-millennial-survey.html.

Depop, "About." https://news.depop.com/who-we-are/about/.

"Depop Environmental Impact Measurement Methodology." Depop, November 2023, https://assets.ctfassets.net/itoh30v6uh9a/6cqSGab95TyWuv3tYVWnPI/23b2047d8365a2274a65dd8bf13437e4/Environmental-Impact-Measurement-Methodology.pdf

"Depop's Climate Impact Plan." Depop Newsroom, 2 Jun 2026, https://news.depop.com/what-we-stand-for/sustainability/depop-takes-action-on-climat

Depop. "Depop Commercial 2026 – Depop Hat New Advert." YouTube, 1 min, 2 sec., https://www.youtube.com/watch?v=65FbZns9hUY.

"Depop on Campus: bringing circularity to your college." Depop, 16 Jun 2025, https://www.depop.com/blog/depop-on-campus.

Depop. "Depop – Where Taste Recognizes Taste." YouTube, 1 min, 2 sec., https://www.youtube.com/watch?v= 4sJeD-0upQg.

Depop. "Easy to Sell." Instagram, February 17, 2026, https://www.instagram.com/reel/DU3J2QUjP-W/.

Depop, "Impact." https://news.depop.com/impact/.

"Depop Partners with Gov Ball as the Style Destination for Festival Season; Shop the In-App Edit and Find Us Onsite." Depop Newsroom, 2 Jun 2026,

https://news.depop.com/company-news/depop-partners-with-gov-ball-as-the-style-destination-for-festival-season-shop-the-in-app-edit-and-find-us-onsite/.

Depop. "Shop resale not retail." Instagram, April 30, 2024, https://www.instagram.com/reel/C6ZJcXBClG9/.

Dockterman, Eliana. "How Depop Made Sustainable Shopping Cool for Gen Z." Time, 19 Aug. 2021, https://time.com/6089003/depop-upcycling-fashion/.

"A New Textiles Economy: Redesigning Fashion's Future." Ellen MacArthur Foundation, 27 Nov. 2017, www.ellenmacarthurfoundation.org/a-new-textiles-economy.

Escalante, Ana. "Save the Earth, but Make It Fashion: Gen Z's Secondhand Shopping, Explained." Who What Wear, 27 Apr. 2025, www.whowhatwear.com/fashion/shopping/gen-z-sustainability-report.

Hall, Izzy. "Sustainability, Reselling, and Second-Hand Shopping: Gen Z's Biggest Obsession." Pion, 2024, www.wearepion.com/blog-posts/sustainability-reselling-gen-zs-obsession.

Hanbury, Mary. "Gen Z Is Fueling the Growth of a New Breed of Secondhand Selling Apps." Business Insider, 5 July 2019, www.businessinsider.com/gen-z-fuels-growth-of-new-secondhand-selling-apps-2019-6.

"How to Become a Posh Ambassador." The Poshmark Blog, https://blog.poshmark.com/how-to-become-a-posh-ambassador/.

McFall-Johnsen, Morgan. "These Facts Show How Unsustainable the Fashion Industry Is." World Economic Forum, 31 Jan. 2020, www.weforum.org/stories/2020/01/fashion-industry-carbon-unsustainable-environment-pollution/

"The State of Fashion 2026: When the Rules Change." McKinsey & Company, 17 Nov. 2025, www.mckinsey.com/industries/retail/our-insights/state-of-fashion.

Moizant, Fanny, et al. "Resale's next Chapter: How Fashion and Luxury Brands Can Win in the Secondhand Market." BCG Global, 9 Oct. 2025, www.bcg.com/publications/2025/how-fashion-luxury-brands-can-win-secondhand-marke

"New Feature Alert: Rate Your Purchase." The Poshmark Blog, 27 Jun 2014, https://blog.poshmark.com/2014/06/27/new-feature-alert-rate-your-purchase/.

Park, Seoryung. "Th(r)ifted: Bridging the Rift in Fashion's Evolution to Save the Earth." Fashion Law Journal, 2 Sept. 2024, https://fashionlawjournal.com/thrifted-bridging-the-rift-in-fashions-evolution-to-save-the-earth/.

Pinto, Katherine and Royo-Vela, Marcelo. "Understanding Second-Hand Clothing Consumption: A Literature Review and Proposed Conceptual Model." MDPI, Sustainability, 2026, 18, 4795, pages 1-19, https://doi.org/10.3390/su18104795.

Poshmark, "About." https://poshmark.com/about/.

Poshmark, "Posh Authenticate." https://poshmark.com/posh_authenticate?srsltid=AfmBOop2I4_aWfnXlGTLoY6aXI0kl9eluHVAQBanfpsyX-f8qX5eMsru.

"Posh for the Planet: Invest in Secondhand, Invest in Your Future." The Poshmark Blog, 18 Apr 2022, https://blog.poshmark.com/2022/04/18/posh-for-the-planet-invest-in-secondhand-invest-in-your-future/.

Poshmark. "Girl, Quit Daydreaming." TV spot, 02 Mar 2026, https://www.ispot.tv/ad/gT7F/poshmark-girl-quit-daydreaming.

Poshmark, “Posh Protect.” https://poshmark.com/posh_protect/.
Poshmark. “Start Cashing In: Luxury Goods.” TV spot, 7 Oct 2024, https://www.ispot.tv/ad/fDhI/poshmark-start-cashing-in-luxury-goods.
Poshmark. “The Perfect Side Hustle: Smart List AI.” TV spot, 25 Mar 2026, https://www.ispot.tv/ad/TpED/poshmark-the-perfect-side-hustle-smart-list-ai.
Poshmark. “Vibe, Check.” TV spot, 16 Mar 2026, https://www.ispot.tv/ad/gGz2/poshmark-vibe-check-song-by-byron-juane.
Poshmark. “Vibe, Check: Every Era.” TV spot, 18 Mar 2026, https://www.ispot.tv/ad/gKc2/poshmark-vibe-check-every-era.
“The Power of Secondhand: How Resale Slows Consumption – Depop Displacement Research.” Depop, 2022, https://news.depop.com/download/docs/DepopDisplacementResearch2022.pdf.
Prisco, Anna, et al. “Sustainability-Driven Fashion: Unpacking Generation Z’s Second-Hand Clothing Purchase Intentions.” Journal of Retailing and Consumer Services, vol. 85, July 2025, p. 104306, https://doi.org/10.1016/j.jretconser.2025.104306.
The RealReal, “About.” https://therealreal.com/about/.
The RealReal, “Authentication.” https://therealreal.com/authentication/.
The RealReal. “A Good Gift.” TV spot, 11 Nov 2025, https://www.ispot.tv/ad/BJtc/the-realreal-a-good-gift.
The RealReal. “Jackie the Talking Gucci Bag.” TV spot, 4 Feb 2026, https://www.ispot.tv/ad/gtjz/the-realreal-jackie-the-talking-gucci-bag.
The RealReal. “The Last Real Man.” TV spot, 6 Apr 2026, https://www.bestadsontv.com/ad/190089/The-RealReal-The-Last-Real-Man.
The RealReal. “It’s Resale: First Time Consigners.” TV spot, 3 Jul 2026, https://www.ispot.tv/ad/g0PV/the-realreal-its-resale-first-time-consignors.
“The RealReal 2025 Resale Report.” The RealReal, 2025, aL8cH2GNHVfTOy-t_TheRealReal2025ResaleReport-1-.pdf,
The RealReal, “Social Impact.” https://investor.therealreal.com/social-impact/.
The RealReal. “Spring 2023 Campaign – Luxury Is Yours to Define.” YouTube, 30 seconds., https://www.youtube.com/watch?v=a71pz6imXdI.
Roshitsh, Kaley. “The Resale Market: Who’s Playing, Who’s Leading, Who’s Emerging.” Women’s Wear Daily, 23 Aug. 2021, https://wwd.com/sustainability/business/feature/ebay-etsy-mercari-thredup-therealreal-clothing-resale-players-market-1234894315/.
Schenkman, Lauren. “US Secondhand Market Projected to Reach $73B by 2028: Report.” Fashion Dive, 27 Mar. 2024, www.fashiondive.com/news/thredup-2024-resale-report-global-data/711500/.
“Shaping a Circular Future: Our 2025 Report.” Vestiaire, 2025, https://assets.vestiairecollective.com/documents/impact_report_2025.pdf.
Szűcs, Róbert Sándor and Földi, Kata. “Second-Hand Clothes and Sustainability: A Literature Review.” Evfolyam, vol. XIII, Special Issues 1, 15 Aug. 2025, pages 30–40, https://doi.org/10.24387/CI.SI.2025.1.5.
Sicurella, Savannah. “When Second Hand Becomes Vintage: Gen Z Has Made Thrifting a Big Business.” NPR, 18 June 2021,

www.npr.org/2021/06/18/1006207991/when-second-hand-becomes-vintage-gen-z-has-made-thrifting-a-big-business.
Thanasi-Boce, Marsela and Al-Issa, Nermain. “The Secondary Luxury Market: A Systematic Review and Theoretical Integration.” Cogent Business & Management, vol. 12, no. 1, 9 July 2025, pages 1-30, https://doi.org/10.1080/23311975.2025.2523405.
“2024 Impact Report.” ThredUp, 2024, https://ir.thredup.com/static-files/26045ed8-c4f9-4b15-8c01-6257da8ba8b0.
ThredUo, “About.” https://thredup.com/about/.
ThredUp. “it doesn’t need to be Earth Month to know that reducing closing waste AND getting paid for it is a good thing to do.” Facebook, 30, Apr 2026, https://www.facebook.com/thredup/videos/1502039988115104/.
“ThredUp Resale Report.” ThredUp, 2025, https://cf-assets-tup-thredup.com/resale_report/2025/ThredUp_Resale_Report_2025.pdf.
ThredUp. “Sell your clothes with ThredUp – Turn extra clothes into extra cash or store credit.” Facebook, 30 Jul 2025, https://www.facebook.com/thredup/videos/1263123682509981/.
ThredUp. “Thrift Your Next Chapter.” Facebook, 13 Jan 2026, https://www.facebook.com/thredup/videos/1215304823872344/.
ThredUp. “y’all know how much I love a good vintage find…this is ur sign to start shopping on @thredup.” Facebook, 20 Mar 2026, https://www.facebook.com/thredup/videos/2140193896824200/.
ThredUp. “Your spring closet spruce up made simple and sustainable with the @thredup clean out bag.” Facebook, 30 Apr 2026, https://www.facebook.com/thredup/videos/1555772819499175/.
Vestiaire. “Baguette” Instagram, 31 Jul 2026, https://www.instagram.com/p/DbdpOxgDXQV/?hl=en&img_index=1.
“Vestiaire Collective Launches ‘Vestiaire Vibes’.” Crash. https://www.crash.fr/vestiaire-collective-launches-vestiaire-vibes/.
Vestiaire. “Think First Buy Second” YouTube, 15 sec., https://www.youtube.com/watch?v=zfQFbG9n5i8.
Vestiaire. “What’s rarer than a Himalayan Birkin?” Instagram, 30 Jul 2026, https://www.instagram.com/p/DbbKk5sjZFE/?hl=en&img_index=1.
“Impact Report.” Vinted, 2025, https://press-center-static.vinted.com/Impact_Report_2025_ENG_11da0e2c91.pdf.
Vinted. “More Space.” TV spot, 20 Oct 2025, https://www.ispot.tv/ad/BsGi/vinted-more-space.
Vinted. “Naomi.” TV spot, 2 Jul 2026, https://www.ispot.tv/ad/g0m8/vinted-naomi.
Vinted. “New Again.” YouTube, 1 min., 20 sec., https://www.youtube.com/watch?v=RtRVJAr6YeI.
Vinted. “Too Many?.” YouTube, 45 sec., https://www.youtube.com/watch?v=yoVa6zgXM-c.
Vinted. “The Vinted Equation.” https://company.vinted.com/sustainability.
Vinted. “We take your safety seriously.” https://company.vinted.com/safety.
Waxman, Olivia B. “People Have Been Reusing Clothes Forever but Thrift Shops Are Relatively New. Here’s Why.” Time, 17 Aug. 2018, https://time.com/5364170/thrift-store-history/.
“Zara Larsson Brings Her Signature Style to Depop’s New Global Campaign.” Depop Newsroom, 16 July 2026, https://news.depop.com/company-news/zara-larsson-brings-her-signature-style-to-depops-new-global-campaign/w Global Campaign.

The Conceptual Frameworks of AI Productivity Paradox: A Theoretical Review

By Chau Nguyen

Abstract

Following the widespread adoption of artificial intelligence (AI) across the global economy in the early twenty-first century, numerous academic investigations into its impacts have been conducted, thereby providing robust empirical evidence of substantial AI productivity gains at the micro-level. However, these improvements are not matched by clear AI-driven growth trends in official macroeconomic statistics. This prompts modern economic literature to propose different theoretical explanations for this gap, also known as the modern Solow paradox or the "AI productivity paradox". In this review, disparate theoretical frameworks attempting to explain this phenomenon are identified and presented as possible mechanisms. Hence, it helps provide a more cohesive understanding of how micro-level efficiency gains have not transformed into measurable aggregate macroeconomic growth.

Keywords economics, artificial intelligence, general-purpose technology, productivity, economic growth, total factor productivity, AI productivity paradox

Introduction

The first two-and-a-half decades of the twenty-first century have witnessed revolutionary milestones in artificial intelligence (AI) development, most prominently the public release of large language models like ChatGPT in the early 2020s. This has facilitated the widespread adoption of AI across a diverse range of sectors, with 88% of organisations in the 2025 McKinsey & Company Survey reporting that they use AI in at least one business function (*Artificial Intelligence Index Report 2026* 193). This pervasiveness, along with its capability for continuous improvement and innovation spawning, has made AI widely regarded as the next General Purpose Technology (GPT), defined as a profound technological innovation with transformative effects on the entire economy (Calvino et al. 7).

Since then, many empirical studies have investigated the economic impacts of AI mass adoption, producing a robust empirical basis of substantial microeconomic productivity improvements. However, macroeconomic indicators, particularly gross domestic product (GDP) and aggregate total factor productivity (TFP), have not experienced transformative growth as expected from turning points in AI development. Observation of this phenomenon has led economists to believe that this is a modern reprisal of the "Solow paradox", which describes how US macroeconomic productivity growth slowed between the 1970s and 1980s despite massive investments in information technology (Brynjolfsson et al., Artificial Intelligence and the Modern Productivity Paradox 24). However, with more definitive micro-level evidence, this "AI productivity paradox" is better understood as the disconnect between microeconomic productivity gains and stagnant aggregate macroeconomic growth.

The modern economic literature has proposed different theoretical frameworks to explain the mechanisms driving this gap. However, the existing literature on the AI productivity paradox discusses these frameworks disparately, producing a scattered understanding of this phenomenon. To fill this gap, this theoretical review brings these fragmented models together as different explanations for the "AI productivity paradox". Thereby revealing a key tension between structural limitations and transitional frictions as two possible channels driving this paradox. Hence, it maps out a more cohesive insight into how micro-level productivity gains have not yet translated into aggregate macroeconomic growth.

The remainder of the paper is structured as follows. Section 2 establishes the empirical foundations that constitute the paradox. Section 3 describes the methodology of this review paper. Section 4 presents how different hypotheses and their theoretical support interpret the AI productivity paradox. Section 5 discusses the tensions that arise from the theoretical mapping of the previous section. Section 6 concludes with an outline of the research landscape produced from the descriptive synthesis in this review, as well as providing new implications and directions for future research.

Empirical foundations

Microeconomic productivity gains of AI are efficiency improvements – ranging from higher output per unit to operational cost savings – achieved at task, worker and firm levels from the AI implementation. Following the widespread enterprise adoption of generative AI tools in the early 2020s, new empirical studies have shown that it significantly boosts productivity at this micro-scale.

For instance, a controlled online experiment by Noy and Zhang found that ChatGPT can considerably raise mid-level professional writing performance, as task completion duration decreased by 0.8 standard deviations (SDs) and task output quality increased by 0.4 SDs (187). These outcomes illustrate that AI assistance can yield measurable task-level productivity improvements. Beyond tasks, the analysis of large real-world datasets also illustrates notable worker-level gains. Among the most prominent papers using this approach is Brynjolfsson et al., which found that support from generative AI improved labour productivity at customer-service agents by 15%, as measured by the number of issues solved per hour (23). This conclusion aligns with findings of Damioli et al., whose global-scale business dataset analysis inferred a positive correlation between firm-level AI patenting activities and worker productivity (1). Meanwhile, field research investigating firm-level productivity often faces methodological challenges in translating individual gains to broad corporate output, but a supportive body of evidence is growing (ILO 3). For example, a positive causality between AI adoption and firm-level total factor productivity was identified in Chinese manufacturing firms, with a 1% increase in penetration of the former resulting in a 14.2% increase in the latter. These results were proven to be rigorous even after being tested by endogeneity analysis and serial robustness checks (Gao and Feng 8934). This relationship mirrored the conclusions of a same-year German

enterprises survey, which established a strong positive association between AI adoption and firm productivity (Czarnitzki et al. 205).

To assess the macroeconomic impact of AI, economists primarily consider its effect on gross domestic product (GDP), a monetary measure of the total market value of all final goods and services produced in an economy, and aggregate total factor productivity (TFP), which measures how efficiently an economy transforms inputs to output (Sickles and Zelenyuk 100). In contrast to the substantial availability of solid microeconomic evidence, macroeconomic statistics on AI remain limited and show minimal improvements from this technological implementation. Parteka and Kordalska investigated this relationship using an augmented productivity growth framework embedded in global panel data on AI patents and bibliometrics (102764). Their results indicated that AI development has a negligible effect on aggregate productivity growth, characterising this technological innovation as a concentrated process rather than a large-scale driver for growth, at least in the near term (102774). Katz and Jung also reflected a similar weak impact on economic growth, estimating generative-AI-induced productivity gains to raise GDP by only 0.008% on average for sampled nations (410). Despite being consistent with the consensus, it noted that the situation may turn out differently in the long-term.

Observing a mismatch in evidence at the micro and macro levels, economists have noted a resemblance to the 1970s-1980s IT productivity paradox in the US, where massive IT investments met a macroeconomic productivity slowdown. Hence, this phenomenon is named the "AI productivity paradox", a modern version of the "Solow paradox" (Brynjolfsson et al., Artificial Intelligence and the Modern Productivity Paradox 24).

Methodology

This paper follows a theoretical review approach, which synthesises the existing literature to identify different theories on the mismatch between measurable microeconomic productivity and the absence of macroeconomic growth improvements. By presenting these frameworks as different explanations for the failure, this study aims to provide a more cohesive overview of the underlying drivers of the puzzling "AI productivity paradox".

To capture the current discourse on this phenomenon, a literature search was conducted across multiple databases, including Google Scholar, JSTOR, ScienceDirect, SSRN Scholarly Papers, and NBER Working Papers. The keywords included in these searches are "artificial intelligence", "productivity", "economic growth", "productivity paradox", and their variations (e.g. "firm-level productivity" or "total factor productivity". Given that artificial intelligence has only been foundationally integrated into the economy since the early 2010s and achieved economic prevalence in the early 2020s, the search was limited to papers published within the last 15 years to ensure the relevance of the discussion.

The search process resulted in 44 potential research papers. These sources were screened for ultimate use based on an assessment of their "quality". Specifically, pre-2025 working papers were checked to determine whether they had been published, while the ones uploaded from 2025

onwards were checked for the reputation of their repositories. For instance, the NBER Working Paper series held prestige within economics research, as a majority of its research drafts were authored by well-known field experts and eventually published in peer-reviewed journals. Sources that have passed this basic screening are prioritised for inclusion in the review. After this selection, the sample included 15 research papers, consisting of recently uploaded working papers from reliable repositories (e.g. NBER, IZA Institute of Labour Economics), articles from peer-reviewed journals, and published book sections.

Then, theoretical explanatory frameworks for the "AI productivity paradox" were identified and synthesised among the selected literature. They were classified into the four thematic classical hypotheses of this phenomenon, as proposed in (Brynjolfsson et al. 25), namely 1) false hopes, 2) redistribution, 3) mismeasurement, and 4) implementation lag.

Theoretical frameworks

False Hopes

The "False Hopes" hypothesis, as its name implies, posits that expectations about the economic potential of artificial intelligence have been fundamentally exaggerated. The premise argues that AI is capable of emulating human intelligence and has significantly improved performance of specific isolated tasks (e.g. coding, customer support). However, it would not be as economically transformative as breakthrough technologies (e.g. electricity, steam engine) of the previous Industrial Revolutions, leaving its aggregate impact insignificant. Hence, it suggests that future macroeconomic data would illustrate persistently sluggish AI-driven aggregate growth, since the underlying optimistic expectations were incorrect (Brynjolfsson et al. 28).

Task-based model

The task-based model, developed by Acemoglu in The simple macroeconomics of AI, can be considered the most important modern theory supporting this hypothesis. In this framework, the production of final goods and services is illustrated by the production function:

$$Y(z) = \left(\int_0^N y(z)^{\frac{\sigma-1}{\sigma}} dz \right)^{\frac{\sigma}{\sigma-1}} \quad (1)$$

where $Y(z)$ indicates aggregate output from performing a continuum of tasks $z \in [0, N]$, and $0 \leq \sigma \leq 1$ represents the substitution elasticity across them. Meanwhile, individual task $y(z)$ can be allocated to either labour or capital, which have different productivities between tasks but are perfect substitutes within a specific one. The allocation of these inputs depends upon their relative comparative advantage (e.g. relative costs, technical implementation feasibility) as factors of production. Execution of an individual task is represented by the function:

$$y(z) = A_l \gamma_l(z) l(z) + A_k \gamma_k(z) k(z) \text{ for any } z \in [0, N] \quad (2)$$

where

A_l, A_k = labour-augmenting and capital-augmenting productivity

γ_l, γ_k = task productivity schedules of labour and capital

$l(z), k(z)$ = labour and capital allocated for task z

Consequently, AI-driven productivity gains are evaluated by 2 primary task-related mechanisms. The first stream is automation, as AI displaces labour in tasks where its implementation is more cost-efficient. The second is task complementarity, as AI is deployed to routine and data-heavy subtasks (e.g. data processing, pattern recognition), and labour is devoted to activities where empathy, physical dexterity, and complex reasoning are of greater value (e.g. strategic direction, contextual evaluation). This augmentation can improve the productivity of not only unautomated tasks but also marginal labour output. Under Hulten's theorem, as long as these AI microeconomic productivity gains are cost reductions, its GDP and aggregate TFP gains can be estimated by the proportion of tasks under AI exposure and average task-level cost savings. This effect on aggregate TFP is illustrated by the equation:

$$\Delta \ln TFP = \int_{z \in N} \lambda_z \cdot \Delta \ln A(z) dz \tag{3}$$

where λ_z measures the Domar weight of each AI-affected task (i.e. the significance of a task in GDP, derived from its wage share multiplied by the factor income share), and $\Delta \ln A(z)$ denotes the micro-level cost reduction or productivity gains of each specific instance. Meanwhile, the total impact of these AI-exposed tasks is given by:

$$\Delta \ln Y = \Delta \ln TFP + s_K \Delta \ln K \tag{4}$$

in which s_K is the capital share in GDP, and $\Delta \ln K$ is the improvement in capital stock.

Using data from several studies recent to its developmental time, the framework estimated that AI economic integration should only yield a TFP growth rate of 0.53 – 0.66% and a GDP growth rate of 0.93 – 1.16% from 2023 to 2033. Several task-based factors are proposed to explain this current wedge between pronounced microeconomic productivity gains and insignificant macroeconomic growth. Firstly, *easy-to-learn tasks*, characterised by their standardised inputs, objective outcome metrics, and simple action-outcome mapping, are the ones whose large AI-driven productivity gains have been documented in microeconomic experiments. However, for the remaining large share of *hard-to-learn tasks*, the absence of a simple mapping has made it difficult to train AI models to automate or augment them. This may have created a substantial upward bias in the forecasts and expectations on macroeconomic growth. Secondly, the current AI boom is dominated by “so-so automation”, which refers to technologies that have strong displacement effects on human labour, but weak broader productivity improvements. This automation arises to save production costs for firms, but it does

not yield significant macroeconomic price reductions or impose fundamentally transformative increases in total task output (Acemoglu and Restrepo 3). Hence, macroeconomic statistics stall.

Redistribution

Another hypothesis holding a negative view of AI adoption is "Redistribution", which suggests that observed microeconomic productivity improvement is a product of zero-sum market dynamics rather than net value creation. Specifically, it argues that corporate investments in AI are carried out so that firms can become more efficient individually (i.e. lower production costs and boost profits), enabling them to be more competitive and capture larger market shares. In this context, firm-level productivity enhancement is driven by wealth shifts from other parties, such as competitors (larger shares of the existing pool of profits) and labour (replaced by AI to save production costs). Hence, there will be no net increase in aggregate economic wealth and productivity. This explains why microeconomic experimental studies have shown massive task-level efficiency gains, but aggregate output and productivity growth remain stagnant (Brynjolfsson et al., Artificial Intelligence and the Modern Productivity Paradox 29).

Product variety model

The product-variety model is a foundational framework introduced in the endogenous growth theory (Romer 10), which explains long-term economic growth as fuelled by continuous creation of new types of specialised intermediate inputs – referring to goods and services rather than simply producing existing goods more efficiently. The production process, as described by the framework, is shown by the function:

$$Y^S = L^{1-\alpha} \int_0^N x_i^\alpha di \tag{5}$$

where

Y = aggregate national final output
L = labour allocated to production of final goods
N = quantity of different product varieties
x_i = the quantity of intermediate good i being produced, for which $i \in [0, N]$
α = output elasticity parameter of each input

To model the macroeconomic impacts of AI and highlight how its adoption creates a zero-sum redistribution of benefits, Gries and Naudé introduced two modifications to this standard model. Firstly, they incorporated AI as both an innovation that creates new product varieties (expanding N) and an automating input that enables capital (K) to displace labour (L) in producing both final and intermediate goods (raising α). Secondly, they imposed demand-side constraints, requiring firms to effectively sell all new output (Y) from AI adoption for economy-wide productivity to grow. Thus, grounding the neoclassical supply-side theoretical growth model in Keynesian reality.

As a constraint, the aggregate demand is specified through the equation:

$$Y^D = C + I = c_w \cdot W + c_k \cdot \Pi + I \quad (6)$$

where

W = total wages paid to human labour ($W = w \cdot L$)
Π = total profits earned by capital and AI owners
C = total consumption (consumption component of aggregate demand)
I = total investment (investment component of aggregate demand)
c_w = marginal propensity to consume (MPC) for workers
c_k = marginal propensity to consume for capital owners

Using these functions, the model frames the AI productivity paradox as being caused by bottlenecks in the demand-side channel. With AI replacing human workers in expanding product varieties N and producing final output, demand for labour falls. This leads wages W to stagnate relative to profit Π, shifting a greater share of national income to capital owners and causing the labour income share to drop. Given that the marginal propensity to consume of labour is much higher than that of capital owners, this shift in income share will significantly reduce aggregate consumption. As a result, aggregate demand will fall below potential capacity $(Y^D < Y_P)$. So, while firms can produce considerably more product varieties N with the adoption of artificial intelligence, it is not matched by a sufficient increase in aggregate demand to consume the extra potential output. As production capacity is underutilised, aggregate output is constrained by low consumption, resulting in stagnation in both measured output and TFP. Ultimately, this leads to the AI productivity paradox, where microeconomic productivity gains fail to assemble into measurable improvements in macroeconomic statistics.

Mismeasurement & Implementation Lag

The "Mismeasurement" hypothesis holds a technologically optimistic outlook on AI, presuming that conventional macroeconomic metrics are incapable of accounting for the true extent of its economic benefits (Brynjolfsson et al., Artificial Intelligence and the Modern Productivity Paradox 29). In particular, it argues that these tools were originally designed to quantify market-based physical production for a 20th-century industrial economy, and AI has generated economic value in a way that defies this convention. Firstly, a significant portion of its consumer surplus improvement is unpriced. AI services like integrated search bars or free chatbots are enjoyed at minimal or no direct cost, delivering great consumer utility (e.g. greater convenience) that is not recorded as paid market transactions. Thus, they are omitted in macroeconomic statistics (Coyle and Poquiz 2–3). Secondly, current investment primarily goes into intangible assets (e.g. software, research and development, corporate reorganisation and proprietary data) to maximise the capability of AI adoption. However, in the traditional framework, these expenditures are viewed as short-run production costs rather than long-term capital investment. Since registered monetary inputs are matched by unrecorded capital formation, growth statistics become artificially flat in the short run (Brynjolfsson et al., The

Productivity J-Curve 335). Overall, this argument states that artificial intelligence is delivering real economic improvements, but flawed accounting is responsible for the gap between microeconomic and macroeconomic statistics. However, it has been heavily criticised, as more economic studies illustrate that statistical error is an insufficient contributor to macroeconomic slowdowns (Syverson 165).

The final hypothesis – "Implementation Lag" – holds a more balanced perspective, being optimistic about the economic potential of artificial intelligence while still acknowledging the reality of stagnant macroeconomic growth. It suggests that AI, as the next General Purpose Technology (GPT), will bring transformative macroeconomic advancements to the economy, which stay invisible in the short run. This considerable delay between development and benefit materialisation emerges for two main reasons. The first is that a long time period is required for the AI stock to accumulate sufficiently and yield a measurable aggregate effect. The other factor is that complementary intangible assets (e.g. labour retraining, organisational restructuring, data infrastructure) are necessary to extract the full capability of artificial intelligence, and it takes time and massive investments to develop and implement them. To sum up, while the transformative potential of the new technology is clearly recognisable, the pathway to realise it can be lengthy and exhaustive. Hence, this explanation resolves the paradox by bridging two seemingly contradictory facets as different stages of an underlying technological development process. This hypothesis currently dominates research consensus as the most convincing explanation of the AI productivity paradox (Brynjolfsson et al., Artificial Intelligence and the Modern Productivity Paradox 31).

Productivity J-Curve

The Productivity J-Curve, as developed by Brynjolfsson et al. in The Productivity J-Curve, bridges these two hypotheses together. It illustrates how the adoption of artificial intelligence, as a new general-purpose technology (GPT), would experience a delay in real-world returns, as well as distorted statistical representations of its impacts.

Traditional growth accounting employs a generalised production function, where aggregate output Y is the product of total factor productivity A and a function F that has constant returns to scale in its inputs K (capital) and L (labour):

$$Y = A \cdot F(K, L) \tag{7}$$

Measured productivity growth is then denoted by the equation:

$$\begin{gathered} g_A = g_Y - \left(\frac{rK}{Y}\right) g_K - \left(\frac{wL}{Y}\right) g_L \\ g_Y = \frac{dY}{Y}, g_K = \frac{dK}{K}, g_L = \frac{dL}{L}, g_A = \frac{dA}{A} \end{gathered} \tag{8}$$

where

g = measured growth rate

r = factor prices of capital
w = factor prices of wages

However, in reality, firms are also investing heavily in unmeasured complementary intangible capital, denoted as U, to adjust the production process so that the new GPT can be smoothly integrated into the economy. Hence, the true output Y^* of an economy encompasses both tangible output Y and intangible investment I_U. In addition, the adjusted aggregate output function becomes the product of total factor productivity A^* and a constant returns to scale function F^* of tangible capital K, labour L, and intangible capital U:

$$Y^* = Y + I_U = A^* \cdot F^*(K, U, L) \tag{9}$$

This results in a new equation illustrating the true productivity growth g_A^* for the intangible-inclusive economy:

$$g_A^* = \left(\frac{Y}{Y^*}\right)\left(g_Y - \left(\frac{rK}{Y}\right)g_K - \left(\frac{r_U U}{Y}\right)g_U - \left(\frac{wL}{Y}\right)g_L\right) + \left(\frac{I_U}{Y^*}\right)g_I \tag{10}$$

where r_U is the factor price of intangible capital.

Considering the impact of intangible capital investment on the measurement of underlying productivity growth, the framework explains the dynamics of the AI productivity paradox as follows. After AI arrives in an economy, measured productivity growth will experience a J-curve trajectory, which can be structured into two stages. In the first stage, firms accelerate intangible investment (e.g. process redesign, human capital, data curation, organisational restructuring). As a result, physical resources like labour and capital are diverted from producing final output towards building complementary intangible capital. According to the traditional accounting method, these allocated resources are registered in aggregate inputs, but the newly created intangible assets are not included in aggregate output. Thus, both measured output and productivity underestimate the actual economic performance, giving rise to the issues described in the mismeasurement hypothesis. At the second stage, intangible capital investment has become sufficiently accumulated and mature, enabling it to be used in producing final output. However, their intangible nature makes them unregistered as inputs, creating an apparent mismatch between missing capital expenditures and a significant increase in measurable tangible output. Given that TFP is computed as the ratio between aggregate output and aggregate input, measured TFP exceeds true productivity in this phase. Altogether, they form the basis of the implementation lag hypothesis, illustrating how organisations need to develop complementary structures for many years, or even decades, to materialise AI improvements in macroeconomic statistics.

Discussion

	False Hopes (H1)	Redistribution (H2)	Mismeasurement (H3)	Implementation Lag (H4)
Mechanism	AI gains are localised and superficial, thus vanishing at scale	AI gains are due to zero-sum wealth redistribution rather than net value creation	Invisible improvements by AI are improperly measured by traditional metrics	Extensive and time-consuming complementary investments are necessary for AI gains to appear at scale
Prediction on AI gains	Small	Redistributed	Incorrectly measured	Delayed
Evidence	High adoption rate but persistently stagnant national statistics	Massive gains for beneficiaries (e.g. high profits for AI firms) and matching losses for losers (e.g. lower incomes for displaced labour)	Considerable improvements (e.g. intangible capital, unpriced welfare) not recorded in national metrics	Productivity growth surges following introduction of historical GPTs in the long term
Limitation	Dismisses weighty micro-level proof as being superficial	Underestimating the spillover effects that can fuel aggregate expansion	Mismeasurement gaps are insufficient to account for the persistent slowdown	Difficult to determine whether current slowdown reflects a permanent stagnation or temporary lag
Relationship	Competing with H2, H3, and H4	Competing with H1, H3, and H4	Competing with H1 and H2, complementary with H4	Competing with H1 and H2, complementary with H3

The four hypotheses offer both competing and complementary explanations for the AI productivity paradox. By investigating how their underlying theories bridge microeconomic realities and stagnant macroeconomic statistics, the review reveals a key tension: whether current macro-level stagnation is structural or transitional. The pessimistic view considers sluggish macroeconomic data as a permanent reality, but disagrees on the root cause. “False Hopes” contends that AI capabilities are fundamentally exaggerated and have limited depth to drive macroeconomic transformation. Conversely, “Redistribution” acknowledges that AI can drive powerful microeconomic gains, but depicts their delivery as macroeconomically redistributive rather than additive. The optimistic view believes that AI has a transformative but temporarily concealed potential, and is unified under the Productivity J-Curve framework. Here, macro-stagnation is fuelled by two distinct but compatible mechanisms. The underlying one is “Implementation Lag”, where resources are diverted from producing final goods and services to restructuring the entire production engine. The more apparent one is “Mismeasurement”, where

traditional macroeconomic metrics have failed to account for the intangible capital produced for the restructuring. Overall, the optimistic camp has been more consolidated with the development of the Productivity J-curve, while the pessimistic camp remains more divisive due to competing theories.

Conclusion

In this paper, the theoretical research landscape on the "AI productivity paradox", particularly how AI-driven microeconomic productivity improvements have not accumulated into measurable macroeconomic growth, is summarised and structured into four different thematic hypotheses. "False Hopes" and "Redistribution" share a similar pessimistic stance on AI, though they interpret the gap through separate mechanisms. The former, "False Hopes", presumes that underlying expectations about AI potential have been overly optimistic. So, AI-driven microeconomic productivity gains are inherently superficial, and macroeconomic stagnation is a permanent ceiling reflecting the true limits of this tool. It is supported by the task-based model, developed by Acemoglu (13), which has gained significant academic interest in the last two years. The latter, "Redistribution", suggests that observable private microeconomic gains are created from zero-sum benefit redistribution across economic agents, translating into neutral net aggregate gains. This explanation is endorsed by the demand-constrained product-variety model, proposed in Gries and Naude (2). However, this paper has not been published and remains a discussion paper of the IZA Labour Institute. The other two hypotheses present a more optimistic outlook on the macroeconomic impacts of AI, with "Mismeasurement" claiming conventional macroeconomic metrics are incapable of reflecting the true extent of AI-generated economic gains, and "Implementation Lag" states that a considerable amount of time is necessary for corporate complementary investments to mature and enable AI to produce improvements. These two explanations have been unified as complementary mechanisms for the same paradoxical phenomenon in the Productivity J-curve framework, which currently remains the most discussed theory that dominates the consensus (ILO 4). To summarise, comparison of these four thematic hypotheses reveals a theoretical schism: whether macroeconomic statistical stagnation reflects structural limitations or transitional frictions in AI adoption.

As a theoretical review, the main limitation of this paper is that it examines an evolving phenomenon at an early stage, giving rise to several structural weaknesses. Firstly, while the broader body of literature documenting the economic impacts of AI is emerging rapidly, only a limited number of papers tracking the micro-to-macro disconnection of AI gains underwent rigorous peer review and reached publication status. This left a narrowly scoped base of secondary literature to be pooled from. Secondly, since macroeconomic statistics are slow to reflect structural shifts, evaluating macroeconomic stagnation using data from the first half of the 2020s is not representative of the true phenomenon. Given that all proposed theoretical frameworks in this review are built on early macroeconomic data, they risk being premature explanations of this phenomenon. Overall, having an evolving research target subjects this

review to being an incomplete reflection. Another important limitation lies in the descriptive synthesis approach. It has comprehensively mapped out and categorised theoretical frameworks into four different thematic hypotheses. However, this review does not manage to critically evaluate their relative explanatory power or weigh their empirical robustness under current economic data and conditions. As some of these hypotheses hold competing or distinct views, while a few can be complementary to each other, some of the underlying relationships between these hypotheses were not resolved. Consequently, the depth of this theoretical review is severely restricted.

Based on the outline from this descriptive synthesis, this theoretical review provides several implications for future research direction. One of them is carrying out a critically evaluative approach to the identified four main hypotheses of the AI productivity paradox. This evaluation will encompass an examination of the relative logical and explanatory power of these explanations, as well as testing their empirical strength using a more sufficiently developed and mature pool of real-world data and evidence in the future. By then, the most convincing theory to explain this phenomenon can be identified. Another approach, which is generally built upon the findings from the critical synthesis, is the construction of a unified meta-framework, integrated from the identified disparate theories. The existence of such a framework enables the most cohesive understanding of the underlying mechanisms of the micro-to-macro disconnection, providing better insights for policymakers on resolving this paradox to extract the optimal capability of AI.

Works Cited

Acemoglu, Daron. 'The Simple Macroeconomics of AI'. Economic Policy, vol. 40, no. 121, Jan. 2025, pp. 13–58. Silverchair, https://doi.org/10.1093/epolic/eiae042.

Acemoglu, Daron, and Pascual Restrepo. 'Automation and New Tasks: How Technology Displaces and Reinstates Labor'. Journal of Economic Perspectives, vol. 33, no. 2, May 2019, pp. 3–30. www.aeaweb.org, https://doi.org/10.1257/jep.33.2.3.

Artificial Intelligence Index Report 2026. Stanford University, 2026, p. 193. Artificial Intelligence Index.

Brynjolfsson, Erik, Daniel Rock, et al. 'Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics'. The Economics of Artificial Intelligence: An Agenda, University of Chicago Press, 2018, pp. 23–57. National Bureau of Economic Research, https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/artificial-intelligence-and-modern-productivity-paradox-clash-expectations-and-statistics.

Brynjolfsson, Erik, Danielle Li, et al. 'Generative AI at Work'. The Quarterly Journal of Economics, vol. 140, no. 2, May 2025, pp. 889–942. Silverchair, https://doi.org/10.1093/qje/qjae044.

Brynjolfsson, Erik, Daniel Rock, et al. 'The Productivity J-Curve: How Intangibles Complement General Purpose Technologies'. American Economic Journal: Macroeconomics, vol. 13, no. 1, Jan. 2021, pp. 333–72. www.aeaweb.org, https://doi.org/10.1257/mac.20180386.

Calvino, Flavio, et al. 'Is Generative AI a General Purpose Technology?: Implications for Productivity and Policy'. OECD Artificial Intelligence Papers, June 2025. www.oecd.org, https://doi.org/10.1787/704e2d12-en.

Callen, Tim. 'Gross Domestic Product: An Economy's All'. IMF, https://www.imf.org/en/publications/fandd/issues/series/back-to-basics/gross-domestic-product-gdp. Accessed 9 Sept. 2026.

Coyle, Diane, and John Lourenze S. Poquiz. 'Making AI Count: The Next Measurement Frontier'. Working Paper no. 34330, National Bureau of Economic Research, Oct. 2025. Working Paper Series. National Bureau of Economic Research, https://doi.org/10.3386/w34330.

Czarnitzki, Dirk, et al. 'Artificial Intelligence and Firm-Level Productivity'. Journal of Economic Behavior & Organization, vol. 211, July 2023, pp. 188–205. ScienceDirect, https://doi.org/10.1016/j.jebo.2023.05.008.

Damioli, Giacomo, et al. 'The Impact of Artificial Intelligence on Labor Productivity'. Eurasian Business Review, vol. 11, no. 1, Mar. 2021, pp. 1–25. Springer Link, https://doi.org/10.1007/s40821-020-00172-8.

Gao, Xueyuan, and Hua Feng. 'AI-Driven Productivity Gains: Artificial Intelligence and Firm Productivity'. Sustainability, vol. 15, no. 11, Jan. 2023, p. 8934. www.mdpi.com, https://doi.org/10.3390/su15118934.

Gries, Thomas, and Wim Naudé. ‘Artificial Intelligence, Jobs, Inequality and Productivity: Does Aggregate Demand Matter?’ SSRN Scholarly Paper no. 3301777, Social Science Research Network, 17 Dec. 2018. papers.ssrn.com, https://papers.ssrn.com/abstract=3301777.

Katz, Raúl, and Juan Jung. ‘The Macroeconomic Effects of Generative AI’. Structural Change and Economic Dynamics, vol. 79, Aug. 2026, pp. 400–11. DOI.org (Crossref), https://doi.org/10.1016/j.strueco.2026.05.013.

Noy, Shakked, and Whitney Zhang. ‘Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence’. Science, vol. 381, no. 6654, July 2023, pp. 187–92. www.science.org, https://doi.org/10.1126/science.adh2586.

Parteka, Aleksandra, and Aleksandra Kordalska. ‘Artificial Intelligence and Productivity: Global Evidence from AI Patent and Bibliometric Data’. Technovation, vol. 125, July 2023, p. 102764. DOI.org (Crossref), https://doi.org/10.1016/j.technovation.2023.102764.

Romer, Paul M. ‘Endogenous Technological Change’. Journal of Political Economy, vol. 98, nos 5, Part 2, Oct. 1990, pp. S71–102. journals.uchicago.edu (Atypon), https://doi.org/10.1086/261725.

Sickles, Robin C., and Valentin Zelenyuk. Measurement of Productivity and Efficiency: Theory and Practice. Cambridge University Press, 2019, https://doi.org/10.1017/9781139565981.

Syverson, Chad. ‘Challenges to Mismeasurement Explanations for the US Productivity Slowdown’. Journal of Economic Perspectives, vol. 31, no. 2, May 2017, pp. 165–86. www.aeaweb.org, https://doi.org/10.1257/jep.31.2.165.

The Aggregation Paradox of AI: Why Do Micro-Economic Productivity Gains from AI Disappear at Scale. International Labour Organisation, https://doi.org/10.54394/00034342.

CSR Disclosure and Stock Returns: Evidence from Selected UAE-Listed Real Estate Firms
By Riaan Shah

Abstract

This study examines whether corporate social responsibility (CSR) disclosure is associated with the stock returns of real estate firms listed in the United Arab Emirates. Real estate is one of the most economically important and publicly visible sectors in the UAE, and listed developers operate under close scrutiny from investors, regulators and the public. CSR disclosure is measured using a disclosure index built through content analysis of annual and sustainability reports across five categories relevant to property firms: environmental responsibility, workforce and employee safety, customer and tenant responsibility, ethics and governance, and community engagement. Each category is scored on a binary basis and the index is the unweighted average of the five items. The sample covers five listed firms over the period 2016 to 2025, giving 40 firm-year observations. CSR disclosure in year *t* is used to predict stock returns in year *t+1*, providing a lagged specification that partially reduces contemporaneous reverse-causality concerns. Using pooled ordinary least squares, the relationship between CSR disclosure and next-year returns is negative and statistically insignificant ($\beta = -0.150$, $p = .512$). Adding a control for firm valuation leaves the coefficient insignificant ($\beta = -0.237$, $p = .297$). The results provide no evidence that stronger CSR disclosure predicts higher subsequent stock returns in this sector. One category was disclosed in every firm-year and another in almost all observations, limiting the discriminatory power of the composite index.

Introduction

The real estate industry is among the most economically important and publicly visible industries in the United Arab Emirates. Rapid urbanisation, sustained infrastructure expansion and the strategic ambition to position Dubai and Abu Dhabi as global business and tourism hubs have placed property development at the frontline of economic transformation (Buckley and Hanieh, 2014). Listed real estate developers and property investment companies operate at the crossroads of mega-project delivery, long-term ownership of assets and capital-intensive investment. These firms face close scrutiny from investors, regulators and the wider public, and that scrutiny increasingly extends to how they manage their social and environmental responsibilities.

Capital markets have increasingly incorporated corporate social responsibility (CSR) into firm valuation and portfolio construction. Stronger CSR performance may influence stock performance through several channels. It can appeal to socially and environmentally conscious investors, support better risk management, build stakeholder confidence, and lower long-run operational costs (Masoud and Vij, 2021). These mechanisms matter particularly in real estate. Environmental responsibility relates to energy efficiency, sustainable building standards and climate resilience. Social responsibility relates to health and safety practices, labour conditions

and community engagement. Governance relates to transparency, regulatory compliance and ethical management. If CSR performance signals better management quality or longer-term strategic thinking, this could be reflected in stock returns. The existing literature, however, does not show a consistent relationship between CSR and financial performance, which raises the question of whether stronger CSR engagement actually translates into better stock performance.

The UAE offers a distinctive setting in which to examine this question. The real estate sector is closely connected to global investment flows, and sustainability reporting and investor attention have grown in importance. Combined with the scale and visibility of property development in the country, this makes the sector a timely environment in which to test whether CSR performance affects stock returns or whether the relationship is more complicated than is often assumed.

This study investigates whether CSR performance is associated with subsequent stock returns among UAE-listed real estate companies. The main research question of this study asks:is CSR disclosure associated with subsequent annual stock returns among selected UAE-listed real estate firms? The study contributes sector-specific evidence from a market that remains underrepresented in research on the pricing of corporate responsibility. By focusing on a highly visible, capital-intensive sector within an emerging but globally integrated economy, the paper offers insight into whether CSR disclosure is reflected in subsequent stock-market performance. The findings carry practical relevance for investors making portfolio decisions, for corporate leaders weighing the capital-market implications of sustainability strategies, and for regulators assessing whether reporting frameworks influence financial outcomes.

Literature Review

CSR and stock returns

Corporate social responsibility refers to companies taking responsibility for their social, environmental and ethical impacts beyond what the law requires. Corporate social performance refers to the measurable outcomes of these activities. CSR may include reducing environmental harm, treating employees fairly, engaging ethically with communities and consumers, and improving corporate governance. Corporate social performance is the term more often used in empirical research, where the aim is to assess how effectively firms implement these responsibilities (Wood, 1991).

Studies in this field measure stock performance in two main ways. The first is unadjusted returns, which is simply the profit an investor makes from holding the stock over a period. The second is abnormal returns, which capture the part of a stock's return that sits above or below what would be expected given market movements and common risk factors. Event studies rely on abnormal returns. Researchers use narrow event windows and measures such as cumulative abnormal returns to estimate how the market responds to a specific CSR-related event, such as the publication of a sustainability report, a scandal or a lawsuit.

The relationship between CSR and stock performance has not been settled empirically.

Whether CSR improves corporate financial performance is ultimately an empirical question, because theory can support different directions depending on the circumstances and the mechanism involved (von Arx & Ziegler, 2014). On the positive side, stronger CSR performance could be rewarded if it signals higher management quality, lower long-run regulatory, reputational and operational risk, or attracts investors who are willing to hold the stock for ethical reasons.

Large-scale reviews suggest that CSR is more often linked with positive or at least non-negative financial performance. Friede, Busch and Bassen (2015) reviewed more than 2,000 empirical studies and found that around 90 per cent of results were non-negative, with a substantial share positive. This does not mean every study finds a benefit, or that CSR produces higher returns in all settings. Results differ because CSR is measured differently across datasets, time periods and countries, because industries face different stakeholder pressures, and because causality is difficult to establish. Financially strong firms may simply be better placed to invest in CSR in the first place. Working in the other direction, event-study evidence often shows that negative CSR events, such as environmental violations or social misconduct, produce negative abnormal returns, which is consistent with markets punishing reputational damage and expected future costs (Krüger, 2015). The state of the literature is best described as mixed: there is a substantial body of evidence that CSR has non-negative effects overall, but this varies by method and setting.

Why CSR could affect returns

CSR may be linked to stock returns through several economic and behavioural channels. Trust is one such mechanism. CSR engagement can help firms build credibility among stakeholders including customers, employees, regulators and local communities. This trust functions as an intangible resource that becomes especially valuable during periods of stress, such as economic crises or firm-specific problems. Where trust is high, stakeholders may be more willing to maintain their relationship with the firm and absorb temporary losses. CSR may therefore stabilize cash flows and reduce downside risk in a way that is reflected in stock prices. Lins, Servaes and Tamayo (2017) provide direct evidence for this channel, showing that during the 2008–2009 financial crisis firms with high CSR intensity earned stock returns four to seven percentage points higher than firms with low CSR intensity, alongside higher profitability and growth. A related study of the corporate bond market found no general relationship between social capital and bond spreads, but a clear benefit for high-performing firms during the same crisis period (Amiraslani et al., 2023). Both findings suggest that CSR matters most where social approval and long-term relationships carry economic weight.

A second channel works through investor preferences. Financial markets do not consist only of investors who care about risk and return. Many investors prefer to hold shares in companies that align with their ethical, social or environmental values, and some are willing to accept slightly lower expected returns to do so. Where these preferences drive demand for high-CSR stocks, prices are bid up in the present. Higher current prices mechanically imply

lower expected future returns, holding cash flows constant. From an asset-pricing perspective, this means CSR may be associated with lower expected returns not because it destroys value, but because it raises current valuations through preference-based demand. This mechanism becomes more relevant as demand from socially responsible investors increases.

A third mechanism is risk management. CSR activities can reduce the likelihood or severity of adverse events such as environmental accidents, labour disputes, regulatory fines, litigation or reputational scandals. By reducing exposure to these outcomes, CSR may lower firm-specific risk and possibly systematic risk. Risk-averse investors may value this insurance-like quality, which again raises prices and lowers expected returns. Empirical evidence supports this reading, since markets tend to react more strongly to negative CSR events than to positive ones, suggesting investors are particularly sensitive to downside risk and reputational damage (Krüger, 2015).

Taken together, these mechanisms suggest that CSR may affect stock returns through trust, preference-driven demand and risk reduction. Importantly, they do not all point in the same direction. A negative association between CSR and future returns is not necessarily evidence that CSR is costly to shareholders, since it is equally consistent with CSR being valued and already reflected in current prices.

Real estate and the UAE setting

Real estate is a particularly relevant sector in which to examine CSR and financial performance, because property projects have direct and visible effects on the environment and on society. Construction affects land use, energy consumption, pollution and local communities. Unlike some other industries, these effects are long-lasting, since buildings and infrastructure remain in place for decades. CSR is therefore not an optional activity in real estate. It is closely tied to how companies operate and how they perform financially.

CSR is also financially important in real estate because the sector is capital-intensive and high-profile. Developers require large investments, and their projects are closely watched by governments, investors and the public. A company that follows good CSR practices, such as using sustainable materials or reducing energy use, may improve its reputation and attract investors. Poor practices can lead to legal problems, project delays or loss of trust. CSR can therefore affect both risk and returns in this sector.

The UAE is a distinctive setting for this question. Real estate plays a major role in the national economy, with rapid development and large-scale projects such as the Burj Khalifa and Palm Jumeirah. These projects are highly visible internationally, which increases the value of maintaining a strong reputation. The country also faces environmental pressures including extreme heat, high energy consumption and limited water resources, which makes sustainability-related CSR activity more consequential than in some other markets.

The UAE also differs from the United States and Europe in that its CSR regulations and reporting standards are still developing. Investor expectations may differ as well, which means the effect of CSR on financial performance may not resemble that observed in more developed

markets. These differences make the UAE worth studying separately.

Although there is extensive global research on CSR and financial performance, evidence focused specifically on UAE-listed real estate companies is limited. Recent work in other emerging markets reports that composite disclosure measures often have an insignificant effect on firm performance (Biju et al., 2025; Shobhwani & Lodha, 2024). This study examines whether CSR performance is linked to future stock returns in the UAE real estate sector, and in doing so adds evidence from a rapidly developing market.

Methodology

This study uses a quantitative research design to examine the effect of CSR disclosure performance on the stock returns of UAE-listed real estate firms between 2016 and 2025. Following the mono-method quantitative approach described by Saunders et al. (2016), it draws on secondary data from two sources: CSR disclosures taken from annual and sustainability reports, and share price data used to calculate returns. The relationship is then tested statistically through regression analysis.

The sample comprises five listed real estate firms and 40 firm-year observations. The firms are Emaar Properties, Deyaar Development, Al Mazaya Holding, DAMAC Properties and TECOM Group, all of which were quoted on the DFM during at least part of the study period. Al Mazaya Holding is dual-listed, trading on both the Dubai Financial Market and Boursa Kuwait.

Firms were included where they had a real estate primary activity, a listing on a UAE exchange during the study period, at least one annual or sustainability report per included year, and available share price data. The number of observations differs across firms because of their listing histories. Emaar, Deyaar and Al Mazaya contribute ten observations each, DAMAC seven, and TECOM three. DAMAC was delisted from the Dubai Financial Market in 2022 following a takeover by a vehicle owned by its founder, and TECOM listed in 2022, which explains the shorter series for both firms.

Table 1. *Sample composition*

Firm	Exchange	Observations
Emaar Properties	DFM	10
Deyaar Development	DFM	10
Al Mazaya Holding	DFM / Boursa Kuwait	10
DAMAC Properties	DFM	7
TECOM Group	DFM	3
Total		40

CSR performance is measured through a CSR disclosure index constructed by content

analysis of annual and sustainability reports. Five categories were selected for their relevance to real estate firms:

1. Environmental responsibility — green buildings, energy efficiency, waste and water initiatives
2. Workforce and employee safety — training, health and safety metrics, worker welfare
3. Customer and tenant responsibility — complaints handling, quality and service standards
4. Ethics and governance — anti-corruption, whistleblowing, compliance culture
5. Community engagement — local investment, philanthropy, stakeholder consultation

Each category is scored on a binary basis, with disclosure receiving a score of 1 and non-disclosure a score of 0. The index is the unweighted average of the five items and ranges from 0 to 1, where 1 indicates disclosure across every category.

This measure has limitations that should be stated at the outset. It captures what firms report rather than what they do, and binary scoring records whether a topic is addressed without capturing the depth or quality of the disclosure. A brief mention and a detailed audited section receive the same score. Equal weighting also assumes that all five categories carry the same informational value, which is a simplification made for transparency.

The dependent variable is the firm's annual stock return in year $t+1$, calculated from year-end share prices. Pairing CSR disclosure in year t with returns in the following year reflects the fact that CSR information for a given financial year only becomes public when the report is released, usually in the first half of the following year. Using same-year returns would assume investors had access to information they did not yet hold.

The lagged structure also reduces concerns about reverse causality. A firm whose share price performed well in a given year may have had more resources available to spend on CSR activity and reporting in that same year. Placing the return window after the disclosure year means this effect would have to work through persistence in returns rather than operating at the same time. It reduces the concern rather than eliminating it.

Two models are estimated using pooled ordinary least squares. The first examines the relationship between CSR disclosure and next-year returns on its own:

$$\mathbf{Return}_{i,\ t+1} = \alpha + \beta_1 \mathbf{CSR}_{i,t} + \varepsilon_{i,t}$$

The second adds a control for firm valuation in the year of CSR measurement:

$$\mathbf{Return}_{i,\ t+1} = \alpha + \beta_1 \mathbf{CSR}_{i,t} + \beta_2 \mathbf{Valuation}_{i,t} + \varepsilon_{i,t}$$

Pooled ordinary least squares were chosen over panel methods because of the size of the sample. Firm fixed effects would use four degrees of freedom out of 40 observations and would restrict the analysis to variation within firms over time, which is small for a disclosure index that

tends to rise and then stabilize. Year fixed effects would use a further nine. With only five firms, clustered standard errors would also be unreliable. Analysis was carried out in Jamovi.

Results

The final dataset included 40 firm-year observations across the five developers. On average, next-year stock returns were positive across the sample at roughly 5 per cent. Four of the five firms had positive average returns, led by Emaar and Deyaar, followed by TECOM and then Al Mazaya. DAMAC was the only company with a negative average return, and it also showed by far the most variable returns, which fits with its sharp share-price decline and eventual delisting. The mean CSR disclosure score across the firms was approximately 0.6, meaning firms disclosed on about three of the five categories on average. Disclosure varied noticeably between companies. TECOM scored highest, disclosing across every category, while DAMAC and Al Mazaya scored lowest.

Some categories showed very little difference between firms. Ethics and governance were disclosed in every observation, and community engagement in almost all of them. Most of the variation in the overall score therefore came from the environmental, workforce-safety and customer responsibility categories. In practice this means a five-item index behaved closer to a three-item index, since two of the categories could not distinguish between firms.

The first regression tested whether CSR disclosure alone could explain next-year returns. It explained 1.1 per cent of the variation in next-year returns, and the relationship between CSR disclosure and returns was negative but not statistically significant.

Table 2. Model 1 — Model fit measures

R	R^2	Adj. R^2	AIC	BIC	F	df1	df2	p
0.107	0.0114	−0.0146	28.5	33.6	0.437	1	38	.512

Note. Model estimated using a sample size of N = 40.

Table 3. Model 1 — Coefficients, dependent variable: return in year t+1

Predictor	Estimate	SE	95% CI lower	95% CI upper	t	p	Std. estimate
Intercept	0.127	0.132	−0.140	0.393	0.961	.342	
CSR index	−0.150	0.226	−0.608	0.308	−0.661	.512	−0.107

The adjusted R^2 is slightly negative, which means the model performs marginally worse than one containing no predictor at all once the extra parameter is accounted for. CSR disclosure

alone did not meaningfully predict next-year returns in this sample.

A second regression added firm valuation as a control. This model explained more of the variation than the first, at 9.5 per cent, but remained statistically insignificant overall.

Table 4. Model 2 — Model fit measures

R	R^2	Adj. R^2	AIC	BIC	F	df1	df2	p
0.309	0.0953	0.0464	26.9	33.7	1.95	2	37	.157

Note. Model estimated using a sample size of N = 40.

Table 5. Model 2 — Coefficients, dependent variable: return in year t+1

Predictor	Estimate	SE	95% CI lower	95% CI upper	t	p	Std. estimate
Intercept	0.10709	0.12819	−0.153	0.36682	0.835	.409	
CSR index	−0.23742	0.22446	−0.692	0.21739	−1.058	.297	−0.169
Valuation	0.00399	0.00216	−3.75e−4	0.00836	1.852	.072	0.296

The CSR disclosure coefficient again remained negative and statistically insignificant, while valuation showed a positive link with next-year returns that was close to, but not quite, significant at the 5 per cent level. Firm valuation was therefore a stronger predictor than CSR disclosure, though still a weak one.

The standardised coefficients give a sense of scale. In the first model, a one standard deviation increase in CSR disclosure is associated with a fall in next-year returns of 0.107 standard deviations. In the second model this rises to 0.169 standard deviations. By comparison, a one standard deviation increase in valuation is associated with a rise of 0.296 standard deviations, roughly twice the size of the CSR effect.

The confidence intervals matter more than the point estimates. In the first model, the 95 per cent interval for the CSR coefficient runs from −0.608 to 0.308. Since the index ranges from 0 to 1, this means that moving a firm from no disclosure to full disclosure is associated with a change in next-year return somewhere between roughly −61 and +31 percentage points. An interval this wide cannot separate a large negative effect from a substantial positive one. Overall, the results provide no statistically significant evidence that higher CSR disclosure is associated with stronger subsequent stock returns in this sample. They also cannot rule out that such a relationship exists.

Discussion

The central finding of this study, that CSR disclosure is not significantly related to subsequent stock market performance, sits against a large body of research that generally reports a positive link. The most influential review, Friede, Busch and Bassen (2015), combined more than two thousand studies and found that around 90 per cent reported a relationship between

responsibility factors and financial performance, with the large majority of these being positive. On the surface, the present result does not confirm this broad business case for CSR.

The gap is smaller than it first appears, because the outcome measured here is narrower. Much of the aggregated evidence uses accounting profitability or valuation ratios rather than realised share returns. A firm can be more profitable because of responsible practice while its share price already reflects that expectation, which would produce a positive accounting result and no return predictability at the same time. Since this study uses next-year share returns, the null result speaks to whether CSR information predicts price movements that are not already incorporated, which is a stricter test.

The finding aligns more closely with evidence from other emerging markets. Shobhwani and Lodha (2024), examining Indian listed firms, report that overall disclosure scores have an insignificant impact on return on assets, and Biju et al. (2025) report a comparable pattern. Composite indices often produce insignificant overall results because positive and negative effects at the individual category level offset one another. That mirrors the pattern here, where ethics and community disclosure were almost universal and most variation came from the environmental, safety and customer categories.

The negative, though insignificant, direction of the coefficient is also not unusual. Brammer, Brooks and Pavelin (2006) found that UK firms scoring higher on a composite social performance indicator earned lower stock returns, a result they attribute to specific components rather than to social performance in general. Their finding fits the traditional shareholder view associated with Friedman (1970), which treats CSR spending as a cost rather than a source of value. It is equally consistent with the preference-based mechanism described in Section 2.2, where a negative association reflects CSR already being priced in rather than value being destroyed. This study cannot distinguish between those two readings, since doing so would require evidence on valuation levels rather than returns.

There is also a methodological warning from von Arx and Ziegler (2014), who show that an apparently significant negative CSR coefficient can largely disappear once the estimation moves from a simple market model to a multi-factor specification. This suggests that studies using raw returns without adjusting for risk factors, as here, are vulnerable to attributing to CSR what belongs to systematic risk exposure.

Several explanations are consistent with the result, and they should be treated as possibilities rather than conclusions. The most immediate is sample size. With 40 observations drawn from five firms, the analysis has limited power to detect anything but a large effect. The width of the confidence intervals reported in Section 4.4 makes this concrete. Absence of statistical significance in a sample this small is weak evidence that no relationship exists.

A second explanation concerns measurement. The index captures disclosure rather than conduct. If the correspondence between what firms report and what they actually do is loose, any true relationship between responsible practice and returns would be pulled toward zero. Binary scoring compounds this, since it cannot distinguish a token mention from substantive reporting.

A third explanation follows from the disclosure pattern itself. Every firm in the sample

reported on ethics and governance, and almost every firm on community engagement. Where all firms clear the same threshold, disclosure cannot explain why their returns differ. This convergence is a reasonable response to a shared regulatory and social environment, but it removes exactly the variation an analysis of this kind depends on.

Finally, the reporting environment in the region is still developing. Where disclosure practices are less standardised, the information reaching investors is harder to compare across firms, which weakens any signal it might carry. Macroeconomic and property-cycle conditions are also likely to dominate annual returns in a sector as cyclical as UAE real estate, leaving relatively little variation for firm-level disclosure to explain.

Limitations

The main limitation is the small number of UAE-listed real estate firms, which caps the precision of any estimate. The sample is drawn entirely from firms quoted on the Dubai Financial Market and does not include Abu Dhabi-listed developers, so it should not be read as representative of the sector as a whole. One sample firm, Al Mazaya Holding, is dual-listed and headquartered outside the UAE.

The panel is also unbalanced. DAMAC's later share price movements were shaped by a takeover and delisting rather than ordinary market pricing, and TECOM is observed only over three years following its listing, so firm characteristics are partly confounded with the periods in which each firm is observed.

Pooled ordinary least squares treats observations as independent when returns across UAE developers are driven by common annual shocks. This tends to understate standard errors, which means the reported p-values are, if anything, conservative in the direction of the reported conclusion. Finally, the design establishes association rather than causation. The lagged structure reduces reverse causality but cannot rule out omitted factors such as management quality or investor sentiment that could influence both disclosure and returns.

Conclusion

This study asked whether CSR performance is associated with the stock returns of publicly listed real estate firms in the UAE. It combined a CSR disclosure index, built through content analysis of corporate reports across five categories, with annual share returns for five listed developers over the period 2016 to 2025, producing 40 firm-year observations. Disclosure in each year was used to predict returns in the following year.

The estimated relationship was negative and statistically insignificant, both on its own ($\beta = -0.150$, $p = .512$) and with a control for firm valuation ($\beta = -0.237$, $p = .297$). Neither model was significant overall, and firm valuation proved a stronger, though still weak, predictor of next-year returns than disclosure. The finding should be read narrowly. The study finds no evidence that CSR disclosure predicts subsequent stock returns in this sector. It does not show that no such relationship exists, since the confidence intervals are wide enough to accommodate effects of meaningful size in either direction. With five firms and forty observations, this is as

much a statement about what a small sample can detect as about the market itself.

Two features of the disclosure data are particularly important for interpreting these estimates. Ethics and governance were disclosed by every firm in every year, and community engagement by nearly all, which means a composite index of this kind measures a common baseline as much as it measures differences between firms. Standardising the content of disclosure, rather than only the topics covered, would produce information better able to distinguish one firm from another. And the small, unstable population of listed real estate firms in the UAE places a practical ceiling on the precision that studies of this market can achieve.

For investors, the results give no basis for treating a composite disclosure score in this sector as a signal about future returns. For companies, there is no evidence that stronger disclosure is either rewarded or penalised by the market. Within this sample, the convergence in several disclosure categories may indicate that certain CSR topics function more as common reporting expectations than as meaningful differentiators between firms. This might be particularly useful for regulators. Extending the analysis to a larger sample, covering both UAE exchanges or the wider Gulf region, would allow these questions to be answered with more confidence than five firms permit.

Works cited

Amiraslani, H., Lins, K. V., Servaes, H., & Tamayo, A. (2023). Trust, social capital, and the bond market benefits of ESG performance. Review of Accounting Studies, 28(2), 421–462.

Buckley, M., & Hanieh, A. (2014). Diversification by Urbanization: Tracing the Property‐Finance Nexus in Dubai and the Gulf. International journal of urban and regional research, 38(1), 155-175.

Biju, A. V., et al. (2025). ESG-firm performance nexus: Evidence from an emerging economy. Business Strategy and the Environment. https://doi.org/10.1002/bse.4152

Brammer, S., Brooks, C., & Pavelin, S. (2006). Corporate social performance and stock returns: UK evidence from disaggregate measures. Financial Management, 35(3), 97–116. https://doi.org/10.1111/j.1755-053X.2006.tb00149.x

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), 210–233. https://doi.org/10.1080/20430795.2015.1118917

Friedman, M. (1970, September 13). The social responsibility of business is to increase its profits. The New York Times Magazine.

Krüger, P. (2015). Corporate goodness and shareholder wealth. Journal of Financial Economics, 115(2), 304-329.

Lins, K. V., Servaes, H., & Tamayo, A. (2017). Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. The Journal of Finance, 72(4), 1785–1824. https://doi.org/10.1111/jofi.12505

Masoud, N., & Vij, A. (2021). The effect of mandatory CSR disclosure on firms: empirical evidence from UAE. International Journal of Sustainable Engineering, 14(3), 378-389.

Saunders, M., Lewis, P., & Thornhill, A. (2016). Research methods for business students (7th ed.). Pearson.

Shobhwani, K., & Lodha, S. (2024). Impact of ESG disclosure scores on financial, operating and market-based performance: Evidence from NSE-100 companies. NMIMS Management Review. https://doi.org/10.1177/09711023241306219

von Arx, U., & Ziegler, A. (2014). The effect of corporate social responsibility on stock performance: New evidence for the USA and Europe. Quantitative Finance, 14(6), 977–991. https://doi.org/10.1080/14697688.2013.815796

Wood, D. J. (1991). Corporate social performance revisited. Academy of Management Review, 16(4), 691-718.

Relativistic Effects from Special Relativity to Black-Hole Accretion Disks: Theory, Paradoxes, and Numerical Visualization By Beatrice Lee (student)[1 *]

Abstract

Special relativity was developed to resolve a fundamental conflict between classical mechanics and electromagnetism: while Newtonian mechanics assumes absolute space and time and allows velocities to add in a simple way, Maxwell's theory predicts that electromagnetic waves propagate in vacuum with a fixed speed, the speed of light. Einstein's special theory of relativity replaces absolute time and space with a spacetime structure in which the speed of light is invariant in all inertial reference frames. This leads to time dilation, length contraction, relativity of simultaneity, relativistic Doppler shifts, and aberration. This paper reviews the basic principles of special relativity and illustrates their meaning through classical thought experiments and paradoxes, including the ladder-and-barn paradox, the hit-rod paradox, and the twin paradox. These examples show that apparent contradictions are resolved when simultaneity, causality, and proper time are treated relativistically. In addition to the theoretical discussion, the paper considers a simplified numerical visualization of an accretion disk around a non-rotating black hole. Although black holes require general relativity for a complete description, the radiation from rapidly orbiting disk material also displays important special relativistic effects, especially Doppler shifts, aberration, and relativistic beaming. By combining theoretical analysis with simulated images of an accretion disk, the paper demonstrates how relativistic effects move from abstract thought experiments to observable astrophysical phenomena.

Keywords Physics, Astrophysics, K Special Relativity, KBlack Holes, Accretion Disks

Introduction

In the late nineteenth century, a profound scientific challenge emerged from the tension between Isaac Newton's classical mechanics and James Clerk Maxwell's electromagnetic theory. Newtonian mechanics, formulated most famously in ***Philosophiae Naturalis Principia Mathematica*** in 1687, described space and time as absolute backgrounds against which motion takes place. In this framework, velocities are relative and additive. For example, if a person walks at $5km/h$ inside a train moving at $50km/h$ relative to the ground, the person's speed relative to the ground is simply $55km/h$. This intuitive rule works extremely well for everyday objects such as cars, trains, projectiles, and planets, whose speeds are very small compared with the speed of light.[1]

Maxwell's theory of electromagnetism created a difficulty for this classical picture. Maxwell's equations imply that light is an electromagnetic wave propagating in vacuum with a fixed speed, usually denoted by c, where $c \approx 3 \times 10^8 m/s$.

1. St. Paul's Co-educational College, 33 MacDonnell Road, Hong Kong
* Corresponding author email: leeyybeatrice@gmail.com

In ordinary Newtonian reasoning, if a source of light moved toward an observer, one might expect the measured speed of light to be $c + v$, where v is the speed of the source. However, experiments showed no such simple addition of velocities for light. The speed of light appeared to remain the same regardless of the motion of the source or observer.

To resolve this tension, nineteenth-century physicists proposed the existence of the luminiferous ether, an invisible medium through which light waves were assumed to propagate. This idea was motivated by analogy with sound waves, which require air or another material medium. If light required such a medium, then the Earth's motion through the ether should produce a measurable difference in the speed of light in different directions. The Michelson-Morley experiment of 1887 was designed to detect this "ether wind". Using an interferometer, Michelson and Morley compared the travel times of light beams moving along perpendicular paths. The expected shift in the interference pattern was not observed.[2] This null result strongly challenged the ether hypothesis and suggested that the speed of light was not measured relative to any ordinary material medium.

Einstein's 1905 paper ***On the Electrodynamics of Moving Bodies*** offered a radical solution. Instead of modifying Maxwell's equations or preserving an undetected ether, Einstein proposed two postulates.

First, the laws of physics are the same in all inertial reference frames. Second, the speed of light in vacuum is the same for all inertial observers, independent of the motion of the source or observer. These two postulates form the foundation of special relativity.[3]

The consequences of these postulates are counterintuitive from the point of view of everyday experience. Time is no longer absolute: a moving clock is measured to run more slowly than a clock at rest relative to the observer. Length is also relative: an object moving at relativistic speed is contracted along the direction of motion. Events that are simultaneous in one inertial frame need not be simultaneous in another.[1,3] These effects are not flaws in measurement, but necessary features of a consistent spacetime structure.

A central quantity in special relativity is the Lorentz factor,

$$\gamma = \frac{1}{\sqrt{1-v^2/c^2}},$$

where v is the relative speed between inertial frames. At everyday speeds, $v \ll c$, so $\gamma \approx 1$, and Newtonian mechanics is recovered as an excellent approximation. At speeds close to c, however, γ becomes significantly greater than one, and relativistic effects can no longer be ignored.

These consequences are developed in more detail below, where time dilation, length contraction, and the relativity of simultaneity are derived from the Lorentz transformations.

Special relativity is often introduced through such thought experiments: lightning strikes on a moving train, rods and ladders, synchronized clocks, and twins travelling along different paths through spacetime. These examples are useful because they isolate the essential principles of relativity. The ladder-and-barn paradox demonstrates the relativity of simultaneity. The hit-rod

paradox illustrates that mechanical signals and causal influences cannot propagate faster than light. The twin paradox shows that elapsed proper time depends on the path taken through spacetime.

However, relativistic effects are not only philosophical or mathematical curiosities. They become physically significant in systems where matter moves at a substantial fraction of the speed of light. One of the most important examples is an accretion disk around a compact object such as a black hole. In the inner regions of such a disk, gas orbits at very high speeds. The radiation emitted by this gas is affected by relativistic Doppler shifts, aberration, and beaming.[7] The approaching side of the disk can appear brighter and blueshifted, while the receding side can appear dimmer and redshifted.

A complete description of a black hole requires general relativity, because gravity near the black hole is not merely a force but a manifestation of curved spacetime. General relativistic effects such as gravitational redshift and light bending are essential for understanding the observed image of the disk.[8] Modern horizon-scale observations with the Event Horizon Telescope have directly resolved compact emission structures around the supermassive black holes in M87 and at the Galactic Center.[16,17] Nevertheless, special relativity remains crucial locally: the matter in the disk moves rapidly, and the transformation of frequency, intensity, and emission direction depends on special relativistic physics. For this reason, black-hole accretion disks provide a natural bridge between the principles of special relativity and observable high-energy astrophysical phenomena.

The aim of this paper is to review the foundations of special relativity and to show how its effects appear in a concrete astrophysical example: a simplified visualization of an accretion disk around a non-rotating black hole. This connects idealized thought experiments with a physical system in which relativistic motion directly shapes the observed image.

Foundations of Special Relativity: Lorentz Transformations, Time Dilation, Length Contraction, and Relativity of Simultaneity

Special relativity follows from two postulates: the laws of physics are the same in all inertial reference frames, and the speed of light in vacuum has the same value c for all inertial observers.[1,3] These assumptions replace the Newtonian idea of absolute time with a four-dimensional spacetime description, in which space and time coordinates depend on the observer's inertial frame. Events are described by coordinates such as

$$(ct, x, y, z),$$

where t is time and (x, y, z) are spatial coordinates. The factor c is included so that the temporal coordinate ct has the same physical dimension as the spatial coordinates.

For simplicity, consider two inertial reference frames, S and S'. The frame S' moves with constant velocity v along the x-axis of S. The origins of the two frames coincide at $t=t'=0$. In classical mechanics, the transformation between these frames would be the Galilean transformation,

$$x' = x - vt, t' = t.$$

However, this transformation does not preserve the speed of light. If a light pulse satisfies $x = ct$ in frame S, the Galilean transformation would give a different speed in frame S'. Therefore, Galilean transformations must be replaced by Lorentz transformations.[1,6]

The Lorentz transformations between the two frames are

$$x' = \gamma(x - vt),$$
$$t' = \gamma\left(t - \frac{vx}{c^2}\right),$$
$$y' = y, z' = z,$$

where

$$\gamma = \frac{1}{\sqrt{1-v^2/c^2}}$$

is the Lorentz factor. The inverse transformations are

$$x = \gamma(x' + vt'),$$
$$t = \gamma\left(t' + \frac{vx'}{c^2}\right).$$

The Lorentz factor measures how strongly relativistic effects appear. When $v \ll c$, we have $\gamma \approx 1$, and the Lorentz transformations reduce approximately to the Galilean transformations. When v becomes comparable to c, however, γ becomes significantly larger than one, and relativistic effects become important.

Spacetime Intervals and Invariance

The most important quantity preserved by Lorentz transformations is not spatial distance or time interval separately, but the spacetime interval. For two nearby events in flat spacetime, the interval is

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2.$$

For finite separations between two events, this may be written as

$$\Delta s^2 = -c^2\Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2.$$

Different inertial observers may measure different values of Δt and Δx, but they agree on the value of Δs^2. This invariance is the geometric foundation of special relativity.[6]

The interval also allows us to classify the relation between events. If

$$c^2\Delta t^2 > \Delta x^2 + \Delta y^2 + \Delta z^2,$$

the separation is timelike, meaning that one event can causally influence the other by a signal travelling slower than or equal to the speed of light. If

$$c^2\Delta t^2 = \Delta x^2 + \Delta y^2 + \Delta z^2,$$

the separation is lightlike, meaning that the events can be connected by a light signal. If

$$c^2\Delta t^2 < \Delta x^2 + \Delta y^2 + \Delta z^2,$$

the separation is spacelike, meaning that no causal signal travelling at or below the speed of light can connect the events.

This classification is essential for resolving many apparent paradoxes in special relativity. Although different observers may disagree about the order of spacelike-separated events, they cannot disagree about the order of causally connected events. Causality is therefore preserved.

Time Dilation

One direct consequence of the Lorentz transformations is time dilation. Consider a clock at rest in frame S'. Since the clock remains at the same position in its own frame, the two tick events occur at

$$\Delta x' = 0.$$

Using the inverse Lorentz transformation for time,

$$\Delta t = \gamma\left(\Delta t' + \frac{v\Delta x'}{c^2}\right),$$

and substituting $\Delta x' = 0$, we obtain

$$\Delta t = \gamma \Delta t'.$$

The time interval $\Delta t'$ measured in the rest frame of the clock is called the proper time and is often written as $\Delta\tau$. Therefore,

$$\Delta t = \gamma \Delta \tau.$$

Since $\gamma \geq 1$, the moving clock is observed to run more slowly.[1,6] In other words, a time interval measured by an observer who sees the clock moving is longer than the time interval measured in the clock's own rest frame.

This result can also be understood through the light-clock thought experiment. Imagine a clock consisting of two mirrors with a photon bouncing vertically between them. In the rest frame of the clock, the photon travels straight up and down. To an observer who sees the clock moving horizontally, however, the photon follows a diagonal path. Since the speed of light must remain c In both frames, the longer diagonal path requires more time. Thus, the moving light clock ticks more slowly.

Time dilation is not an illusion caused by delayed observation or signal travel time. It is a real difference between the time coordinates assigned by different inertial frames. It has been confirmed experimentally in particle lifetime measurements, atomic clocks on aircraft, and high-energy particle accelerators.[4,5]

Length Contraction

Another consequence of the Lorentz transformations is length contraction. Consider a rod at rest in frame S', with proper length

$$L_0 = x_2' - x_1'.$$

To measure the length of this moving rod in frame S, an observer must record the positions of both ends at the same time in S, so that

$$\Delta t = 0.$$

From the Lorentz transformation for position,

$$\Delta x' = \gamma(\Delta x - v\Delta t),$$

and using $\Delta t = 0$, we find

$$\Delta x' = \gamma \Delta x.$$

Since $\Delta x' = L_0$ and $\Delta x = L$, this gives

$$L_0 = \gamma L,$$

or

$$L = \frac{L_0}{\gamma}.$$

Thus, an object moving relative to an observer is measured to be shorter along the direction of motion.[1,6] This effect is called length contraction.

Length contraction occurs only along the direction of relative motion. Dimensions perpendicular to the motion are unchanged. As with time dilation, the effect is negligible at everyday speeds because $\gamma \approx 1$, but it becomes significant at velocities close to the speed of light.

It is important to note that length contraction depends on simultaneity. To measure the length of a moving object, the observer must determine the positions of both ends at the same time in their own frame. Since simultaneity itself is relative, different observers may assign different lengths to the same object without contradiction.[6]

Relativity of Simultaneity

Perhaps the most conceptually important consequence of special relativity is the relativity of simultaneity. In Newtonian mechanics, if two events occur at the same time in one frame, they occur at the same time in every frame. In special relativity, this is not true.

From the Lorentz transformation for time,

$$t' = \gamma\left(t - \frac{vx}{c^2}\right),$$

the time difference between two events in frame S' is

$$\Delta t' = \gamma\left(\Delta t - \frac{v\Delta x}{c^2}\right).$$

Suppose two events are simultaneous in frame S, so that $\Delta t = 0$.
Then

$$\Delta t' = -\gamma\frac{v\Delta x}{c^2}.$$

If the two events occur at different positions, so that $\Delta x \neq 0$, then generally $\Delta t' \neq 0$. Therefore, two spatially separated events that are simultaneous in one inertial frame are not simultaneous in another inertial frame moving relative to the first.[1,6]

This result explains Einstein's train-and-lightning thought experiment. Imagine a train moving at high speed past a platform. Lightning strikes the front and back of the train. An observer standing on the platform may see the two strikes as simultaneous if light from both strikes reaches the midpoint of the platform at the same time. However, an observer at the midpoint of the moving train is moving toward the light from one strike and away from the light from the other. Since the speed of light is the same in all inertial frames, the train observer concludes that the two strikes were not simultaneous.[3,6]

The same idea can be visualized using spacetime diagrams. In a spacetime diagram, time is usually plotted vertically and one spatial coordinate horizontally. The worldline of a stationary object is vertical, while the worldline of a moving object is tilted. Light rays form lines at 45° if the units are chosen so that $c=1$. The time axis of a moving frame is tilted relative to the time axis of the stationary frame, and the space axis of the moving frame is also tilted. These tilted axes represent the fact that different inertial observers slice spacetime into space and time in different ways.

An example helps illustrate this point. Suppose two volcanoes, one to the left and one to the right of a stationary observer, erupt simultaneously in the stationary frame. A spaceship moving relativistically relative to the observer will not generally assign the same time to the two eruptions. Events that lie on a horizontal line in the stationary frame do not lie on a horizontal line in the moving frame. This is not because the spaceship observer receives the light signals at different times, but because the coordinate time assigned to distant events depends on the observer's inertial frame.

The relativity of simultaneity is the key to resolving several classical paradoxes, including the ladder-and-barn paradox. In that paradox, a ladder moving at relativistic speed may be short enough to fit inside a barn in the barn's frame because of length contraction. In the ladder's own frame, however, the barn is contracted and the ladder does not fit. The apparent contradiction is resolved because the events corresponding to closing the two barn doors are simultaneous in the barn frame but not simultaneous in the ladder frame.[6]

Spacetime Diagrams and the Geometry of Lorentz Transformations

Spacetime diagrams provide a geometric way to understand Lorentz transformations. In a simple diagram with one spatial dimension, the vertical axis represents time, usually ct, and the horizontal axis represents position, x. A light ray satisfies

$$x = \pm ct.$$

If we use units where $c = 1$, the light rays have slopes of ± 1 and form the light cone. No massive object can have a worldline outside the light cone, because that would correspond to motion faster than light.

For an observer at rest in frame S, the time axis is the line $x=0$. For an observer moving with velocity v relative to S, the origin of the moving frame satisfies

$$x = vt.$$

This line is the time axis of the moving frame S'. The faster the frame moves, the more its time axis tilts toward the light cone.

The space axis of the moving frame is defined by events that are simultaneous in S', meaning $t' = 0$.

Using $t' = \gamma\left(t - \frac{vx}{c^2}\right)$, we get $t = \frac{v}{c^2}x$.

Thus, the space axis of S' is also tilted relative to the space axis of S. This tilt represents the relativity of simultaneity. Events that are simultaneous in the moving frame lie along this tilted line, not along the horizontal axis of the stationary frame.

This geometric picture explains why time dilation, length contraction, and relativity of simultaneity are not separate assumptions. They are different consequences of the same spacetime geometry. The Lorentz transformation is the mathematical expression of this geometry.[1,6]

Relativistic Doppler Effect and Aberration

In addition to changing measurements of time and length, special relativity also changes the observed frequency and direction of light emitted by moving sources. These effects become especially important in astrophysical systems, including accretion disks around compact objects.

If a source emits radiation with frequency ν_{em}, an observer measures a frequency

$$\nu_{obs} = D\nu_{em},$$

where D is the relativistic Doppler factor. For a source moving with speed v relative to the observer, this factor can be written as

$$D = \frac{1}{\gamma(1-\beta cos\alpha)},$$

where

$$\beta = \frac{v}{c},$$

and α is the angle between the velocity of the emitting material and the direction of photon emission toward the observer. If the material moves toward the observer, the radiation is blueshifted and its frequency increases. If the material moves away from the observer, the radiation is redshifted and its frequency decreases.[7]

Relativistic aberration is the corresponding change in the apparent direction of light due to the motion of the source or observer. Radiation emitted by rapidly moving material tends to be concentrated in the direction of motion. This effect is closely related to relativistic beaming: the approaching side of a moving source can appear significantly brighter than the receding side.[7]

These effects are local consequences of special relativity. In the later section on accretion disks, they will be used to explain why the image of a disk around a black hole is not symmetric when viewed at high inclination. Even before including general relativistic light bending, the rapidly orbiting matter in the disk produces Doppler shifts, aberration, and intensity enhancement on the approaching side of the disk.

These effects are not merely conceptual. In systems where matter moves at a substantial fraction of the speed of light, especially in compact astrophysical objects, Doppler shifts, aberration, and beaming become directly relevant to the radiation observed by a distant observer.

Classical Paradoxes and Relativistic Consistency

Special relativity is often considered counterintuitive because it contradicts the everyday experience of space and time. In ordinary life, velocities are much smaller than the speed of light, so the Lorentz factor is extremely close to one and relativistic effects are practically invisible. As a result, concepts such as time dilation, length contraction, and the relativity of simultaneity appear paradoxical when first encountered.

However, the classical paradoxes of special relativity are not contradictions in the theory. Instead, they are situations in which Newtonian intuition is applied outside its domain of validity. Once the Lorentz transformations and the spacetime interval are used consistently, the apparent contradictions disappear. The following examples illustrate how special relativity preserves causality and logical consistency.

The Ladder-and-Barn Paradox

The ladder-and-barn paradox is one of the clearest demonstrations of the relativity of simultaneity. Consider a ladder with proper length L_0 moving at relativistic speed v toward a barn with proper length B_0. Suppose that in the rest frame of the barn, the moving ladder is length-contracted so that its observed length is

$$L = \frac{L_0}{\gamma},$$

where

$$\gamma = \frac{1}{\sqrt{1-v^2/c^2}}.$$

If $L < B_0$, then in the barn frame the entire ladder can momentarily fit inside the barn. In this frame, it seems possible to close the front and back doors of the barn simultaneously while the ladder is inside.

The paradox appears when the same situation is described in the rest frame of the ladder. In that frame, the ladder has its full proper length L_0, while the barn is moving and therefore length-contracted:

$$B = \frac{B_0}{\gamma}.$$

If the barn is even shorter in the ladder frame, how can the ladder ever fit inside it?

The resolution is that simultaneity is relative. The statement that both barn doors close at the same time is true only in the barn frame. In the ladder frame, the two door-closing events are not simultaneous. From the Lorentz transformation for time,

$$\Delta t' = \gamma\left(\Delta t - \frac{v\Delta x}{c^2}\right).$$

If the two door events are simultaneous in the barn frame, then $\Delta t = 0$.

Therefore,

$$\Delta t' = -\gamma \frac{v \Delta x}{c^2}.$$

Since the two doors are located at different positions, $\Delta x \neq 0$, so

$$\Delta t' \neq 0.$$

Thus, the ladder observer does not see the two doors close simultaneously. One door closes before the other. There is no contradiction, because each frame gives a different but internally consistent description of the same spacetime events. Thus, the paradox is resolved by recognizing that length contraction and simultaneity cannot be separated.[6]

The Hit-Rod Paradox and the Finite Speed of Signals

Another apparent paradox concerns the transmission of force through a solid object. Consider a long rod initially at rest. If one end of the rod is struck, it may seem natural from a Newtonian point of view to imagine that the entire rod begins moving at once. If this were true, then different observers might disagree about the order of events in such a way that the far end of the rod begins moving before the near end is struck. This would violate causality.

The resolution is that no material object is perfectly rigid in special relativity. A push or impact cannot be transmitted instantaneously through a rod. Instead, the disturbance propagates through the material as a mechanical signal, typically at the speed of sound in that material. This speed is always much smaller than the speed of light.

Let the rod have proper length L_0, and let a mechanical signal propagate through it with speed u, where $u < c$.

If the left end of the rod is struck at event A, the right end begins to respond at a later event B. In the rest frame of the rod, the time delay between these events is approximately

$$\Delta t = \frac{L_0}{u}.$$

The spacetime separation between the two events satisfies

$$c^2 \Delta t^2 > L_0^2,$$

because $u < c$. Therefore, the separation between the impact event and the response event is timelike. Timelike-separated events have the same causal order in all inertial frames. This means that every inertial observer agrees that the left end is struck before the right end responds.

This resolves the paradox. The mistake is to assume that a rod can move as a perfectly rigid object. In special relativity, perfect rigidity is impossible because it would require signals to propagate faster than light.[6] A real rod deforms: the struck end begins to move first, and the disturbance travels through the rod at a finite speed.

The hit-rod paradox therefore illustrates the causal structure of special relativity. Events connected by physical signals must remain inside or on the light cone. No force, mechanical disturbance, or information can propagate outside the light cone. This principle guarantees that cause always precedes effect.[6]

The Twin Paradox

The twin paradox demonstrates another important feature of special relativity: elapsed time depends on the path taken through spacetime. Consider two twins. One remains on Earth, while the other travels in a spaceship at high speed to a distant star and then returns. According to time dilation, the travelling twin's clock runs more slowly relative to Earth during the journey. When the twins reunite, the travelling twin is younger.[6]

At first, this seems paradoxical. From the point of view of the travelling twin, Earth appears to move away and then return. One might therefore expect the Earth twin to be younger instead. The apparent contradiction comes from treating the two twins symmetrically. They are not symmetric. The Earth twin remains approximately in one inertial frame, while the travelling twin changes inertial frames when turning around.

The proper time measured by a clock moving along a worldline is

$$\Delta\tau = \int \sqrt{1 - \frac{v^2(t)}{c^2}}\, dt,$$

where $v(t)$ is the speed of the clock relative to the chosen inertial frame. For motion at constant speed v, this reduces to

$$\Delta\tau = \frac{\Delta t}{\gamma}.$$

Thus, if the travelling twin moves at relativistic speed for a coordinate time Δt in the Earth frame, the elapsed proper time on the spaceship is smaller.

For example, suppose the spaceship travels at

$$v = 0.9c.$$

Then

$$\gamma = \frac{1}{\sqrt{1-0.9^2}} \approx 2.29.$$

If the Earth twin measures a total travel time of 10 years, then the travelling twin experiences approximately

$$\Delta\tau = \frac{10}{2.29} \approx 4.36 years.$$

The travelling twin is therefore younger when they return.

The twin paradox is best understood geometrically. In spacetime, each twin follows a different worldline between the same departure and reunion events. The Earth twin follows a nearly straight worldline, while the travelling twin follows a broken or curved worldline due to the turnaround. The proper time is analogous to the length of a path through spacetime, but because of the Minkowski geometry of spacetime, the inertial path between two timelike-separated events has the greatest proper time. Therefore, the twin who changes inertial frames experiences less time.[1,6]

Lessons from the Paradoxes

These paradoxes show that apparent contradictions arise when Newtonian assumptions are applied to relativistic situations. The ladder-and-barn paradox illustrates the relativity of simultaneity, the hit-rod paradox emphasizes causal signal propagation, and the twin paradox demonstrates the path-dependence of proper time. Together, they prepare the transition from idealized thought experiments to physical systems in which relativistic effects become observable.

From Special Relativity to Relativistic Astrophysics

The previous sections introduced the basic principles of special relativity and showed how apparent paradoxes are resolved when space and time are treated consistently. These examples are conceptually useful, but they are also idealized. Trains, ladders, rods, and clocks help isolate the logic of the theory, yet they do not always make clear where such effects become important in real physical systems.

In everyday conditions, relativistic effects are usually extremely small. For example, if an object moves with speed $v \ll c$, the Lorentz factor

$$\gamma = \frac{1}{\sqrt{1-v^2/c^2}}$$

is very close to one. In this limit, time dilation, length contraction, and relativistic Doppler shifts are negligible, and Newtonian mechanics provides an excellent approximation. This is why ordinary experience does not reveal the relativistic structure of spacetime directly.

Relativistic effects become significant when velocities approach a substantial fraction of the speed of light. Such conditions occur naturally in high-energy astrophysical environments. Examples include relativistic jets, neutron stars, compact binary systems, and accretion disks around black holes.[7,8] In these systems, matter can move rapidly enough that special relativistic effects influence the radiation received by a distant observer.

An accretion disk around a black hole provides a particularly useful example. A black hole is surrounded by a strong gravitational field, so a complete description of photon motion and spacetime geometry requires general relativity. However, the gas in the disk also moves at high orbital velocities. Therefore, several effects that shape the observed image of the disk are local consequences of special relativity.

The most important special relativistic effects in this context are the Doppler shift, relativistic aberration, and relativistic beaming. If part of the disk moves toward the observer, the emitted radiation is blueshifted. If another part of the disk moves away from the observer, the radiation is redshifted. In addition, radiation from rapidly moving material is not emitted isotropically in the observer's frame. Instead, it is beamed preferentially in the direction of motion. As a result, the approaching side of the disk can appear brighter than the receding side.[7]

These effects are described by the relativistic Doppler factor introduced earlier. In the present context, it determines how the orbital motion of the disk material changes the observed frequency and brightness of the emitted radiation. Blueshifted regions are enhanced, while redshifted regions are dimmed.

In the environment of a black hole, these special relativistic effects occur together with general relativistic effects. Gravitational redshift reduces the energy of photons escaping from deep in the gravitational field. Light bending changes the apparent position of emitting regions in the disk and can allow the observer to see parts of the disk that would be hidden in a Newtonian picture. Near the photon sphere, light may orbit the black hole before escaping, contributing to narrow ring-like features in the image.[8,14]

It is useful to keep this distinction clear: special relativity describes the local effects of rapid orbital motion, while general relativity describes gravitational redshift, light bending, the event horizon, and the photon sphere. The accretion disk example is therefore not a full model of black-hole accretion, but a physically meaningful visualization of relativistic effects.

For this reason, the following sections consider a simplified model of a geometrically thin accretion disk around a non-rotating black hole. The black hole is described by the Schwarzschild metric, while the disk material is assumed to move on approximately circular orbits.[9,12] Photons are traced from the observer backward toward the disk in order to determine which parts of the disk are visible and how their radiation is shifted and amplified.

This model provides a bridge between the conceptual foundations of relativity and observable astrophysical images. In the earlier sections, relativity appeared through thought experiments and paradoxes. In the accretion disk model, the same principles affect the appearance of a real physical system: the observed image depends on the velocity of the emitting gas, the inclination of the disk, the gravitational field of the black hole, and the curved trajectories of photons.

Accretion Disk Around a Schwarzschild Black Hole: Physical Model

To illustrate how relativistic effects appear in a physical system, we consider a simplified model of a geometrically thin accretion disk around a non-rotating black hole. The goal is not to describe the full physics of accretion, but to visualize how relativistic motion and curved spacetime influence the observed image.

We assume that the black hole is non-rotating and electrically neutral. In this case, the spacetime outside the black hole is described by the Schwarzschild solution.[8,9] This is the simplest exact black-hole solution of Einstein's field equations and provides a useful first model for studying photon trajectories and gravitational redshift.

The Schwarzschild metric is

$$ds^2 = -\left(1 - \frac{2GM}{rc^2}\right)c^2dt^2 + \left(1 - \frac{2GM}{rc^2}\right)^{-1}dr^2 + r^2(d\theta^2 + sin^2\theta d\phi^2),$$

where M is the black hole mass, G is the gravitational constant, c is the speed of light, and r is the Schwarzschild radial coordinate. The Schwarzschild radius is

$$R_s = \frac{2GM}{c^2}.$$

This radius defines the event horizon of a non-rotating black hole.[8,9] Once a photon or particle crosses this surface, it cannot return to a distant observer.

Black Hole and Disk Geometry

We model the accretion disk as geometrically thin and optically emitting.[10,12] The disk is assumed to lie in the equatorial plane, $\theta = \frac{\pi}{2}$.

This means that the vertical thickness of the disk is neglected compared with its radial size. Such a model is idealized, but it is useful because it allows the main relativistic effects to be isolated and visualized clearly.

The disk extends from an inner radius R_{in} to an outer radius R_{out}. For a Schwarzschild black hole, the innermost stable circular orbit, or ISCO,[11,12] is located at

$$R_{ISCO} = \frac{6GM}{c^2}.$$

In the present model, the inner edge of the disk is taken to be approximately $R_{in} = R_{ISCO}$.

The outer radius R_{out} is chosen to be large enough that the simulated image includes the visible emitting region of the disk. The precise value of R_{out} mainly affects the size and extent of the outer disk, while the strongest relativistic effects occur near the inner radius.

The gas in the disk is assumed to move on approximately circular orbits around the black hole. In a Newtonian approximation, the Keplerian orbital velocity is

$$v_K(r) = \sqrt{\frac{GM}{r}}.$$

Although this formula is only approximate near a black hole, it captures the important physical fact that orbital speeds increase at smaller radii. Near the inner edge of the disk, the orbital velocity becomes a significant fraction of the speed of light.[11,12] Therefore, special relativistic effects such as Doppler shifts, aberration, and beaming become important.

It is useful to introduce

$$\beta(r) = \frac{v_K(r)}{c},$$

and the corresponding Lorentz factor

$$\gamma(r) = \frac{1}{\sqrt{1-\beta^2(r)}}.$$

These quantities describe the local relativistic motion of the emitting disk material. Even though the global spacetime is curved, the emission at each point of the disk can be treated locally using special relativity in the instantaneous rest frame of the orbiting gas.

Emission and Relativistic Frequency Shifts

The disk brightness is described by a simple power-law emissivity profile,

$$I_{em}(r) = I_0\left(\frac{r}{R_{in}}\right)^{-q},$$

with a typical choice $q=3$. This makes the inner disk brighter than the outer disk and emphasizes the region where relativistic effects are strongest. The normalization I_0 is not important because the calculation focuses on relative brightness and image morphology.

Radiation emitted by the disk is shifted before it reaches the observer. Several effects contribute to the total shift. First, the orbital motion of the gas produces a special relativistic Doppler shift. If the emitting material moves toward the observer, the radiation is blueshifted. If it moves away from the observer, the radiation is redshifted.

The relativistic Doppler factor is

$$D = \frac{1}{\gamma(1-\beta cos\alpha)},$$

where α is the angle between the velocity of the emitting gas and the direction of photon emission toward the observer. The observed frequency due to this local special relativistic effect is

$$\nu_{obs} = D\nu_{em}.$$

Second, photons climbing out of the gravitational field of the black hole lose energy. For a static emitter in Schwarzschild spacetime, the gravitational redshift factor[8,9] is

$$g_{grav} = \sqrt{1 - \frac{2GM}{rc^2}}.$$

This factor becomes smaller near the black hole, meaning that photons emitted close to the event horizon are strongly redshifted.

In a simplified model, the total redshift factor may be written schematically as

$$g = \frac{\nu_{obs}}{\nu_{em}} \approx D g_{grav}.$$

This expression separates the local special relativistic Doppler factor from the gravitational redshift. A more complete general relativistic treatment computes g from the photon four-momentum and the four-velocities of the emitter[12,13] and observer, but the simplified form is sufficient for explaining the main physical contributions.

The observed intensity is also affected by the frequency shift. In relativistic radiative transfer, the quantity

$$\frac{I_\nu}{\nu^3}$$

is invariant along a photon trajectory.[7,13] Therefore, the observed specific intensity transforms as

$$I_{obs} = g^3 I_{em}.$$

This relation is important because even a moderate frequency shift can produce a strong change in observed brightness. Regions of the disk that are blueshifted are also brightened, while redshifted regions are dimmed. This is one of the main reasons why a highly inclined accretion disk appears asymmetric. These relations describe the expected effect of relativistic frequency shifts on the observed radiation. In the numerical visualization presented below, however, we calculate the photon trajectories and the geometry entering the Doppler factor, rather than a complete general relativistic map of the observed intensity.

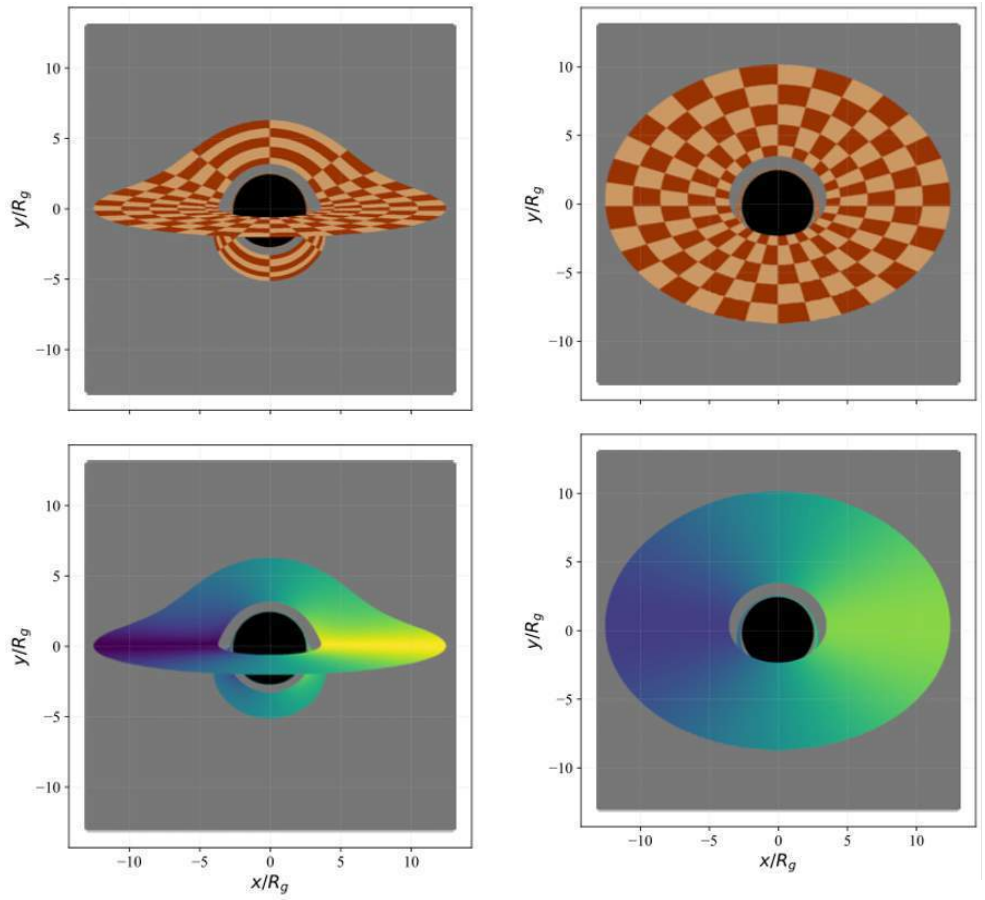


Fig.1: Numerical backward ray-tracing visualization of a geometrically thin accretion disk around a Schwarzschild black hole for two viewing inclinations. The upper panels show the apparent disk geometry distorted by gravitational light bending, while the lower panels show the angle between the photon propagation direction and the local orbital velocity of the disk material. This angle indicates where stronger Doppler blueshifting or redshifting is expected; the lower panels do not represent observed intensity.

Light Bending and Model Limitations

In Newtonian physics, photons travel in straight lines. Near a black hole, however, spacetime curvature bends photon trajectories.[8,14] This effect is essential for forming the observed image of the disk. For photon motion in Schwarzschild spacetime, one may describe the radial trajectory using the inverse radius

$$u = \frac{1}{r}.$$

A useful form of the photon orbit equation is

$$\left(\frac{du}{d\phi}\right)^2 = \frac{1}{b^2} - u^2 + \frac{2GM}{c^2}u^3,$$

where b is the impact parameter of the photon. The first two terms correspond to straight-line propagation in flat spacetime, while the final term represents the correction due to the gravitational field of the black hole.[8,14]

Light bending has several important consequences. It changes the apparent position of the disk on the observer's image plane. It can make the far side of the disk visible above and below

the black hole. It also produces strong distortions near the photon sphere. For a Schwarzschild black hole, the photon sphere is located at[8,14]

$$R_{ph} = \frac{3GM}{c^2}.$$

Photons near this radius can orbit the black hole before either escaping to infinity or falling through the event horizon. This behavior contributes to ring-like structures in the image.[14,15]

The appearance of the disk depends strongly on the inclination angle i, defined as the angle between the observer's line of sight and the axis perpendicular to the disk plane. A face-on disk has $i = 0^{\circ}$, while an edge-on disk has $i = 90^{\circ}$.

At small inclination, the line-of-sight component of the orbital velocity is small, and the disk appears nearly symmetric. At larger inclination, one side of the disk moves more directly toward the observer, while the opposite side moves away. Therefore, Doppler shifts and relativistic beaming become much stronger.[7,13]

As the inclination increases, the observed image also becomes more distorted by light bending. The far side of the disk can appear lifted above the black hole, and multiple photon paths may contribute to the observed brightness near the central shadow.

The model is deliberately simplified: it uses a Schwarzschild rather than Kerr black hole, treats the disk as thin and stationary, and does not include magnetic fields, turbulence, absorption, scattering, or time variability. Nevertheless, it captures the effects needed here: local Doppler shifts and beaming from the orbital motion of the gas, together with gravitational redshift and light bending from the Schwarzschild geometry.

Numerical Visualization and Ray-Tracing Procedure

To connect the theoretical discussion with a concrete physical example, we calculated the apparent image of a geometrically thin accretion disk around a Schwarzschild black hole. The purpose of the calculation is not to construct a complete astrophysical model of an accretion disk, but to demonstrate how gravitational light bending changes the apparent disk geometry and how this geometry determines the expected Doppler effect.

Observer and image plane

The calculation uses backward ray tracing. Instead of emitting photons from the disk in many different directions, a virtual image plane is placed in front of a distant observer and one photon trajectory is traced backward from each pixel toward the black hole. Each calculated trajectory therefore corresponds directly to one pixel of the resulting image.

All distances in the numerical calculation are expressed in units of the Schwarzschild radius,

$$r_s = \frac{2GM}{c^2},$$

with $r_s = 1$ in the numerical code. The image plane is placed at a distance

$$r_{obs} = 80r_s$$

from the black hole. Its size is $26r_s \times 26r_s$ and it is divided into 951×951 pixels. Thus, approximately 9×10^5 Photon trajectories are calculated for each image.

For each pixel, the initial photon direction is taken toward the black hole. The orientation of the image plane relative to the disk is changed to represent different viewing inclinations.

Numerical integration of photon trajectories

Photon propagation is calculated in Schwarzschild spacetime. Each photon trajectory lies in a plane containing the center of the black hole. Introducing the dimensionless inverse radial coordinate

$$u = \frac{r_s}{2r} = \frac{GM}{c^2 r},$$

the photon trajectory equation can be written as

$$\frac{d^2u}{d\phi^2} = 3u^2 - u,$$

where ϕ is the angular coordinate in the plane of the photon trajectory. The equation is solved numerically. At each step, the code stores two consecutive points of the photon trajectory, corresponding to u_{j-1} and u_j. A short trial displacement along the current photon direction is first used to determine the next angular increment. The size of this trial displacement is adaptive, $\Delta r_{trial} = 0.01r$, so that the spatial scale used to construct the next step becomes smaller as the photon approaches the black hole.

The trajectory equation is then advanced using a finite-difference scheme. For two consecutive angular increments $\Delta\phi_1$ and $\Delta\phi_2$, we define $\xi = \frac{\Delta\phi_1}{\Delta\phi_2}$.

The inverse radial coordinate of the next point is calculated as

$$u_{j+1} = \left(1 + \frac{1}{\xi}\right)u_j - \frac{u_{j-1}}{\xi} + (3u_j^2 - u_j)\Delta\phi_1\Delta\phi_2.$$

The new radial coordinate is then obtained from

$$r_{j+1} = \frac{r_s}{2u_{j+1}}.$$

Repeating this procedure step by step produces the curved photon trajectory from the observer toward the vicinity of the black hole.

Intersection with the accretion disk

The accretion disk is assumed to be geometrically thin and to lie in the equatorial plane. Its inner boundary is taken at the innermost stable circular orbit of a Schwarzschild black hole,

$$R_{in} = R_{ISCO} = \frac{6GM}{c^2} = 3r_s,$$

while in the numerical visualization the outer boundary is $R_{out} = 12r_s$.

After each numerical step, the code checks whether the segment connecting two consecutive points of the photon trajectory crosses the disk plane. If such a crossing occurs between R_{in} and R_{out}, the ray is identified as intersecting the disk and the intersection position is recorded. Otherwise, the trajectory continues until it reaches one of the stopping conditions of the calculation.

This procedure determines which point of the disk corresponds to each pixel of the observer's image and automatically includes the distortion of the apparent disk geometry caused by Schwarzschild light bending.

Geometry of the Doppler effect

For each ray intersecting the disk, the calculation also determines the angle α between the local photon direction and the direction of orbital motion of the disk material. For circular motion in the disk plane, the direction of the Keplerian velocity can be written as

$$\hat{v}_K = (-\sin\phi, \cos\phi, 0).$$

The angle is then obtained from $\cos\alpha = \hat{k} \cdot \hat{v}_K$, where $\hat{k}$ is the local direction of photon propagation.

This angle determines the geometrical dependence of the relativistic Doppler factor,

$$D = \frac{1}{\gamma(1-\beta\cos\alpha)}.$$

Consequently, the calculated map of α shows which regions of the disk are expected to experience stronger Doppler blueshifting or redshifting.

It is important to distinguish this quantity from observed intensity. The present numerical calculation directly determines photon trajectories, disk-intersection points, and the angle α; it does not calculate a complete general-relativistic intensity map I_{obs}. Modern black-hole imaging calculations use more complete general-relativistic radiative-transfer methods to calculate spectra, intensity, and polarization along photon trajectories.[18] The numerical results should therefore be interpreted as a visualization of gravitational light bending and of the geometry responsible for the expected Doppler asymmetry.

Results and Discussion: Relativistic Appearance of the Accretion Disk

Figure 1 shows the results of the backward ray-tracing calculation for two viewing inclinations. The upper panels demonstrate how photon trajectories in Schwarzschild spacetime determine the apparent geometry of the disk, while the lower panels show the angle α between the local photon direction and the Keplerian orbital velocity.

For a disk viewed at lower inclination, the apparent image is more nearly symmetric. At higher inclination, gravitational light bending produces a much stronger distortion of the disk image. Photon trajectories passing close to the black hole are strongly deflected, allowing radiation from parts of the disk that would be partly hidden in a flat-space picture to reach the

observer. The far side of the disk can therefore appear displaced and strongly curved around the central dark region.

The lower panels illustrate the connection between the ray-tracing geometry and special-relativistic Doppler effects. The angle α varies across the image because different parts of the disk have different directions of orbital motion relative to the photons that reach the observer. Since

$$D = \frac{1}{\gamma(1-\beta cos\alpha)},$$

regions where the orbital velocity is directed more nearly toward the observer have a larger Doppler factor, whereas regions moving away have a smaller Doppler factor.

The lower panels of Figure 1 should therefore not be interpreted as maps of observed brightness. They show the geometrical quantity α. Nevertheless, this quantity identifies the physical origin of the expected brightness asymmetry. Radiation from the approaching side of an inclined disk is expected to be blueshifted and enhanced, whereas radiation from the receding side is expected to be redshifted and dimmed.

The calculation therefore illustrates two different relativistic effects. General relativity determines the curved photon trajectories and hence the apparent geometry of the disk. Special relativity determines how the rapid orbital motion of the disk material affects the radiation emitted toward the observer. The present calculation directly visualizes the first effect and the geometrical factor entering the second.

This distinction is important because the present numerical model is not a complete relativistic radiative-transfer calculation. A quantitative prediction of the observed brightness would require the frequency shift to be calculated consistently along each photon trajectory and the corresponding transformation of the emitted specific intensity. Such a calculation would allow the expected Doppler asymmetry discussed here to be converted into an actual observed-intensity map.

The accretion-disk example nevertheless provides a useful connection with the earlier discussion of special relativity. The thought experiments considered above demonstrate how measurements of space, time, and simultaneity depend on relative motion. In the disk, the same relativistic principles have observable consequences for photons emitted by matter moving at a substantial fraction of the speed of light, while the strong gravitational field additionally curves the photon trajectories.

Conclusion

This paper reviewed the main principles of special relativity, beginning with the conflict between Newtonian mechanics and Maxwell's electromagnetism. The invariance of the speed of light requires Lorentz transformations rather than Galilean transformations, leading to time dilation, length contraction, relativity of simultaneity, and relativistic changes in radiation.

The classical paradoxes considered in the paper demonstrate that special relativity is internally consistent when its principles are applied carefully. The ladder-and-barn paradox shows that length contraction must be understood together with the relativity of simultaneity. The

hit-rod paradox illustrates that no mechanical signal or causal influence can propagate faster than light. The twin paradox shows that elapsed proper time depends on the path followed through spacetime. These examples help clarify why relativistic effects may appear paradoxical from a Newtonian point of view, while remaining fully consistent within the geometry of spacetime.

The final part of the paper connected these ideas to a physical system in which relativistic effects become important: an accretion disk around a black hole. Although a complete description of a black hole requires general relativity, the rapid orbital motion of the disk material also makes local special-relativistic effects important. Doppler shifts, aberration, and relativistic beaming are therefore expected to produce a strong asymmetry between the approaching and receding sides of an inclined disk.

The numerical calculation used backward ray tracing in Schwarzschild spacetime to visualize the propagation of photons from the disk to a distant observer. It demonstrates how gravitational light bending

distorts the apparent disk geometry and determines, for each visible region of the disk, the angle between the photon direction and the local orbital velocity. This angle controls the geometrical dependence of the relativistic Doppler factor and therefore identifies regions where blueshifting or redshifting is expected.

The numerical model is intentionally simplified. It assumes a non-rotating Schwarzschild black hole and a geometrically thin disk, and the images presented here do not constitute a complete general-relativistic radiative-transfer calculation. A natural next step would be to calculate the frequency shift directly from the photon four-momentum and the four-velocities of the emitter and observer and to use it to construct quantitative maps of the observed intensity. The model could also be extended to a rotating Kerr black hole and used to study systematically how the relativistic effects change with viewing inclination.

Despite these simplifications, the accretion-disk visualization provides a useful bridge between the conceptual foundations of special relativity and their astrophysical applications. It shows how the motion of rapidly orbiting matter and the curved trajectories of photons together determine what a distant observer can see near a black hole.

Data and Code Availability

The numerical results presented in this work were produced using a custom ray-tracing code developed for this study. The code is available from the authors upon reasonable request.

AI Acknowledgement

During the preparation of this work, the authors used ChatGPT (OpenAI) to improve the English language and readability of the manuscript. The scientific ideas, physical model, and numerical approach used to calculate and visualize the accretion disk are original work by the authors and were developed independently of AI tools. After using ChatGPT, the authors reviewed and edited the resulting text as necessary and took full responsibility for the content of the publication.

Works Cited

1. Taylor, E. F.; Wheeler, J. A. Spacetime Physics, 2nd ed.; W. H. Freeman and Company: New York, 1992.
2. Michelson, A. A.; Morley, E. W. On the Relative Motion of the Earth and the Luminiferous Ether. Am. J. Sci. 1887, 34, 333–345. https://doi.org/10.2475/ajs.s3-34.203.333.
3. Einstein, A. Zur Elektrodynamik bewegter Körper. Ann. Phys. 1905, 322, 891–921. https://doi.org/10.1002/andp.19053221004.
4. Ives, H. E.; Stilwell, G. R. An Experimental Study of the Rate of a Moving Atomic Clock. J. Opt. Soc. Am. 1938, 28, 215–226. https://doi.org/10.1364/JOSA.28.000215.
5. Hafele, J. C.; Keating, R. E. Around-the-World Atomic Clocks: Observed Relativistic Time Gains. Science 1972, 177, 168–170. https://doi.org/10.1126/science.177.4044.168.
6. Rindler, W. Introduction to Special Relativity, 2nd ed.; Clarendon Press: Oxford, 1991.
7. Rybicki, G. B.; Lightman, A. P. Radiative Processes in Astrophysics; Wiley: New York, 1979.
8. Misner, C. W.; Thorne, K. S.; Wheeler, J. A. Gravitation; W. H. Freeman and Company: San Francisco, 1973.
9. Schwarzschild, K. Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie. Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.) 1916, 189–196.
10. Shakura, N. I.; Sunyaev, R. A. Black Holes in Binary Systems. Observational Appearance. Astron. Astrophys. 1973, 24, 337–355.
11. Bardeen, J. M.; Press, W. H.; Teukolsky, S. A. Rotating Black Holes: Locally Nonrotating Frames, Energy Extraction, and Scalar Synchrotron Radiation. Astrophys. J. 1972, 178, 347–369. https://doi.org/10.1086/151796.
12. Page, D. N.; Thorne, K. S. Disk-Accretion onto a Black Hole. Time-Averaged Structure of Accretion Disk. Astrophys. J. 1974, 191, 499–506. https://doi.org/10.1086/152990.
13. Cunningham, C. T. The Effects of Redshifts and Focusing on the Spectrum of an Accretion Disk around a Kerr Black Hole. Astrophys. J. 1975, 202, 788–802. https://doi.org/10.1086/154033.
14. Luminet, J.-P. Image of a Spherical Black Hole with Thin Accretion Disk. Astron. Astrophys. 1979, 75, 228–235.
15. Falcke, H.; Melia, F.; Agol, E. Viewing the Shadow of the Black Hole at the Galactic Center. Astrophys. J. Lett. 2000, 528, L13–L16. https://doi.org/10.1086/312423.
16. Event Horizon Telescope Collaboration. First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole. Astrophys. J. Lett. 2019, 875, L1. https://doi.org/10.3847/2041-8213/ab0ec7.
17. Event Horizon Telescope Collaboration. First Sagittarius A* Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole in the Center of the Milky Way. Astrophys. J. Lett. 2022, 930, L12. https://doi.org/10.3847/2041-8213/ac6674.
18. Prather, B. S.; Dexter, J.; Mościbrodzka, M.; et al. Comparison of Polarized Radiative Transfer Codes Used by the Event Horizon Telescope Collaboration. Astrophys. J. 2023, 950, 35. https://doi.org/10.3847/1538-4357/acc586.

Invisible in the Data: Middle Eastern and North African (MENA) Women and Pharmacogenomic Uncertainty in Clinical Drug Research

By Arianna M. Brown

Abstract

Historically, individuals of Middle Eastern and North African (MENA) descent were classified by the US government as "white" for purposes of collecting and presenting data on race and ethnicity. Only recently, in 2024, did the federal government create "Middle Eastern or North African" as a new, distinct category. For clinical drug research conducted by the government or by private parties following the government's pre-2024 classifications, this has resulted in the aggregation of data regarding subjects of MENA descent and subjects of white European descent. At the same time, women have long been underrepresented in clinical research, and the available evidence suggests that individuals of MENA descent are underrepresented as well. While the field of pharmacogenomics has established that genetic makeup may affect the way an individual's body processes drugs, which can allow clinicians to make more personalized decisions about the prescription and dosing of medications, these decisions depend on the quality and breadth of the data concerning particular patient populations. There has been some specialized pharmacogenomic research focusing on certain portions of the MENA population, but the research has not been extensive. The confluence of this data aggregation and underrepresentation in clinical research represents a clinical evidence gap affecting MENA women and raises an important, insufficiently studied question as to whether MENA women are at a higher risk of adverse drug reactions (ADRs). I argue—drawing on scholarship on the theory of Orientalism, the racialization of individuals of MENA descent, and the colonial foundations of Western medicine—that the clinical evidence gap affecting MENA women and the uncertainty concerning their ADR risk can be understood as a structural outcome rather than merely the result of isolated oversights. A more equitable approach requires disaggregated data collection and reporting, efforts to increase the participation of women and people of MENA descent in clinical research, and attentiveness to the variations in the way MENA women may process particular drugs. This approach would also advance the promise of more personalized medicine.

Keywords Pharmacogenomics, Middle Eastern and North African (MENA), Orientalism, colonial medicine, adverse drug reactions (ADRs)

Introduction

Modern medicine increasingly holds the promise of personalization. The field of pharmacogenomics studies how genetic variation can affect an individual's response to medication and can guide clinicians in their selection of a particular drug or dosage. But the quality of personalized medicine depends on the quality and breadth of the data that informs these clinical decisions. Among other concerns, without data of sufficient quality and breadth for

a particular patient population, there is potentially a risk that they may experience a higher incidence of adverse drug reactions (ADRs). There is an important but insufficiently studied question as to whether deficiencies in data affecting women of Middle Eastern and North African (MENA) descent mean they are at a higher risk of ADRs. There are specific genetic variants seen in this population (or, more accurately, in subgroups within this population) that, if not accounted for when researching and prescribing drugs, can lead to unknown and severe ADRs (Al-Mahayri et al. 1).

The underlying data deficiencies—a clinical evidence gap—stem both from unwarranted data aggregation and from underrepresentation in clinical drug research. As to data aggregation, historically, individuals of MENA descent were classified by the US government as "white" for purposes of collecting and presenting data on race and ethnicity. It was not until very recently, in 2024, that the federal government created "Middle Eastern or North African" as a new seventh category, distinct from the "white" category, while acknowledging that, in this context, the categories are sociopolitical constructs and not an attempt to define race and ethnicity biologically or genetically (United States, Federal Register 22182, 22185-86, 22191-92). For clinical drug research conducted by the government or by private parties following the government's pre-2024 classifications, this has had the effect of aggregating data regarding subjects of MENA descent and subjects of White European background (or subjects of other backgrounds who did not fall neatly into any of the other five categories and therefore were placed in the "white" category).

In addition to the problem of data aggregation, MENA women have been underrepresented in clinical drug research. It is well established that women are generally underrepresented in this research, and the available evidence suggests that the MENA population as a whole is underrepresented as well (Abouhala et al., "Facilitators and Barriers" 359-60; Abouhala et al., "Our Voices Will Get Louder" 418, 428). More recently, scholars have written and analyzed the sex-based disparities in clinical research and their effects (Bøttern et al.; Zucker et al., "Pervasive Neglect of Sex Differences"; Zucker and Prendergast), but there has been relatively little research on the underlying mechanisms that cause each sex to respond differently to treatments. Many of the historical justifications for lower representation of women in clinical research were unsound (Zucker et al., "Pervasive Neglect of Sex Differences" 1-2). There has been some specialized pharmacogenomic research focusing on certain portions of the MENA population (Fleischer and Sadek; Kalbasi; Neyshaburinezhad et al.; Jithesh et al.), but the research has not been extensive.

In this paper, I examine the potential causes and foundations of the clinical evidence gap affecting MENA women and the uncertainty concerning their ADR risk. I argue that they can be understood as a structural outcome, reflecting vestiges of "Orientalist" and colonialist ideas and values that became embedded in Western medicine, rather than merely the result of isolated oversights. The current situation is connected to an institutional history in which Western authorities determined which populations were prioritized, whose knowledge was credible, and whose bodies would be considered the reference point for medical research and clinical

decisions. The theory of Orientalism helps to explain how white Europeans developed a cultural and political construction of MENA populations, and particularly MENA women, as a deprioritized "other." The history of colonial medicine helps to explain how similar hierarchies, whether intentionally or not, became a facet of research practices in the Western medical field.

The paper is laid out in four main sections. First, I explore existing literature addressing the racialization of the MENA population, inequalities in healthcare, inadequate clinical research data specific to MENA women, and the growing pharmacogenomic approach to the prescription and dosing of drugs. Second, I discuss scholarship on Orientalism and colonialism and how related ideas and values historically influenced Western medicine. Third, I discuss the clinical evidence gap facing MENA women, owing to two distinct problems: underrepresentation in clinical research and the aggregation of data concerning MENA research subjects and non-MENA subjects in a broad "white" category. I argue that this clinical evidence gap reflects remnants of Orientalist and colonialist ideas and values that historically became embedded in Western medicine and have persisted, even if unknowingly. Fourth, I provide an concrete example of why the clinical evidence gap matters. I discuss pharmacogenomics, the study of how a person's DNA affects their body's response to medications, and examine two studies—of populations in Iran and Qatar—that illustrate why subjects of European descent should not simply be presumed to represent patients in various MENA populations given empirical genetic variations affecting biological processing of certain drugs. I conclude by offering thoughts on deliberate steps that can be taken to work toward an approach that is more equitable for MENA women while advancing the promise of more personalized medicine.

Literature Review

Formal imperial control by Western governments was often effectuated through colonization of foreign regions that the West deemed inferior or in need of saving. A key characteristic of colonialism "involves the arrogation of sovereignty by a conquering power" (Steinmetz 79). Existing literature tends to analyze the process of globalization through this lens, focusing on the structures the colonizers have imposed to subjugate the "inferior" population in terms of economy, labor and production, culture, law, race, and more (Bates and Huerta 71-2). Historian Partha Chatterjee coined the term "rule of difference" to explain the social construction of power where "colonial states divide their subjects into different tribal or racial groups in an effort to enhance control, but at the same time the colonized are subsumed by the colonial state under a single, overarching category" (Steinmetz 80). Sociologist Anibal Quijano argues that one of the elementary features of globalization is "the social classification of the world's population around the idea of race" (551). Ahmad Sa'di discusses the ways colonial hierarchies show up in the modern, globalized world.

The racialization of people of color and the racial hierarchies imposed upon them have often been analyzed through colonial-projects and white-superiority constructs, but it has not been thoroughly analyzed through a public health lens focusing on the MENA population. Scholars such as Maghbouleh have analyzed the contentious classification of the MENA

population throughout the 20th century, and how the classification as "White" was not reflective of public sentiment, but rather a strategy for gaining citizenship in the discriminatory US legal process (Maghbouleh et al. 2). Hadi Khoshneviss presents a key argument about MENA categorization as "White," explaining how the "legally white" MENA community has been marginalized and deemed an "inferior white" and the "other" (120).

Public health has always been shaped by the colonial systems around them: funding allocation, segregation, and medical and scientific justification are used to strengthen hierarchies. Races were conceptualized using Western medicine as justification that was to be unquestioned, and diseases were claimed to pervade specifically Arab countries to perpetuate prejudices about the group (Amster E709). Du Bois, in his 1911 work "The Negro Race in the United States of America," considered the racial underpinnings of health disparities (Patel 121). Du Bois described the "global color line" as a fundamental construct that shapes culture, politics, and more against "darker" races (Patel 119). In fact, research has found a concrete correlation between race and ethnicity and the occurrence of medical errors (Suen et al. 2163). Existing, though limited, studies have identified a general hesitation in many MENA communities to participate fully in healthcare and research (Abouhala et al., "Facilitators and Barriers" 360; Abouhala et al., "Our Voices Will Get Louder" 419). Existing literature discussing diversity in clinical research typically categorizes its data into six main racial or ethnic groups: White, Asian, Black, Latino/a, and recently American Indian or Alaska Native, and Native Hawaiian or Pacific Islander. A major gap in the literature is reporting on the Middle East and North Africa as its own category. There is already an underrepresentation of MENA participants in clinical trials, but even when there is participation, their data gets hidden in the "White" category (Abouhala et al., "Facilitators and Barriers" 359). Therefore, a significant underrepresentation of MENA bodies appears in healthcare, which both fundamentally harms MENA patient experiences and perpetuates the structures in place that prevent their participation and trust.

Existing literature has begun to endorse the use and continued research on a pharmacogenomic approach to medicine. The variants of enzymes that metabolize most prescribed medications are regionally and ethnically quite variable, and clear differences in genetic variant frequencies can be seen across MENA, even more so when compared to Europeans (Jithesh et al. 5). A 2021 study by Neyshaburinezhad et al. concluded that, contrary to the commonly held belief that Iranians share great similarities with European whites, their frequencies of different genotypic variants of the nine main enzymes are significantly distinct from any of the five other general regional groupings (10-11). There have been only a few dedicated studies on the pharmacogenomics of MENA populations, and this new research suggests there is a substantial need for more MENA-specialized research (Neyshaburinezhad et al. 2). This being said, most do not incorporate this data into the larger picture to begin to separate their differences from the "white" category. In their book Inflamed, Marya and Patel analyze the construction of Western medicine and how the hierarchies that colonizers embedded into modern Western medicine have had persistent effects (1-20, 186-88, 209).

Historically, women's representation in clinical research has been inadequate. Existing literature has more recently explored the neglect of female bodies and experiences in the field, and reports a continued disproportionate representation compared to the reference population (Bøttern et al. 3-5; Zucker et al., "Pervasive Neglect of Sex Differences"). A historic justification for this disparity relies on refuted claims of increased variability in female bodies and concerns about reproductive health (Zucker et al., "Pervasive Neglect of Sex Differences" 2-4). Few studies have tracked specific sex-based differences in reactions to medications in drug trials. This neglected area is critical to understanding why and how female metabolisms differ and the subsequent treatment adjustments that need to be made. The available evidence suggests that MENA individuals, both women and men, are also underrepresented in clinical research (Abouhala et al., "Facilitators and Barriers" 359-60; Abouhala et al., "Our Voices Will Get Louder" 418, 428). MENA women are therefore affected by both limited sex-specific data and limited MENA-specific data. This paper aims to analyze the potential causes and effects of this gap.

Historical Orientalist and Colonial Values in Western Medicine

Edward Said described Orientalism as "a style of thought based upon an ontological and epistemological distinction made between 'the Orient' and (most of the time) 'the Occident'" (2). Said viewed the late eighteenth century as a rough starting point for Orientalism as a Western institution for describing, governing, and exercising authority over the Orient (3, 5). The "Orient" included the Middle East and North Africa. Through the construction of this dichotomy by white Europeans, the MENA region was flattened to one meaning encompassing the various cultures, religions, and ethnicities (Said 3). European Orientalists constructed various narratives about this "other," and, over time, the narratives were sustained and deepened through repetition and confirmation bias (Said 6-8, 14, 16, 20). The MENA region and particularly MENA women were characterized as exotic, meek, sensual, and mysterious (Said 6-7, 26, 187). Westerners focused on the parts of the culture that were entertaining for them, while ignoring or downplaying the genuine, rich culture that existed throughout the region (185-88). The fetishization of Middle Eastern women eventually turned to pity as the narrative shifted to vilify the men as perpetrators and paint the women as oppressed, using the veil as a visual symbol of this perceived and forced placidity and helplessness (Aljuaid 21). The Orientalist perception of MENA women was one of meekness and helplessness (Said 187, 206-07). In this way, Orientalism not only acted to solidify the perceived authority of the West but also allowed Orientalists to take on the self-created role of white saviors (Said 18-21, 206-07).

Said did not argue that Europeans simply held inaccurate stereotypes of the Orient, but rather that Orientalism was in part a system for producing knowledge about a region that European powers claimed the right to describe, classify, and govern. Said wrote: "It is Europe that articulates the orient" (57). Orientalists imposed their systems onto those they deemed "other," in an act of "intellectual colonialism" (Aljuaid 22). This intellectual colonialism reinforced imperial colonial power structures. The dichotomy created was meant to establish the

West's superiority by painting the Middle East as the opposite of what the West claims to be: civilized and culturally and intellectually superior (Aljuaid 18, 20). Classification and categorization are not neutral; they influence who and what becomes visible, what questions researchers and scholars ask, and whose interests are prioritized. Colonial governments often used racial, tribal, and cultural classifications to organize populations and justify their authority and preeminence. Scholars have described this as a "rule of difference" that classified populations, linked "race" to labor and resources, and positioned proximity to whiteness as superior (Bates and Huerta 73).

For Said, colonial power determined who possessed the authority to know and dictate the Orient; for Marya and Patel, colonial medicine determined whose body and knowledge were prioritized and could constitute authoritative medical truth. Global health has been entangled with colonialism from the beginning of its Westernization. The concept of "colonial medicine," as discussed by Marya and Patel, refers to a Western model of medicine shaped by colonial power, where medical frameworks are used to divide and empower the "dominant" group, Western knowledge is prioritized through conformity, and disease is constrained to the body (Marya and Patel 17-20). Under this conception, the West's "superior" information overruled existing knowledge (indigenous medicine), and Western power structures were imported into medicine (Marya and Patel 19). Colonial medicine divided rational human subjects—white European men—from "nature"—women and people of color—connecting this hierarchy to race, gender, and medicine. In this system, Europe exercised the power to describe, interpret, and classify (Said 20, 57). Under the "colonial medicine" theory, medicine was not a benevolent tool, but an implementation of European control over narrative and the question of whose evidence and knowledge counts (Marya and Patel 64-70, 206-09; Said 20, 57).

Tropical medicine was one historical manifestation of colonial medicine. When colonizers took up residence in foreign regions, specifically in tropical countries, they were susceptible to new diseases particular to the area. To combat this problem, and to enable the continuation of the colonial project, tropical medicine was established. By co-opting and reorganizing medical knowledge, often taking information from native communities, Western elites enforced who was worthy and credible of dictating a field as noble and factual as medicine. In this new medical framework, "bourgeois men's bodies [were] a metonym for everyone's bodies" (Marya and Patel 209), and flattening MENA to one homogenous connotation empowered the West to control the narrative (Said 6, 33).

The US specifically also played a role in establishing racial hierarchies in the field of medicine. For example, in the late 1840s, white scientist Dr. James Marion Sims, considered the "father of modern gynecology," conducted clinical research by nonconsensually using enslaved black women as test subjects—an unconscionable and atrocious practice (Amster E708; Shende et al. 4). Throughout much of the history of medicine in the US, white men have been the model patients, and the ones whose bodies were considered when procedures and treatments were formulated. This has had a significant effect insofar as the US houses a large number of globally renowned academic and scientific institutions (Shende et al. 4).

The Clinical Evidence Gap Affecting MENA Women: Underrepresentation and Data Aggregation, Understood as a Structural Outcome

The colonial roots of Western medicine have effects to this day. While the structures of formal imperial control are largely in the past, vestiges of certain ideas and values that were embedded in medicine in the colonial era have persisted, even if unintentionally. As discussed, Orientalism should be understood not merely as a collection of prejudices, but rather as a set of institutional structures that afforded Western authorities, among other advantages, control over classification and categorization, which in turn influenced who and what became visible, what questions researchers and scholars asked, and whose interests were prioritized. MENA women have faced a clinical evidence gap due to underrepresentation in clinical research and aggregation of their data in a broader "white" category. More precisely, MENA women are affected by two evidence gaps at once: limited sex-specific data and limited MENA-specific data. In this sense, MENA women are "invisible in the data."

As to underrepresentation in clinical research, although there has been a marginal increase in women's inclusion in clinical trials in the past 50 years (Zucker et al., "Pervasive Neglect of Sex Differences" 2), a study done by Bøttern et al. found that female participation in drug trials has remained unchanged in the last 20 years, and they are still grossly underrepresented in clinical trials (940). There are barriers specific to women's participation that point to how systems in the US and around the world continue to reflect patriarchal values which prioritize men and their well-being over women's. Researchers, up until recently, have incorrectly held that men are representative of the whole population, and that sex is not relevant in this research. Now, sex-based differences in pain signaling, drug metabolism, and other areas are garnering more attention (Zucker et al., "Pervasive Neglect of Sex Differences" 1). While the reasoning often forwarded to explain or justify women's exclusion was that women are more variable because of their hormonal fluctuation, empirical evidence proves that there is no significant inherent increased variability in women, nor is fertility risk a reason to exclude them from research (Zucker et al., "Pervasive Neglect of Sex Differences" 2). Even when women have been included in clinical research, most studies have not provided sex-specific data analyses (Zucker and Prendergast, "Sex Differences in Pharmacokinetics" 2).

The exclusion of women from research risks their safety by not considering the interactions between drugs and, among other sex-based differences, women's hormones, making it potentially unsafe for women to use the drug once available to consumers. When drugs are tested, tailored, and manufactured based on trials overrepresenting men, the drugs are already female-biased for ADRs because women metabolize drugs differently. Women chronically experience toxicity due to overdosing, which often results from dosing recommendations based on the male metabolism (Zucker and Prendergast, "Sex Differences in Pharmacokinetics" 2-3). Indeed, women are reported to experience ADRs almost twice as often as men across all drug classes and are significantly more likely to be hospitalized due to an ADR (Zucker and Prendergast, "Sex Differences in Pharmacokinetics" 1, 3).

Additionally, the available evidence suggests that MENA individuals, both women and men, are also underrepresented in clinical research for various reasons, including hesitancy to participate (Abouhala et al., "Facilitators and Barriers" 359-60; Abouhala et al., "Our Voices Will Get Louder" 418, 428).

As to data aggregation, to the extent MENA women have participated in clinical research, their representation has been obscured because their data has been included within a broad "white" category rather than a narrow, more tailored category. Historically, until very recently, individuals of MENA descent were classified by the US government as "white" for purposes of maintaining, collecting, and presenting data on race and ethnicity, under the Office of Management and Budget (OMB) Statistical Policy Directive No. 15 (SPD 15). In 2024, after a nearly two-year process considering changes to SPD 15, the government created a new "Middle Eastern or North African" category, separate and distinct from the "white" category, while acknowledging that the categories are sociopolitical constructs and not an attempt to define race and ethnicity biologically or genetically (United States, Federal Register 22182, 22185-86, 22191-92). In urging this change, some public commenters expressed "that the current [pre-2024] classification of MENA respondents as White does not reflect the reality of many who are MENA," whereas a few opponents of the change expressed that "some individuals from the MENA region of the world do consider themselves to be White" (United States, Federal Register 22185). Within this new MENA category, SPD 15 also authorizes data to be collected for narrower subcategories, including "Lebanese, Iranian, Egyptian, Syrian, Iraqi, Israeli, [or] Another group (for example, Moroccan, Yemeni, Kurdish, etc.)" (United States, Federal Register 22192-93). These detailed categories encompass Arab nationalities in the United States from the Middle East, Arab nationalities in the United States from North Africa, and non-Arab nationalities in the United States from the MENA region (United States, Federal Register 22186-87, 22191-93)—reflecting the diversity even within MENA populations. "[A] number of health organizations expressed the importance of having data available for the detailed groups to measure differences in healthcare outcomes" (United States, Federal Register 22186). (Notably, SPD 15 was also changed to replace the term "Far East" with "Central or East Asia" (United States, Federal Register 22187).)

This historical data aggregation has had meaningful impacts. For clinical drug research conducted by the government or by private parties following the government's pre-2024 classifications, this has had the effect of aggregating data regarding subjects of MENA descent and subjects of white European background (or subjects of other backgrounds who did not fall neatly into any of the other five categories and therefore were placed in the "white" category). Sponsors conducting clinical trials intended to support applications for US Food and Drug Administration (FDA) approval commonly collect and report demographic data using the categories in SPD 15, which allows them to address FDA requirements for submission of safety and efficacy data (United States, Collection of Race and Ethnicity Data in Clinical Trials). Thus, with MENA individuals historically classified as "white," there has been limited tracking of the MENA demographic specifically. As a result of these gaps, there is a serious question as to

whether MENA women may be at an increased risk of experiencing ADRs, although this precise question needs further study.

The clinical evidence gap affecting MENA women and the uncertainty concerning their ADR risk can be understood as echoes of Orientalist and colonialist ideas and values that historically influenced Western medicine, rather than merely the result of isolated oversights (although, to be sure, those may sometimes be contributing factors). The current state of affairs reflects a history in which, as discussed in greater detail above, Western medical institutions prioritized people of European descent, considered white men to be the primary reference point for medical research and clinical decisions, and operated on the assumption that Western authorities could speak for MENA women more reliably than MENA women could speak for themselves. The effects of this history remain visible. It should be acknowledged that many strides have been made since the late eighteenth century as to the perception and treatment of women and people of MENA descent. And the presence of these structural issues does not mean that contemporary researchers intend to reproduce colonial prejudices. Yet the ideas and values from earlier eras can be remarkably persistent, passed down across generations, sometimes unknowingly, even while evolving alongside contemporary thinking and norms.

These structural issues have given rise to surprising and, at times, ironic circumstances. Women were sometimes treated as sufficiently similar to men that their inclusion in drug research was supposedly unnecessary, while also being treated as sufficiently different that their inclusion would complicate research (Marya and Patel 209). For their part, MENA individuals were classified for many years by the US government and in many clinical studies as "white," yet many MENA Americans either have not been perceived or treated socially as white, or have been deemed "inferior white" (Khoshneviss 120). And when MENA participants are reported as white, their presence may effectively become invisible in the data even if they were included in the study.

An Illustration of Why the Clinical Evidence Gap Matters: Pharmacogenomic Uncertainty and the MENA Population

Pharmacogenomics, which studies how a person's DNA affects their body's response to medications, shows us that particular genes can influence enzymes and transporters that are involved in drug absorption, metabolism, and clearance. A genetic variant may affect how an individual processes a drug, which can influence the drug's efficacy and toxicity risk. Pharmacogenomics provides one concrete reason to take seriously the need for greater representation of the MENA population in clinical drug research and the need for disaggregated MENA-specific data. Two illustrative studies show why this is so.

The first study involved gene sequencing data of Iranians. The study focused on the Cytochrome P450 (CYP450) enzyme genes that are responsible for metabolizing over 90% of commonly prescribed drugs. The authors compared the frequencies of actionable variants of nine important CYP450 enzymes in the Iranian population (800 unrelated Iranians, 100 people from each of eight ethnic groups) to those of five major human populations (European,

Latino/Admixed American, African, East Asian, and South Asian) (Neyshaburinezhad et al. 2-3). Although the results were quite granular, the study found that Iranian CYP450 genetic variant frequencies are, overall, significantly distinct from those of the other populations, and concluded that the differences in pharmacogenomics between the two groups are substantial (2, 8-11). The results were not uniformly similar to, or different from, any one comparison group, but certain variants were significantly dissimilar to the European population included in the study (Neyshaburinezhad et al. 2, 8-11).

A larger Qatari study offers a second illustrative example. This 2022 study used 6,045 whole genomes from the Qatari population, and the authors analyzed 2,629 variants in 1,026 genes known to affect 559 drugs or classes of drugs (Jithesh et al. 8). They found that 1,320 variants in 703 genes affecting 299 drugs or classes of drugs were significantly different between the Qatari population and other world populations (Jithesh et al. 2). The authors also measured clinically actionable diplotype frequencies of the gene SLCO1B1 in the Qatari population. The SLCO1B1 gene encodes the OATP1B1 transporter protein, which, among other functions, moves certain drugs, including statins, from the blood into liver cells. A variant that reduces transporter function can affect drug exposure and side effects (Jithesh et al. 2). The study's empirical data concludes that the occurrence of the actionable diplotypes in the Qatari population was more than double that of other world populations being compared (Jithesh et al. 2, 5). A total of nine genes had distinctly different actionable diplotypes in the study's population compared to European data from an extensive genome dataset, with six of them having higher frequencies in the Qatari population (Jithesh et al. 2). The study also concluded there was a diverse distribution of actionable genotype/diplotype frequencies even among Qatari subpopulations (Jithesh et al. 5).

These studies illustrate why European reference frequencies should not simply be presumed to represent, in these cases, patients of Iranian or Qatari descent. It is also the case that MENA, which is a category that itself encompasses a wide diversity of people, should not be treated as having a single genetic profile—indeed, there can be material genetic variation even within a single country. Because specific data for the MENA population (and, in particular, for MENA women) is relatively limited, it introduces serious pharmacogenomic uncertainty for the various constituents of this population and presents the possibility of increased ADRs for particular drugs, an area that warrants further study. More broadly, it presents a challenge for this population to share equitably in the promise of personalized medicine.

Conclusion

I first reviewed existing literature addressing the racialization of the MENA population, inequalities in healthcare, inadequate clinical research data specific to MENA women, and the growing pharmacogenomic approach to the prescription and dosing of drugs. I then discussed the concepts of Orientalism and colonialism and explored how related ideas and values historically influenced Western medicine. Critically, Said did not argue that Europeans simply held inaccurate stereotypes of the "Orient" (which encompassed MENA), but rather that Orientalism

was in part a system for producing knowledge about a region that European powers claimed the right to describe, classify, and govern. This intellectual colonialism claimed the right to classification and categorization, which influences who and what becomes visible, what questions researchers and scholars ask, and whose interests are prioritized. Next, I discussed the clinical evidence gap facing MENA women, owing to two distinct problems: underrepresentation in clinical research and the aggregation of data concerning MENA research subjects and non-MENA subjects in a broad "white" category. I argued that this clinical evidence gap—which frustrates MENA women's ability to reap the benefits of personalized medicine and raises uncertainty about their risk of experiencing adverse drug reactions (ADRs)—reflects remnants of certain Orientalist and colonialist ideas and values that historically became embedded in Western medicine and have persisted, even if unintentional or at times hidden. Next, I discussed pharmacogenomics, the study of how a person's DNA affects their body's response to medications, and examined two studies of populations in Iran and Qatar, respectively. These studies illustrated why subjects of European descent should not simply be presumed to represent patients in various MENA populations given genetic variations affecting biological processing of certain drugs. These studies also emphasize the point that MENA populations are not monolithic. MENA is a category that encompasses a wide diversity of people and does not have a single genetic profile. In fact, the studies showed there can be material genetic variation even within a single country. Because specific data for the MENA population (and, in particular, for MENA women) is limited, it introduces serious pharmacogenomic uncertainty for the constituents of this population. Further study in this area is warranted to explore whether MENA women face a risk of increased ADRs for particular drugs.

It is important for those conducting drug research to be aware of these structural and at times obscured biases and to take deliberate corrective steps to work toward an approach that is more equitable for MENA women. First, researchers should ensure adequate representation by women and individuals of MENA descent in studies—understanding that MENA is itself a broad classification, one including people with a substantial range of genetic variation, and therefore more granular thinking about study design and selection of participants may be necessary. Greater participation may require outreach and educational efforts as well as overcoming practical impediments to participation, such as language barriers, trial locations, and hesitancy on the part of members of the MENA population. Second, researchers should employ disaggregated data collection and reporting—at minimum, tracking the 2024 US government change that created "Middle Eastern or North African" as a new category, distinct from the "white" category, even though these categories are sociopolitical constructs rather than an attempt to define race and ethnicity biologically or genetically. Third, both researchers and clinicians should be attentive to the variations in the way MENA women, as distinct from other patient populations, may process particular drugs. Taking these steps would address the identified shortcomings facing MENA women and would advance the promise of more personalized medicine.

Works Cited

Abouhala, Siwaar, et al. “Facilitators and Barriers to Health Research Knowledge and Participation among Arab/Middle Eastern and North African (MENA) Patients in the US.” Journal of Community Health, vol. 50, no. 2, Apr. 2025, pp. 358–68. DOI.org (Crossref), https://doi.org/10.1007/s10900-024-01423-9.

Abouhala, Siwaar, et al. “‘Our Voices Will Get Louder’: A Qualitative Exploration of Factors Influencing Arab/Middle Eastern North African (MENA) American Patient Participation in US-Based Health Care and Health Research.” Journal of Racial and Ethnic Health Disparities, vol. 13, no. 1, Feb. 2026, pp. 418–31. DOI.org (Crossref), https://doi.org/10.1007/s40615-024-02255-y.

Aljuaid, Sara. “Confronting Orientalism in American Media.” Revista Adelante-Ahead, vol. 12, no. 1, 2023, pp. 17–24.

Al-Mahayri, Zeina N., et al. “Pharmacogenomic Diversity among Arab Populations: A Systematic Review.” iScience, vol. 29, no. 4, Apr. 2026, p. 115191. DOI.org (Crossref), https://doi.org/10.1016/j.isci.2026.115191.

Amster, Ellen J. “The Past, Present and Future of Race and Colonialism in Medicine.” Canadian Medical Association Journal, vol. 194, no. 20, May 2022, pp. E708–10. DOI.org (Crossref), https://doi.org/10.1503/cmaj.212103.

Bates, Taryn, and Rubén Huerta. “Racial Hierarchies Globalized: From Colonialism to Informal Empire.” California Sociology Forum, vol. 6, no. 1, June 2024, pp. 68–82.

Bøttern, Julie, et al. “Sex, Racial, and Ethnic Diversity in Clinical Trials.” Clinical and Translational Science, vol. 16, no. 6, Mar. 2023, pp. 937–45. PubMed Central, https://doi.org/10.1111/cts.13513.

Fleischer, Nicole J., and Katherine Sadek. “Arab, Middle Eastern, and North African Health Disparities Research: A Scoping Review.” Journal of Racial and Ethnic Health Disparities, vol. 12, no. 3, June 2025, pp. 1397–415. DOI.org (Crossref), https://doi.org/10.1007/s40615-024-01972-8.

Jithesh, Puthen Veettil, et al. “A Population Study of Clinically Actionable Genetic Variation Affecting Drug Response from the Middle East.” Npj Genomic Medicine, vol. 7, no. 1, Feb. 2022, p. 10. www.nature.com, https://doi.org/10.1038/s41525-022-00281-5.

Kalbasi, Shaida. “Health Care Discrimination Experiences of Middle Eastern Women.” Critical Studies: An International and Interdisciplinary Journal, vol. 13, no. 1, May 2017, pp. 5–23. ojs.scholarsportal.info, https://doi.org/10.51357/cs.v13i1.128.

Khoshneviss, Hadi. “The Inferior White: Politics and Practices of Racialization of People from the Middle East in the US.” Ethnicities, vol. 19, no. 1, Feb. 2019, pp. 117–35. DOI.org (Crossref), https://doi.org/10.1177/1468796818798481.

Maghbouleh, Neda, et al. “Middle Eastern and North African Americans May Not Be Perceived, nor Perceive Themselves, to Be White.” Proceedings of the National Academy of Sciences, vol. 119, no. 7, Feb. 2022, p. e2117940119. pnas.org (Atypon), https://doi.org/10.1073/pnas.2117940119.

Marya, Rupa, and Raj Patel. Inflamed: Deep Medicine and the Anatomy of Injustice. Picador, 2022.

Neyshaburinezhad, Navid, et al. “Frequency of Important CYP450 Enzyme Gene Polymorphisms in the Iranian Population in Comparison with Other Major Populations: A Comprehensive Review of the Human Data.” Journal of Personalized Medicine, vol. 11, no. 8, Aug. 2021, p. 804. www.mdpi.com, https://doi.org/10.3390/jpm11080804.

Patel, Payal K. “On the Path of the Maharajah of Bwodpur: The Global Problem of the Color Line in W. E. B. Du Bois’s Dark Princess.” CR: The New Centennial Review, vol. 15, no. 2, 2015, pp. 119–56. JSTOR, https://doi.org/10.14321/crnewcentrevi.15.2.0119.

Quijano, Aníbal. “Coloniality of Power, Eurocentrism, and Latin America.” Nepantla: Views from South, vol. 1, no. 3, 2000, pp. 533–80.

Sa’di, Ahmad H. “Orientalism in a Globalised World: Said in the Twenty-First Century.” Third World Quarterly, vol. 42, no. 11, Nov. 2021, pp. 2505–20. DOI.org (Crossref), https://doi.org/10.1080/01436597.2020.1788935.

Said, Edward W. Orientalism. Vintage Books, 1979.

Shende, Priyal, et al. “The Legacy of James Marion Sims: History Revisited.” Cureus, vol. 16, no. 9, p. e69484. PubMed Central, https://doi.org/10.7759/cureus.69484.

Steinmetz, George. “The Sociology of Empires, Colonies, and Postcolonialism.” Annual Review of Sociology, vol. 40, 2014, pp. 77-103.

Suen, Kelly, et al. “Association between Patient Race/Ethnicity, Health Literacy, Socio-Economic Status, and Incidence of Medication Errors: A Systematic Review.” Journal of Racial and Ethnic Health Disparities, vol. 13, no. 3, June 2026, pp. 2161–72. Springer Link, https://doi.org/10.1007/s40615-025-02407-8.

United States, Executive Office of the President, Office of Management and Budget. “Revisions to OMB’s Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity.” Federal Register, vol. 89, no. 62, 29 Mar. 2024, pp. 22182–96.

United States, Food and Drug Administration. Collection of Race and Ethnicity Data in Clinical Trials: Guidance for Industry and Food and Drug Administration Staff. Oct. 2016, www.fda.gov/media/75453/download.

Zucker, Irving, and Brian J. Prendergast. “Sex Differences in Pharmacokinetics Predict Adverse Drug Reactions in Women.” Biology of Sex Differences, vol. 11, art. 32, June 2020. PubMed Central, https://doi.org/10.1186/s13293-020-00308-5.

Zucker, Irving, et al. “Pervasive Neglect of Sex Differences in Biomedical Research.” Cold Spring Harbor Perspectives in Biology, vol. 14, no. 4, Apr. 2022, p. a039156. cshperspectives.cshlp.org, https://doi.org/10.1101/cshperspect.a039156.

Aerodynamic Drag Reduction and Thermal Management in Electric Vehicles: Trade-offs for Energy Efficiency and Driving Range By Ada Yilmaz

Abstract

The growing use of electric vehicles (EVs) has made improving driving range and energy efficiency increasingly important, while also maintaining the temperatures needed for safe and reliable operation. This review examines how aerodynamic drag reduction strategies can balance these requirements by considering their effects on aerodynamic drag, energy consumption, driving range and thermal management. The literature shows that reducing aerodynamic drag can lower energy consumption, particularly at higher speeds. Of the strategies reviewed, body-shape modifications produced the largest reductions, with boat-tail configurations achieving reductions of up to 30%. Underbody optimisation, wheel design and active grille shutters also showed improvements in aerodynamic performance. However, direct evidence of improvements in driving range remains limited, as many aerodynamic studies are based on computational fluid dynamics (CFD) simulations and simplified vehicle models. Thermal management also plays an important role in EV efficiency and performance. Battery thermal management improved temperature control and driving range, while motor cooling improved thermal performance and power output. Cabin thermal management strategies also reduced energy consumption and helped maintain driving range. Integrated thermal management systems showed further potential for improving energy efficiency and driving range. However, reducing airflow can make vehicle cooling more challenging, creating a trade-off between aerodynamic performance and thermal requirements. Active grille shutters demonstrate this balance by reducing drag while adjusting airflow to meet cooling demands. Overall, the literature suggests that considering aerodynamic and thermal management requirements together could improve overall EV efficiency more effectively than treating them separately. Greater integration of aerodynamic and thermal systems, supported by vehicle-level simulation and experimental testing, could help identify designs that improve driving range and energy efficiency without compromising thermal performance.

1. Introduction

As road transport transitions towards lower-emission mobility, improving vehicle energy efficiency and driving range has become an important engineering objective (Tran et al.). The transport sector is a major source of greenhouse gas emissions and makes a large contribution to climate change (‘Broken Record Temperatures Hit New Highs, yet World Fails to Cut Emissions (Again)’; US EPA). Electric vehicles have the potential to lower the amount of emissions associated with transport operations when compared to conventional internal combustion engine (ICE) vehicles (Dan et al.). However, the fact that electric vehicles have a more limited driving range than conventional vehicles is still a significant practical issue affecting their adoption. Because of this, aerodynamic and thermal management methods that reduce energy losses and enhance the overall efficiency of the vehicle have become more important (Ravelli and Ravelli).

This review investigates the following research question: How do aerodynamic drag reduction strategies in electric vehicles balance driving range, energy efficiency and thermal management requirements?

Despite significant improvements in both aerodynamic design and thermal management, existing literature tends to treat these as separate engineering problems. It has been shown that aerodynamic measures such as active grille shutters, underbody optimisation and modifications to the vehicle's shape can reduce drag and enhance energy efficiency (Ravelli and Ravelli; de Souza et al.). Also, progress in thermal management systems has led to improved battery performance and lower energy consumption through techniques such as heat pumps and integrated cooling systems (Dan et al.). However, restricting airflow in order to reduce aerodynamic drag may also have an impact on the cooling of the battery, motor and cabin systems (Patil et al.). For example, active grille shutters alter the amount of airflow entering the vehicle, requiring a balance between aerodynamic drag reduction and supplying cooling air to various vehicle components (Patil et al.). It is therefore important to understand the interaction between aerodynamic design and thermal management when designing EVs that can achieve greater driving range without compromising performance, reliability or safety.

This review examines how current aerodynamic drag reduction strategies balance driving range, energy efficiency and thermal management requirements, arguing that while significant progress has been made, critical gaps remain in how these demands are addressed as an integrated system.

The relevant literature was found using academic search platforms such as Google Scholar and various journal databases; the searches employed a combination of terms involving electric vehicles, aerodynamic drag reduction, body-shape optimisation, thermal management and driving range. Initially, review papers were used to identify the major research themes and relevant studies, after which the original research papers were located.

This paper is structured as follows. Section 2 reviews aerodynamic drag reduction strategies including body shape, wheel design, underbody design and grille shutters, focusing on drag reduction, energy efficiency and EV driving range. Section 3 examines battery, motor and cabin thermal management systems, focusing on thermal performance, energy efficiency, motor performance and driving range. Section 4 integrates the main findings, discusses key limitations and research gaps, and presents the conclusions and future research directions.

Figure 1 summarises the conceptual relationship between aerodynamic strategies, thermal-management requirements and vehicle-level outcomes considered in this review.

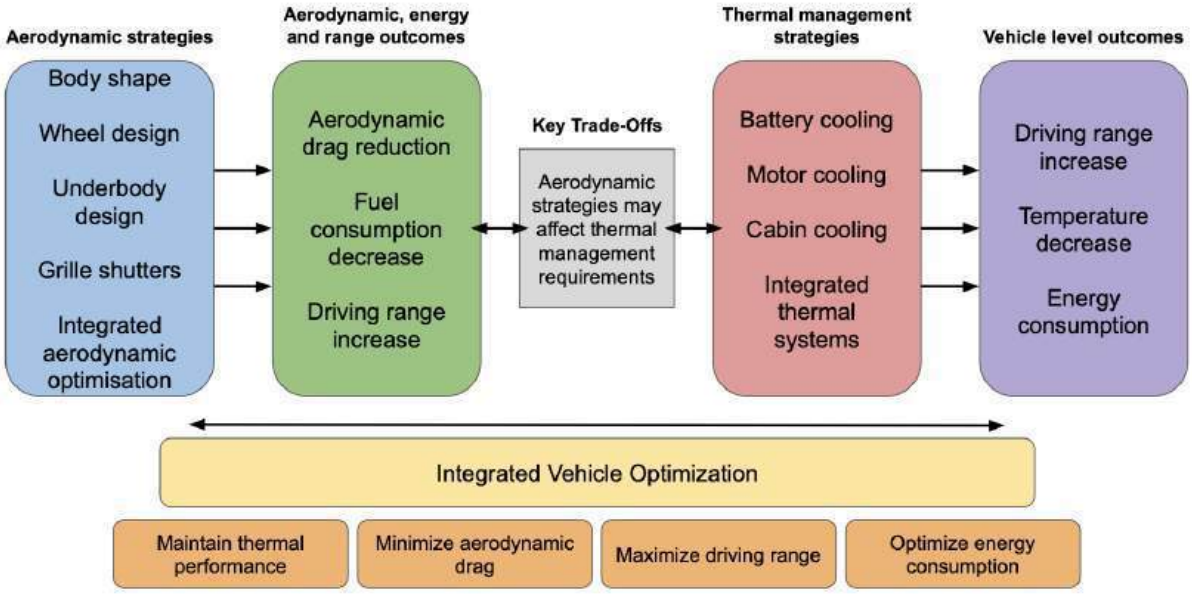


Fig 1: Conceptual framework illustrating the interactions between aerodynamic strategies, thermal management requirements and vehicle level outcomes in electric vehicles.

2. Aerodynamic Design

Aerodynamic drag is an important contributor to energy consumption in battery electric vehicles, particularly during high-speed driving (Seo). The reduction of aerodynamic drag plays a critical role to increase the driving range of electric vehicles (Ren and Fan). Moreover, reducing aerodynamic drag can therefore reduce the energy required to operate an EV. In a battery electric vehicle, Seo found that a 15% reduction in aerodynamic drag reduced energy consumption by 6.0% in urban driving and 11.5% in highway driving (Seo).

The following sections examine the main aerodynamic drag reduction strategies identified in the literature, including body shape optimisation, wheel design, underbody design and grille shutters. For each strategy, findings are reported in terms of drag reduction achieved, energy efficiency improvement and impact on EV driving range. Table 1 provides an overview of the drag reduction strategies, assessing the criteria such as aerodynamic drag reduction, energy efficiency and the impact on the EV driving range.

Table 1. Comparison of aerodynamic drag reduction strategies and their effects on EV performance.

Strategy	**Vehicle model and method**	**Aerodynamic outcome**	**Energy/range outcome**	**Main limitation**	**Reference**
Body shape modification	Ahmed body; CFD	Boat tail 30%; chamfers 5.8%; rounded edges	Not reported	Mostly simplified Ahmed body geometry	(Khalighi et al.; Grandemange et al.; Thacker et al.)

		10% drag reduction			
Wheel design modification	CFD; shape morphing wheel covers	3.6% reduction in C_D	Not reported	CFD based	(W. D. Wang et al.)
Underbody modification	Simplified flat-underbody vehicle; CFD + wind tunnel.	7.05% - 8% drag reduction	Not reported	Simplified model and specific conditions	(Yuan and Wang; Cho et al.)
Active grille shutters	2D/3D vehicle models; CFD	12.23% drag reduction	0.65%-0.89% decrease in fuel consumption	Simplified 2D representation; simulation based	(Patil et al.; Y. Zhang et al.)
Aerodynamic drag reduction	BEV; CFD	15% drag reduction	6.0% lower urban and 11.5% lower highway energy consumption	Simulation based analysis under specified driving conditions	(Seo)
Combined aerodynamic optimisation	CFD; air dam, under-hood cover and rear bumper cover	9.6% reduction in C_D	7km increase in driving range	CFD based	(Gorre et al.)

2.1 Importance of Aerodynamic Drag

Aerodynamic drag is a resistive force which acts against the motion of a vehicle as it travels through the surrounding air (*Aerodynamic Drag - an Overview | ScienceDirect Topics*). At 100 km/h, aerodynamic drag accounts for approximately 75-80% of the total driving resistance (Tran et al.). As vehicle speed increases, aerodynamic drag becomes an increasingly important contributor to vehicle energy consumption (Upadhyaya). In addition, the higher efficiency of electric vehicle powertrains means that aerodynamic drag accounts for a greater proportion of energy losses, making aerodynamic performance a major design priority (Sweeting et al.). Furthermore, even small changes to vehicle geometry can influence the airflow around the vehicle and its aerodynamic performance (Tran et al.). Reducing aerodynamic drag can reduce vehicle energy consumption, with a 20% reduction in aerodynamic drag associated with a 4% decrease in highway fuel consumption (Tran et al.). Similarly, for electric vehicles, reducing aerodynamic drag is particularly important for increasing driving range (Ren and Fan).

2.2 Body Shape

This section examines the effect that vehicle body geometry has on aerodynamic drag by altering the flow of air and the way the wake is formed. Vehicle body shape is the main factor in determining the aerodynamic drag coefficient (Pan). The overall shape of the vehicle body has a greater impact on aerodynamic drag than individual aerodynamic parts such as spoilers and diffusers (Pan). Rear-end geometry is especially important since it affects the structure of the

wake and the aerodynamic drag. Furthermore, vehicles with a fastback body design have narrower wakes and shorter backflows than those with a straight-back design, and straight-back vehicles have drag coefficients that are about 0.03 to 0.05 higher (Pan).

To investigate the influence of body shape on aerodynamic drag, researchers commonly use the Ahmed body. This is a simplified model of a vehicle which reflects the key aerodynamic features of a road vehicle (*Ahmed Body Drag Coefficient*; Aguerre et al.). Much of the literature has been concerned with changing the rear-end shape in order to modify the wake structure and decrease aerodynamic drag. Grandemange et al. reduced the drag by 5.8% by adding chamfers to both the upper and lower edges of the rear surface of a squareback Ahmed body (Grandemange et al.). In a similar way, Thacker et al. obtained a 10% decrease in drag by introducing a rounded edge to a 25° slanted Ahmed body, this being enough to avoid flow separation at the point where the roof meets the rear slant (Thacker et al.). These studies demonstrate that relatively small geometric modifications can lead to a reduction in aerodynamic drag.

Larger rear-end modifications have produced even greater reductions in drag. Khalighi et al. studied a rear cavity and a boat-tail on a squareback Ahmed body and found drag reductions of 18% and 30% respectively (Khalighi et al.). Moreover, Lorite-Díez et al. discovered that curved rear cavities were more effective than straight ones at reducing wake fluctuations and improving base pressure recovery, although the effectiveness of straight cavities diminished under moderate crosswind conditions (Lorite-Díez et al.).

Similar rear-end optimisation strategies have also been studied using the Windsor body, a model used to study fundamental flow physics in road vehicles (Urquhart et al.). Urquhart et al. investigated the effect of a tapered rear cavity on the aerodynamic drag of the Windsor body. This tapered cavity reduced the crossflow in the wake, resulting in a smaller and more balanced wake together with an increase in base pressure (Urquhart et al.). Further investigations reported the effect of rear tapering on the aerodynamic performance of the Windsor model (Howell et al.). In this study the rear taper angle (ranging from 5° to 25°) and the taper length were varied, and the results were compared with those obtained for a baseline squareback configuration. Substantial drag reductions were achieved with moderate taper angles, yet a drag peak was found at a specific taper length (Howell et al.).

These findings are seen in the development of electric vehicles. When the Tesla Model S was being worked on from an aerodynamic point of view, optimisation was carried out first at the level of the general body shape before the various components used in production were improved (Palin et al.). This caused the drag coefficient to decrease by 0.07 between the original styling surface that had been approved and the first wind tunnel model (Palin et al.). The aerodynamic optimisation achieved as a result provided significant benefits to vehicle efficiency and overall range (Palin et al.).

Overall, these studies indicate that rear-end optimisation leads to a reduction in aerodynamic drag, the extent of the drag reduction however depending on the design modifications made.

2.3 Wheel Design

This section examines the way rim geometry, tyre design and wheel cover strategies affect wheel aerodynamic drag, wheel design contributing to the vehicle's aerodynamic drag with about 25% of the total aerodynamic drag of a passenger vehicle arising directly or indirectly from the wheels (Wen et al.).

Rim geometry is therefore an important factor in improving the energy efficiency of vehicles (Brandt et al.). Brandt et al. used a parametric wheel model. Their study showed that the wheel coverage area was the most important parameter, since greater coverage led to a reduction in aerodynamic drag (Brandt et al.). Other important parameters were the outer radial region of the wheel and the depth of the wheel centre (Brandt et al.).

Wheel aerodynamic drag is also affected by tyre geometry. Landstrom et al. found that various combinations of tyres and rims led to variations in drag of 5 to 10 drag counts (Landstrom et al.). Hobeika et al. found that tyre profile and tread pattern affect aerodynamic forces, with main grooves consistently reducing drag and larger tyre profiles producing higher drag values (Hobeika et al.).

There is a trade-off when it comes to optimising wheels between reducing aerodynamic drag and improving brake cooling. Although wheel covers can reduce drag by smoothing the airflow around the wheels, completely enclosed designs may hinder the airflow to the brakes and thus decrease cooling performance (W. D. Wang et al.). Research has been carried out into shape morphing wheel covers for use in electric vehicles in order to deal with this compromise. Wang et al. found that the wheel cover decreased the drag coefficient by 3.6% when fully closed at speeds over 80 km/h, while at the same time increasing the heat dissipation of the brake discs at lower speeds (W. D. Wang et al.).

These studies as a whole indicate that aerodynamic drag can be reduced by optimising the rim shape, the tyre properties, and the wheel covers. It is necessary, though, to balance the aerodynamic benefits against practical considerations such as brake cooling.

2.4 Underbody Design

This section examines how underbody components affect aerodynamic drag. Components such as rear diffusers and smooth panels can improve the airflow beneath the vehicle, helping to reduce aerodynamic drag (Yuan and Wang).

Yuan and Wang examined the effect of underbody structure on aerodynamic drag using numerical simulations and wind tunnel experiments. They discovered that the underbody of the real vehicle caused a number of separation vortices and thus led to an increase in aerodynamic drag of 23.4% when compared with the simplified underbody configuration (Yuan and Wang). Additionally, a flat underbody panel can also improve airflow beneath the vehicle by reducing turbulent separation and pressure drag (Pan). Yuan and Wang found that installing a flat underbody panel, which minimises stagnant and recirculating flow regions, reduced the drag coefficient by approximately 0.012 (Yuan and Wang).

Moreover, to reduce drag, the study investigated wheel spoilers and underbody spoilers designed to control underbody airflow. The optimised wheel spoiler and underbody spoiler reduced the drag coefficient by 4.86% and 7.05%, respectively (Yuan and Wang). Even though the spoilers cause extra drag because of their increased frontal area, they can reduce the total aerodynamic drag by restricting the amount of airflow entering the underbody area and by increasing the speed of the airflow under the vehicle (Yuan and Wang).

Rear diffusers can moreover help to reduce aerodynamic losses by enhancing the pressure recovery in the airflow under the vehicle's rear section. A study reported that when the diffuser shape was optimised with a diffuser angle of 15° and a ground clearance of 50 mm, the drag coefficient was lowered by 0.008 and the lift was also reduced (Pan).

Similarly, Cho et al. examined the effect of various underbody aerodynamic devices, such as an undercover, an under-fin, and a side air dam, on a sedan type vehicle. They found that the undercover reduced the rear wake area by weakening the longitudinal vortex pairs and achieved a drag reduction of approximately 8% (Cho et al.). The under-fin and side air dam only provided additional benefits when used in combination with the undercover (Cho et al.).

Similar underbody optimisation strategies have also been carried out on electric vehicles. A study discovered that combining an air dam, under-hood cover and rear bumper cover reduced the drag coefficient by 9.6% compared with the baseline vehicle (Gorre et al.). The decrease in aerodynamic drag increased the estimated driving range of the electric vehicle by about 7 km (Gorre et al.).

Overall, these studies demonstrate that underbody optimization can reduce aerodynamic drag by controlling airflow, reducing vortex formation and improving pressure recovery beneath the vehicle (Pan; Yuan and Wang; Cho et al.). EV specific studies further show that these aerodynamic improvements can translate directly into increased driving range, although the effectiveness of underbody modifications depends on the geometry of individual components and their interaction with the other aerodynamic features (Yuan and Wang; Cho et al.; Gorre et al.).

2.5 Grille Shutters

This section examines how grille design and active grille shutters balance aerodynamic drag reduction with cooling requirements. Front grille design is likewise affected by styling considerations, these sometimes clashing with aerodynamic optimisation (Pan). Connolly et al. found that prominent upright grille designs raised the drag coefficient by about 0.005 when compared to those which had been aerodynamically optimised, showing the aerodynamic disadvantage of having larger front openings (Connolly et al.).

In order to improve grille design, researchers have examined the relationship between the geometry of the grille opening, the airflow rate and cooling performance. Kim et al. found that an increased grille mass flow rate led to better cooling performance but at the same time reduced aerodynamic performance, showing that it is necessary to balance the demand for airflow with drag reduction (J. M. Kim et al.).

Similarly, active grille shutters (AGS) are devices that alter airflow to minimise aerodynamic drag and have been identified as a method of improving fuel economy (Patil et al.). The system controls airflow through the grille, enabling more airflow to be passed through when the cooling demand from the vehicle's systems is high. Under lower load conditions and moderate environmental conditions, the grille opening can be reduced, although this decreases under-hood airflow (Patil et al.).

The aerodynamic advantages of using active aerodynamic devices have also been shown as a result of numerical simulations. A computational fluid dynamics (CFD) study which looked at the combined use of active grille shutters and air dams found that these measures cut aerodynamic drag by 12.23% when compared to the baseline passenger vehicle (Patil et al.). The study also stated a significant reduction in fuel consumption (Patil et al.). Moreover, Zhang et al. suggested an AGS opening control strategy for a hybrid electric vehicle that took into account the interaction between aerodynamic performance and thermal management. Compared to vehicles without active grille shutters, this strategy lowered the average resistance coefficient and resulted in fuel consumption reductions of 0.65% and 0.89% respectively for initial SOC values of 0.2 and 0.3 (Y. Zhang et al.).

2.6 Key Aerodynamic Findings

The findings from the literature are considered in terms of aerodynamic drag reduction, energy efficiency and EV driving range.

Firstly, the aerodynamic drag reduction strategies reviewed in this section achieved varying levels of drag reduction depending on the design approach and vehicle configuration. Modifications to the body shape resulted in the greatest drag reductions, with boat-tail designs achieving a 30% reduction in drag, as opposed to 5.8% for chamfers and 10% for rounded edges (Grandemange et al.; Thacker et al.; Khalighi et al.). The changes made to the wheels produced smaller reductions, with shape-morphing wheel covers reducing the drag coefficient by 3.6% (W. D. Wang et al.). Drag was reduced by 7.05% using an underbody spoiler and by about 8% using a combination of an underbody cover, an under-fin, and side air-dams (Yuan and Wang; Cho et al.). Active grille shutters combined with air dams reduced aerodynamic drag by 12.23% compared with a baseline passenger vehicle (Patil et al.).

When it comes to energy efficiency, a 20% reduction in aerodynamic drag results in a 4% lower fuel consumption (Tran et al.). According to Zhang et al., the fuel consumption was reduced by 0.65% and 0.89% for a hybrid electric vehicle when the initial state of charge was 0.2 and 0.3 respectively, using an optimised AGS control strategy (Y. Zhang et al.). For the battery electric vehicle studied by Seo, a 15% reduction in aerodynamic drag reduced energy consumption by 6.0% in urban driving and 11.5% in highway driving (Seo).

Finally, the direct effect on EV driving range was demonstrated by an underbody optimisation study, in which combining an air dam, under-hood cover and rear bumper cover reduced the drag coefficient by 9.6% and increased the estimated driving range by approximately 7 km (Gorre et al.).

However, direct measurements of the effect of aerodynamic modifications on EV driving range were limited within the reviewed studies. Most of these studies used CFD simulations and simplified vehicle models such as the Ahmed and Windsor models (Khalighi et al.; Grandemange et al.; Thacker et al.; *Ahmed Body Drag Coefficient*; Aguerre et al.; Urquhart et al.; Howell et al.). This may limit their representation of real world aerodynamic conditions, since factors like crosswinds and variations in yaw angle influence vehicle aerodynamics (Nakashima et al.). Moreover, the literature was not evenly spread out over the various aerodynamic strategies, with more attention being given to body shape and underbody modifications than to wheel design and the other methods.

3. Thermal Management

Thermal management in electric vehicles involves the cooling and heating of batteries, electric motors and power electronics, as well as cabin heating and cooling (Previati et al.; Y. Wang et al.). Battery thermal management is particularly challenging because batteries have a limited operating temperature range and require a low temperature gradient (Previati et al.). Electric motors also require thermal management to maintain their operating temperature, although their high efficiency means that relatively little energy is dissipated as heat (Previati et al.). Cabin heating and cooling can also affect EV driving range because the energy required for both is supplied by the battery (Previati et al.). The sections that follow look at the strategies used for battery, motor and cabin thermal management, with a focus on how these affect energy efficiency, thermal performance and motor performance. While aerodynamic strategies reduce energy consumption by reducing drag, thermal management can also affect EV energy efficiency and driving range, therefore both factors should be considered when evaluating EV efficiency and range.

3.1 Battery Thermal Management

It is necessary to have a battery thermal management system in order to keep EV batteries within their operating temperature range and to reduce the temperature gradients across the battery (Previati et al.). The best operating temperature for EV batteries is about 20 to 40 °C, and temperatures above 50 °C should be avoided since this could cause damage or lead to combustion (Previati et al.). Battery power and energy capacity can be reduced by low temperatures, so it is necessary to heat the batteries or preheat them when the weather is cold (Previati et al.). High charging power can increase battery temperature, and thermal conditioning may be required during charging (Previati et al.).

Battery thermal management systems (BTMS) can be classified as passive, active or hybrid systems according to how heat is removed from the battery (Previati et al.). Passive BTMS regulate battery temperature through natural convection and conductive heat transfer principles. The main advantages of these systems are low complexity and low power consumption, however their limited heat dissipation capacity restricts their application in high power battery systems (Wen et al.).

Active BTMS use external energy sources to control battery temperature, providing greater cooling capacity than passive systems (Wen et al.; Pesaran). The main active BTMS technologies are forced air cooling, liquid cooling and heat-pipe cooling (Wen et al.). Forced air cooling is commonly used in lower power applications due to its simple structure and low cost (Garud et al.). It also has low energy consumption and can be used to provide heated air for preheating the batteries when the weather is cold (Ji and Wang). Additionally, liquid cooling systems, which have become mainstream for high power battery packs, can use either indirect cooling or direct immersion cooling (Wen et al.). The way the cooling system is arranged also has an impact on thermal performance since both the way the cells are arranged and the flow of the coolant affect heat transfer (T. Wang et al.). In a parallel cooling setup, all the cells are exposed to the coolant at the same time, which improves temperature uniformity but at the cost of greater system complexity (Sun and Dixon; K. Chen et al.). Indirect liquid cooling is currently more widely used because of its safety and reliability, while direct immersion cooling has not yet been widely adopted in EVs (Wen et al.). Increasing the complexity of liquid cooling channels can improve cooling performance but also increases pressure loss and pump power requirements whereas simpler channel designs reduce these requirements but may not provide sufficient cooling for fast charging and high power operation (Wen et al.). Finally, heat pipes can improve temperature uniformity and reduce weight, but this effect does not increase in proportion as the number of heat pipes is increased; integrating heat pipes into battery packs is still a difficult task (Wen et al.).

Numerous studies have shown that a single cooling method has certain limitations therefore a hybrid system has been introduced (Liu et al.). Hybrid BTMS combine two or more cooling methods to reduce battery temperature and improve the uniformity of them (Wen et al.). Common hybrid configurations include air cooling combined with liquid cooling, phase change materials or heat pipes (Wen et al.). Wang et al. examined a hybrid system which involved forced air cooling combined with liquid cooling plates and concluded that placing the fans below the cooling plate resulted in advantages regarding thermal performance and cost effectiveness (S. Wang et al.). Li et al. found that when air cooling was combined with indirect liquid cooling the maximum battery temperature remained below 45 °C under conditions of high temperature and high charge-discharge rates, but the internal temperature gradient was more than 5 °C (Li et al.). Hence, while the system managed to keep the maximum temperature low it did not achieve sufficient temperature uniformity (Wen et al.). Although combining cooling methods can improve battery temperature control, some hybrid configurations introduce additional complexity and practical challenges for commercial applications (Garud et al.). For this reason hybrid BTMS have not yet been used in mass-produced electric vehicles and are at present mainly in the stage of theoretical research (Wen et al.).

Moreover, integrated cooling systems combining liquid cooling, phase change materials (PCM) and heat pipes have been developed for extreme operating conditions (Wen et al.). Not only do these systems manage to lower the battery temperature but they also lead to a reduction in energy consumption (Wen et al.).

BTMS design can affect three key vehicle performance indicators: driving range, charging efficiency and power output (Wen et al.). The temperature at which the battery operates has an effect on both capacity retention and energy conversion efficiency, and it has been reported that optimised BTMS can improve energy efficiency and increase driving range by as much as 25% (Wen et al.). During charging, electrochemical reactions produce heat, which raises the thermal load on the battery, and poor thermal management can reduce the amount of charge that is accepted and result in thermal throttling (Bandhauer et al.). Battery temperature also affects discharge performance, particularly under high-current conditions, which can affect power output (Wen et al.). Therefore, BTMS must ensure thermal stability in order to maintain vehicle performance by avoiding power derating caused by excessive heat accumulation (Garud et al.).

BTMS optimization also needs to consider interactions between vehicle subsystems (Previati et al.). Optimising individual thermal management components can provide a partial view of system performance because interactions between subsystems may be neglected (Previati et al.). System level optimization can therefore account for the behaviour of multiple subsystems, although due to the complexity of these systems, special strategies have to be implemented for the optimization process (Previati et al.).

3.2 Motor Thermal Management

Motor thermal management plays a vital role in the performances of electric motors (Bilgin and Emadi). The motor is a key contributor to EV energy output and produces considerable heat during operation (Dan et al.). Motors must exhibit higher speed, power density and torque density to meet the increasing demand for electric vehicles (Dan et al.; Ušakovs et al.). These requirements can cause short circuits, magnet demagnetisation and other thermal issues (X. Zhang et al.). Insufficient cooling can consequently reduce motor power density, durability and EV driving range on a single charge; therefore an effective cooling method must be applied to electric motors (Park and Kim). The traditional motor thermal management design is often based on the motor operational conditions and large safety factors, which can result in oversized and inefficient systems (Y. Yang et al.). Specifically, the two commonly used motor thermal management methods are air and liquid cooling (Dan et al.). Various designs have been investigated for both cooling methods to meet the thermal management requirements of electric motors (Dan et al.).

The air cooling method, which is commonly used for small to medium sized electric motors, removes heat by directing air flowing into the electric motor (Dan et al.; Chang et al.). Air cooling can be obtained in configurations of ambient or conditioned air and forced or natural convection (Previati et al.). In natural convection, ribs or fins are commonly used for cooling outside the motor, with their layouts optimised to increase the convection surface (Dan et al.; Previati et al.). Forced air convection provides greater heat transfer effectiveness but needs fans to circulate air through the motor (Dan et al.; Previati et al.). Furthermore, it is essential to optimise heat transfer in order to achieve good cooling performance within the airflow path and

in the air gap (Dan et al.). Peng and Lai found that the fin pitch ratio had a significant effect on winding temperature, and their optimised fin and fan-cover arrangement lowered the average rise in winding temperature by 4.8 °C (Peng and Lai). Kim et al. investigated air-gap fans in a large capacity, high speed induction motor. Their findings indicated that the fans increased the airflow through the air gap by eliminating stagnant flow and raised the heat transfer coefficients at both the winding surface and at the air gap by as much as 31% and 90% respectively (C. Kim et al.). The overall winding cooling performance was improved by an average of 55% when fans were used on both the front and rear sides of the motor, whereas the rear-side fan was more effective when used by itself (C. Kim et al.). It is also possible to introduce air via the rotor in order to improve the motor's thermal efficiency, especially during continuous high-power operations such as those involved in highway driving (Jaeger et al.). At high speeds, the rotor becomes increasingly important for thermal management because of higher rotor iron losses and the insulating effect of the air gap (Jaeger et al.). Overall, natural convection provides lower heat transfer capability than forced convection, with typical heat transfer coefficients of 5-10 $W/m^2 \cdot K$ compared with 10-300 $W/m^2 \cdot K$ for forced convection (Y. Yang et al.). Forced convection thus provides greater cooling capability, although it requires additional equipment such as fans (Dan et al.; Y. Yang et al.).

On the other hand, liquid cooling is currently the main thermal management system in electric motors because of its good heat transfer properties (Dan et al.). Liquid cooling can be divided into indirect and direct liquid cooling (Dan et al.; Previati et al.). In the case of indirect cooling, the coolant does not make direct contact with the windings and instead flows through jacket-like channels in the motor casing, with water being a commonly used coolant (Dan et al.). Various channel shapes, such as spiral, U-shaped and axial ones, have been studied using CFD (Satrústegui et al.). Wu et al. designed a smoothly spirally twisted cooling channel which resulted in a reduction of the average temperature of up to 10% (Wu et al.). Ethylene glycol solutions can also be used as a coolant as it extends the temperature range of water cooling, while oil provides another alternative that is compatible with motor lubrication systems (Dan et al.). Liquid cooling may also be applied directly to the windings, with cooling channels placed inside the motor slots (Dan et al.). Moreover, Rhebergen et al. found that winding temperatures could be reduced by 65 to 90 °C when serpentine cooling channels were positioned above the windings (Rhebergen et al.). Chen et al. investigated hollow conductors that include an internal coolant path and achieved a winding temperature reduction of up to 33 °C as compared to indirect cooling (Dan et al.; X. Chen et al.). Another approach is to directly cool the motor end windings using oil jet or oil spray cooling (Dan et al.). Davin et al. investigated the effect of the oil flow rate, rotation speed and oil temperature. The study found that oil flow rate had a greater influence on cooling performance, while rotor speed mainly affected local temperature (Davin et al.). Similarly, Wang et al. showed that oil spray cooling could keep the winding temperatures within safe limits even at higher heat loads and produced a more uniform temperature distribution at higher flow rates and spray temperatures (X. Wang et al.). Moreover, the recovery of waste heat from the motor can also help to increase the driving range of an EV when the

recovery is incorporated into the overall thermal management system. He et al. reported an approximately 30 °C reduction in motor and motor controller temperature, while driving range increased by 8.58 km at -5 °C, 6.67 km at -15 °C and 5.05 km at -25 °C (He et al.). In addition, Park et al. found that a direct slot-cooled motor achieved a 30 °C lower temperature than a water jacket cooled motor under highway driving conditions, while also achieving 99% higher maximum torque and 117% higher maximum power (Park et al.). Furthermore, the DSC motor showed 80% greater thermal durability in terms of current density when the coolant pumping power was the same (Park et al.). In summary, the results obtained by Park et al. show the potential of direct liquid cooling to offer better thermal management and motor performance than indirect water-jacket cooling under high power operating conditions (Park et al.).

3.3 Cabin Thermal Management

The increasing adoption of electrified vehicles has made cabin thermal management increasingly important for maintaining operating range, protecting components from ageing and ensuring passenger comfort (Lajunen et al., "Review of Cabin Thermal Management"). Cabin thermal management consists of a heating loop and an air conditioning loop. The heating loop is supplied by an electric heat source, while the air conditioning loop uses a condenser, compressor and evaporator to regulate the cabin climate (Lajunen et al., "Recent Developments"). In EVs, electric water pumps are required to circulate coolant because there is no engine and an electric heater is required for cabin heating (Lajunen et al., "Recent Developments"). The compressor used in the air conditioning system also needs an electric motor (Lajunen et al., "Recent Developments"). Unlike in conventional vehicles, electric vehicles produce very little waste heat for use in heating the cabin (D. Yang et al.; Lajunen et al., "Recent Developments"). As a result, using the heating system can reduce the maximum driving range of the vehicle, which makes it a key challenge to develop an energy-efficient method of cabin heating (D. Yang et al.; Lajunen et al., "Recent Developments").

Since the waste heat produced by the powertrains of electric vehicles is not enough to effectively heat the cabin, it is generally necessary to use electric heaters (Previati et al.). For cabin heating in electric vehicles, Positive Temperature Coefficient (PTC) heaters with a capacity of between 5 and 10 kW are typically used (Shin et al.). Cabin heating systems can be divided into those which use a PTC heater by itself, those that combine a PTC heater with a heat pump, and those that incorporate both a heat pump and waste-heat recovery (Dan et al.). However, cabin thermal management requires a considerable amount of energy, and up to 50% reduction in vehicle range has been reported in cold weather depending on the difference between the outside temperature and the desired cabin temperature (Previati et al.; Ravelli and Ravelli). The requirements for cabin thermal management differ between summer and winter, with temperature, humidity and airflow all playing a role in thermal comfort (Dan et al.). In order to lower heating power consumption, heat pumps can be used when the temperature is above -10 °C, although some systems have been reported to function down to -30 °C (Previati et al.). A heat-pump system that also includes an electric heater increased the driving range by 6 to 22% in

mild ambient temperatures ranging from -5 °C to 10 °C, the exact figure depending on the driving cycle (Lajunen). Waste heat recovery from the powertrain components further cut vehicle energy consumption by 2 to 4% (Lajunen). Guo et al. suggested an integrated thermal-management system which recovers waste heat from the range extender and the electric drive system to heat both the battery and the cabin (Guo et al.). In range extended mode, the integrated system reduced battery heating time and fuel consumption by more than 10.2% compared with the independent system (Guo et al.). In pure electric mode, the proposed control strategy reduced heating energy consumption by 20.95%, while total energy consumption for both propulsion and thermal management was reduced by 2.84% (Guo et al.).

3.4 Key Thermal-Management Findings

The findings from the literature are considered in terms of thermal performance, energy efficiency and motor performance.

The various thermal management strategies resulted in different degrees of temperature control according to the component and the method of cooling. Direct liquid cooling usually gave better thermal control than air cooling, since direct slot cooling not only lowered the motor temperature but also improved motor output (Y. Yang et al.; Rhebergen et al.; Park et al.). Motor waste heat recovery also reduced motor and motor controller temperatures while increasing driving range under low temperature conditions (He et al.).

The evaluation of energy efficiency and driving range was more directly carried out in studies concerning cabin thermal management. Cabin heating can have a significant impact on the range of electric vehicles in cold weather, whereas heat pump systems were found to improve range under mild temperature conditions (Ravelli and Ravelli; Previati et al.; Lajunen). Battery thermal management has also been connected with driving range, since an optimised BTMS has been reported to boost range by as much as 25% (Wen et al.). However, the battery and motor studies generally focus on component level outcomes such as temperature, heat transfer, torque and power, making direct comparison between the three thermal management areas difficult.

Finally, improved thermal management can also enhance motor performance. Direct slot cooling increased the maximum torque, the maximum power and thermal durability when compared with cooling using a water jacket, thus showing the connection between thermal control and motor performance (Park et al.).

In this section, the three components have been considered separately. However, integrating the cabin air conditioning, battery and motor thermal management systems can further improve overall thermal management effectiveness (Dan et al.). For example, Tian et al. proposed an integrated system combining the motor, cabin and battery thermal management systems and reported a maximum 13.8% improvement in coefficient of performance (COP) and a reduction in annual cost of at least 54.8% compared with using heaters alone (Tian et al.). Furthermore, the driving range increased by 33.64% (Tian et al.). Also, Guo et al. stated that an integrated thermal management system made up of the cabin and battery systems reduced heating energy consumption by 20.95% and lowered the total energy consumption for both

propulsion and thermal management by 2.84% in pure electric mode (Guo et al.). Moreover, Guo et al. reported more than 10.2% lower fuel consumption in range extended mode and shorter battery heating time compared with the independent system (Guo et al.).

The connection between thermal management and the aerodynamic findings also shows that aerodynamic studies can directly link drag reduction to energy consumption and driving range, whereas thermal management studies rely on a wider range of component-level and system-level metrics. Future research must therefore make use of integrated vehicle level testing in order to assess aerodynamic and thermal strategies by means of the same kinds of measures, such as energy consumption and driving range. This would allow the total contribution of both strategies to the efficiency and range of electric vehicles to be compared more directly.

Table 2 gives a summary of the main quantitative results reported in the literature regarding battery, motor and cabin thermal management strategies, together with their impact on thermal performance, energy efficiency, motor performance and the EV driving range.

Table 2. Comparison of thermal management strategies and their quantitative effects on EV performance.

Thermal management strategy	**Component**	**Metric**	**Key quantitative findings**	**Vehicle level implementation**	**Reference**
Optimised BTMS	Battery	Driving range	Up to 25% increase	Improved EV range	(Wen et al.)
Natural and forced convection	Motor	Heat-transfer coefficient	Natural 5-10 W/m²·K; forced 10-300 W/m²·K	Improved air cooling	(Y. Yang et al.)
Direct slot cooling	Motor	Temperature, torque, power, durability	Temperature 30 °C decrease; maximum torque 99% increase; maximum power 117% increase; thermal durability 80% increase	Higher motor performance under high thermal load	(Park et al.)
Fin + fan-cover configuration	Motor	Average winding temperature	4.8 °C decrease	Improved motor cooling	(Peng and Lai)
Air gap fans	Motor	Heat-transfer coefficient, cooling performance	Winding up to 31% increase; air gap up to 90% increase; overall cooling 55% increase	Improved air cooling	(C. Kim et al.)
Serpentine direct cooling channels	Motor winding	Winding temperature	65-90 °C decrease	Improved winding cooling	(Rhebergen et al.)
Motor waste heat recovery	Motor, motor controller	Driving range	Up to 8.58 km increase	Improved motor cooling and EV driving range	(He et al.)

PTC heating	Cabin	Heater capacity, range	5-10 kW heater capacity; cabin HVAC associated with up to 50% range reduction in cold weather	High heating demand in cold conditions	(Shin et al.; Previati et al.; Ravelli and Ravelli)
Heat pump + electric heater, waste heat recovery	Cabin	Driving range, energy consumption	Driving range 6-22% increase; waste heat recovery energy consumption 2-4% decrease	Improved cabin heating efficiency and EV range	(Lajunen)
Integrated battery + cabin system	Battery, cabin	Energy consumption, fuel consumption	Heating energy 20.95% decrease; total energy 2.84% decrease; fuel consumption >10.2% decrease	Benefits of integrated thermal management	(Guo et al.)
Integrated battery + motor + cabin system	Battery, motor, cabin	Coefficient of performance, cost, driving range	COP 13.8% increase; annual cost 54.8% decrease; driving range 33.64% increase	Benefits of integrated thermal management	(Tian et al.)

4. Conclusion

This review demonstrates that improving the energy efficiency and driving range of electric vehicles requires a balance between aerodynamic drag reduction and thermal management strategies. Both affect the energy required for EV operation, although the literature measures their effects using different performance metrics.

The aerodynamic strategies reviewed produced varying levels of drag reduction depending on the design approach and vehicle configuration. The largest reported reductions among the selected studies were body shape modifications, while underbody, wheel and active grille shutter strategies also improved aerodynamic performance. However, direct evidence of driving range improvements remained limited, as many studies relied on CFD simulations and simplified vehicle models (Khalighi et al.; *Ahmed Body Drag Coefficient*; Urquhart et al.).

Thermal management showed similar differences between battery, motor and cabin applications. Battery thermal management strategies improved temperature control, energy efficiency and driving range (Feng et al.). Motor thermal management strategies improved heat transfer, temperature control and motor performance, while motor waste heat recovery also showed potential to improve driving range when integrated into the wider thermal management system (Y. Yang et al.; He et al.; Park et al.). Cabin thermal management reduced the energy required for heating and improved driving range, particularly through heat pumps and waste heat recovery (Lajunen; Guo et al.). Integrated thermal management systems combining the battery, motor and cabin provided further improvements in energy efficiency, thermal performance and driving range (Dan et al.; Guo et al.; Tian et al.). The different metrics used across these studies make direct comparison between battery, motor and cabin thermal management difficult.

The optimum EV configuration is not necessarily the configuration with the minimum aerodynamic drag. Vehicle level optimization must also consider thermal management requirements, energy consumption and driving range, as reducing aerodynamic drag can alter airflow available for cooling the battery, motor and other vehicle systems. This is particularly relevant for active grille shutters, where reducing the grille opening can improve aerodynamic performance but also restrict cooling airflow. Similarly, thermal management systems can require additional energy for cooling and heating, meaning that an aerodynamic improvement does not necessarily result in an equivalent improvement in overall vehicle efficiency.

Future research should therefore evaluate aerodynamic and thermal management strategies together using vehicle models and experimental testing. Combining CFD, thermal modelling and real vehicle testing could provide a more consistent basis for comparing these strategies and identifying the most effective approaches for improving overall EV efficiency.

Works Cited

Aerodynamic Drag - an Overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/engineering/aerodynamic-drag? Accessed 4 Sept. 2026.

Aguerre, Horacio J., et al. 'Aerodynamic Study of a Moving Ahmed Body by Numerical Simulation'. Journal of Wind Engineering and Industrial Aerodynamics, vol. 245, Feb. 2024, p. 105635. ScienceDirect, https://doi.org/10.1016/j.jweia.2023.105635.

Ahmed Body Drag Coefficient. https://ansyshelp.ansys.com/public//Views/Secured/corp/v252/en/discovery_tech_show/Discovery/tech_show/ahmed_body_drag/c_ahmed_body_intro.html? Accessed 4 Aug. 2026.

Bandhauer, Todd M., et al. 'A Critical Review of Thermal Issues in Lithium-Ion Batteries'. Journal of The Electrochemical Society, vol. 158, no. 3, Jan. 2011, p. R1. iopscience.iop.org, https://doi.org/10.1149/1.3515880.

Bilgin, Berker, and Ali Emadi. 'Electric Motors in Electrified Transportation: A Step toward Achieving a Sustainable and Highly Efficient Transportation System'. IEEE Power Electronics Magazine, vol. 1, no. 2, June 2014, pp. 10–17. IEEE Xplore, https://doi.org/10.1109/MPEL.2014.2312275.

Brandt, Adam, et al. 'The Effects of Wheel Design on the Aerodynamic Drag of Passenger Vehicles'. SAE International Journal of Advances and Current Practices in Mobility, vol. 1, no. 3, nos 2019-01–0662, Apr. 2019, pp. 1279–99. legacy.sae.org, https://doi.org/10.4271/2019-01-0662.

'Broken Record Temperatures Hit New Highs, yet World Fails to Cut Emissions (Again)'. DOI.org (Crossref), https://doi.org/10.1163/9789004322714_cclc_2023-0252-1155. Accessed 31 July 2026.

Chang, Min, et al. 'Comprehensive Efficiency Analysis of Air-Cooled vs Water-Cooled Electric Motor for Unmanned Aerial Vehicle'. Applied Thermal Engineering, vol. 225, May 2023, p. 120226. ScienceDirect, https://doi.org/10.1016/j.applthermaleng.2023.120226.

Chen, Kai, et al. 'Structure Optimization of Parallel Air-Cooled Battery Thermal Management System with U-Type Flow for Cooling Efficiency Improvement'. Energy, vol. 145, Feb. 2018, pp. 603–13. ScienceDirect, https://doi.org/10.1016/j.energy.2017.12.110.

Chen, Xiao, et al. 'Thermal Modeling of Hollow Conductors for Direct Cooling of Electrical Machines'. IEEE Transactions on Industrial Electronics, vol. 67, no. 2, Feb. 2020, pp. 895–905. IEEE Xplore, https://doi.org/10.1109/TIE.2019.2899542.

Cho, Junho, et al. 'Comparative Investigation on the Aerodynamic Effects of Combined Use of Underbody Drag Reduction Devices Applied to Real Sedan'. International Journal of Automotive Technology, vol. 18, no. 6, Dec. 2017, pp. 959–71. Springer Link, https://doi.org/10.1007/s12239-017-0094-5.

Connolly, Michael Gerard, et al. ‘Drag Reduction Technology and Devices for Road Vehicles - A Comprehensive Review’. Heliyon, vol. 10, no. 13, July 2024, p. e33757. PubMed Central, https://doi.org/10.1016/j.heliyon.2024.e33757.

Dan, Dan, et al. Review of Thermal Management Technology for Electric Vehicles. https://www.mdpi.com/1996-1073/16/12/4693. Accessed 29 July 2026.

Davin, Tanguy, et al. ‘Experimental Study of Oil Cooling Systems for Electric Motors’. Applied Thermal Engineering, vol. 75, Jan. 2015, pp. 1–13. ScienceDirect, https://doi.org/10.1016/j.applthermaleng.2014.10.060.

de Souza, Fenella, et al. ‘Investigation of Drag Reduction Technologies for Light-Duty Vehicles Using Surface, Wake and Underbody Pressure Measurements to Complement Aerodynamic Drag Measurements’. SAE International Journal of Advances and Current Practices in Mobility, vol. 1, no. 3, 2019, pp. 1233–50. PubMed Central, https://doi.org/10.4271/2019-01-0644.

Feng, Xuning, et al. ‘Thermal Runaway Mechanism of Lithium Ion Battery for Electric Vehicles: A Review’. Energy Storage Materials, vol. 10, Jan. 2018, pp. 246–67. ScienceDirect, https://doi.org/10.1016/j.ensm.2017.05.013.

Garud, Kunal Sandip, et al. ‘A Review of Advanced Cooling Strategies for Battery Thermal Management Systems in Electric Vehicles’. Symmetry, vol. 15, no. 7, July 2023, p. 1322. www.mdpi.com, https://doi.org/10.3390/sym15071322.

Gorre, Parandhamaiah, et al. Underbody Drag Reduction Study for Electric Car Using CFD Simulations. SAE International, 14 Jan. 2015, Symposium on International Automotive Technology 2015. saemobilus.sae.org, https://doi.org/10.4271/2015-26-0211.

Grandemange, M., et al. ‘A Study of Wake Effects on the Drag of Ahmed′s Squareback Model at the Industrial Scale’. Journal of Wind Engineering and Industrial Aerodynamics, vol. 145, Oct. 2015, pp. 282–91. ScienceDirect, https://doi.org/10.1016/j.jweia.2015.03.004.

Guo, Rong, et al. ‘An Integrated Thermal Management Strategy for Cabin and Battery Heating in Range-Extended Electric Vehicles under Low-Temperature Conditions’. Applied Thermal Engineering, vol. 228, June 2023, p. 120502. ScienceDirect, https://doi.org/10.1016/j.applthermaleng.2023.120502.

He, Liange, et al. ‘Performance Research of Integrated Thermal Management System for Battery Electric Vehicles with Motor Waste Heat Recovery’. Journal of Energy Storage, vol. 84, Apr. 2024, p. 110893. ScienceDirect, https://doi.org/10.1016/j.est.2024.110893.

Hobeika, Teddy, et al. ‘Investigation of the Influence of Tyre Geometry on the Aerodynamics of Passenger Cars’. SAE International Journal of Passenger Cars - Mechanical Systems, vol. 6, no. 1, nos 2013-01–0955, Apr. 2013, pp. 316–25. legacy.sae.org, https://doi.org/10.4271/2013-01-0955.

Howell, Jeff, et al. ‘Aerodynamic Drag Reduction on a Simple Car-Like Shape with Rear Upper Body Taper’. SAE International Journal of Passenger Cars - Mechanical Systems, vol. 6, no. 1, nos 2013-01–0462, Apr. 2013, pp. 52–60. legacy.sae.org, https://doi.org/10.4271/2013-01-0462.

Jaeger, Markus, et al. ‘Thermal Analysis of an Electrical Traction Motor with an Air Cooled Rotor’. 2018 IEEE Transportation Electrification Conference and Expo (ITEC), 2018, pp. 467–70. IEEE Xplore, https://doi.org/10.1109/ITEC.2018.8450239.

Ji, Yan, and Chao Yang Wang. ‘Heating Strategies for Li-Ion Batteries Operated from Subzero Temperatures’. Electrochimica Acta, vol. 107, Sept. 2013, pp. 664–74. ScienceDirect, https://doi.org/10.1016/j.electacta.2013.03.147.

Khalighi, Bahram, et al. ‘Unsteady Aerodynamic Flow Investigation Around a Simplified Square-Back Road Vehicle With Drag Reduction Devices’. Journal of Fluids Engineering, vol. 134, no. 061101, May 2012. Silverchair, https://doi.org/10.1115/1.4006643.

Kim, Chiwon, et al. ‘Effect of Air-Gap Fans on Cooling of Windings in a Large-Capacity, High-Speed Induction Motor’. Applied Thermal Engineering, vol. 100, May 2016, pp. 658–67. ScienceDirect, https://doi.org/10.1016/j.applthermaleng.2016.02.077.

Kim, J. M., et al. ‘Grille Design for Passenger Car to Improve Aerodynamic and Cooling Performance Using CFD Technique’. International Journal of Automotive Technology, vol. 17, no. 6, Dec. 2016, pp. 967–76. DOI.org (Crossref), https://doi.org/10.1007/s12239-016-0094-x.

Lajunen, Antti. Energy Efficiency and Performance of Cabin Thermal Management in Electric Vehicles. SAE International, 28 Mar. 2017, WCX™ 17: SAE World Congress Experience. saemobilus.sae.org, https://doi.org/10.4271/2017-01-0192.

———. ‘Recent Developments in Thermal Management of Electrified Powertrains’. IEEE Transactions on Vehicular Technology, vol. 67, no. 12, Dec. 2018, pp. 11486–99. IEEE Xplore, https://doi.org/10.1109/TVT.2018.2876315.

———. ‘Review of Cabin Thermal Management for Electrified Passenger Vehicles’. IEEE Transactions on Vehicular Technology, vol. 69, no. 6, June 2020, pp. 6025–40. IEEE Xplore, https://doi.org/10.1109/TVT.2020.2988468.

Landstrom, Christoffer, et al. ‘Aerodynamic Effects of Different Tire Models on a Sedan Type Passenger Car’. SAE International Journal of Passenger Cars - Mechanical Systems, vol. 5, no. 1, nos 2012-01–0169, Apr. 2012, pp. 136–51. legacy.sae.org, https://doi.org/10.4271/2012-01-0169.

Li, Meng, et al. ‘Research on Temperature Control Performance of Battery Thermal Management System Composited with Multi-Channel Parallel Liquid Cooling and Air Cooling’. Ionics, vol. 27, no. 6, June 2021, pp. 2685–95. Springer Link, https://doi.org/10.1007/s11581-021-04033-w.

Liu, Jiahao, et al. ‘Recent Progress and Prospects in Liquid Cooling Thermal Management System for Lithium-Ion Batteries’. Batteries, vol. 9, no. 8, Aug. 2023, p. 400. www.mdpi.com, https://doi.org/10.3390/batteries9080400.

Lorite-Díez, M., et al. ‘Drag Reduction on a Three-Dimensional Blunt Body with Different Rear Cavities under Cross-Wind Conditions’. Journal of Wind Engineering and Industrial

Aerodynamics, vol. 200, May 2020, p. 104145. ScienceDirect, https://doi.org/10.1016/j.jweia.2020.104145.

Nakashima, Takuji, et al. ‘Active Aerodynamics Control of Simplified Vehicle Body in a Crosswind Condition’. The Journal of Engineering, vol. 2020, no. 14, 2020, pp. 1005–11. Wiley Online Library, https://doi.org/10.1049/joe.2020.0062.

Palin, Robert, et al. ‘(PDF) The Aerodynamic Development of the Tesla Model S - Part 1: Overview’. ResearchGate. www.researchgate.net, https://doi.org/10.4271/2012-01-0177. Accessed 12 Aug. 2026.

Pan, Cheng. ‘(PDF) Aerodynamic Considerations and Optimization in Vehicle Body Design’. ResearchGate, Apr. 2026. www.researchgate.net, https://www.researchgate.net/publication/399044538_Aerodynamic_Considerations_and_Optimization_in_Vehicle_Body_Design.

Park, Joonbum, et al. ‘Cooling and Dynamic Performance of Electric Vehicle Traction Motor Adopting Direct Slot Cooling Method’. International Communications in Heat and Mass Transfer, vol. 147, Oct. 2023, p. 106970. ScienceDirect, https://doi.org/10.1016/j.icheatmasstransfer.2023.106970.

Park, Myeong Hyeon, and Sung Chul Kim. ‘Development and Validation of Lumped Parameter Thermal Network Model on Rotational Oil Spray Cooled Motor for Electric Vehicles’. Applied Thermal Engineering, vol. 225, May 2023, p. 120176. ScienceDirect, https://doi.org/10.1016/j.applthermaleng.2023.120176.

Patil, C. Amogh, et al. ‘Characteristics Simulation on Opening of Active Grille Shutters of Automobile Using Computational Fluid Dynamics (CFD)’. [Medan, Indonesia], 2022, p. 020031. DOI.org (Crossref), https://doi.org/10.1063/5.0116749.

Peng, Huan-Sen, and Feng-Hsiang Lai. ‘Investigation of Parameters Affecting Heat Transfer and Fluid Flow of a TEFC Electric Motor by Using Taguchi Method’. IOP Conference Series: Materials Science and Engineering, vol. 491, no. 1, Mar. 2019, p. 012021. Institute of Physics, https://doi.org/10.1088/1757-899X/491/1/012021.

Pesaran, Ahmad. ‘Battery Thermal Management in EVs and HEVs: Issues and Solutions’. Battery Man, vol. 43, Jan. 2001.

Previati, Giorgio, et al. ‘Thermal Management of Electrified Vehicles—A Review’. Energies, vol. 15, no. 4, Jan. 2022, p. 1326. www.mdpi.com, https://doi.org/10.3390/en15041326.

Ravelli, Umberto, and Silvia Ravelli. An Overview of Drag Reduction Methods in Road Cars - ProQuest. https://www.proquest.com/openview/173f6bf08cfaf80e97c0aa9a6db989d1/1?pq-origsite=gscholar&cbl=2048739. Accessed 29 July 2026.

Ren, Chao, and Mengyu Fan. Experimental and Numerical Investigation on the Influence of Different Detailed and Deformed Tires on the Aerodynamics of Electric Vehicles. 9 Jan. 2024, https://doi.org/10.1177/09544070231219293.

Rhebergen, Cody, et al. ‘Enhancement of Electric Motor Thermal Management through Axial Cooling Methods: A Materials Approach’. 2015 IEEE Energy Conversion Congress and

Exposition (ECCE), 2015, pp. 5682–88. IEEE Xplore, https://doi.org/10.1109/ECCE.2015.7310458.

Satrústegui, Marco, et al. ‘Design Criteria for Water Cooled Systems of Induction Machines’. Applied Thermal Engineering, vol. 114, Mar. 2017, pp. 1018–28. ScienceDirect, https://doi.org/10.1016/j.applthermaleng.2016.12.031.

Seo, Jigu. ‘Component-Level Energy Consumption and Range Analysis of Battery Electric Vehicles under Urban and Highway Driving Conditions’. Applied Thermal Engineering, vol. 274, Sept. 2025, p. 126797. ScienceDirect, https://doi.org/10.1016/j.applthermaleng.2025.126797.

Shin, Yoon Hyuk, et al. ‘Performance Characteristics of a Modularized and Integrated PTC Heating System for an Electric Vehicle’. Energies, vol. 9, no. 1, Jan. 2016, p. 18. www.mdpi.com, https://doi.org/10.3390/en9010018.

Sun, Hongguang, and Regan Dixon. ‘Development of Cooling Strategy for an Air Cooled Lithium-Ion Battery Pack’. Journal of Power Sources, vol. 272, Dec. 2014, pp. 404–14. ScienceDirect, https://doi.org/10.1016/j.jpowsour.2014.08.107.

Sweeting, W. J., et al. ‘Factors Affecting Electric Vehicle Energy Consumption’. International Journal of Sustainable Engineering, vol. 4, no. 3, Sept. 2011, pp. 192–201. Taylor and Francis+NEJM, https://doi.org/10.1080/19397038.2011.592956.

Thacker, A., et al. ‘Effects of Suppressing the 3D Separation on the Rear Slant on the Flow Structures around an Ahmed Body’. Journal of Wind Engineering and Industrial Aerodynamics, vols 107–108, Aug. 2012, pp. 237–43. ScienceDirect, https://doi.org/10.1016/j.jweia.2012.04.022.

Tian, Zhen, et al. ‘Performance Evaluation of an Electric Vehicle Thermal Management System with Waste Heat Recovery’. Applied Thermal Engineering, vol. 169, Mar. 2020, p. 114976. ScienceDirect, https://doi.org/10.1016/j.applthermaleng.2020.114976.

Tran, Jonathan, et al. ‘Aerodynamics-Guided Machine Learning for Design Optimization of Electric Vehicles’. Communications Engineering, vol. 3, Nov. 2024, p. 174. PubMed Central, https://doi.org/10.1038/s44172-024-00322-0.

Upadhyaya, Avaneesh. Effect of Cooling Air Intake and Outlet Positioning on the Aerodynamics of BEVs.

Urquhart, Magnus, et al. ‘Drag Reduction Mechanisms on a Generic Square-Back Vehicle Using an Optimised Yaw-Insensitive Base Cavity’. Experiments in Fluids, vol. 62, no. 12, Nov. 2021, p. 241. Springer Link, https://doi.org/10.1007/s00348-021-03334-0.

US EPA, OAR. Fast Facts on Transportation Greenhouse Gas Emissions. Overviews and Factsheets. 25 Aug. 2015, https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions.

Ušakovs, Igors, et al. ‘Experimental Thermal Characterization of the In-Wheel Electric Motor with Loop Heat Pipe Thermal Management System’. Case Studies in Thermal Engineering, vol. 47, July 2023, p. 103069. ScienceDirect, https://doi.org/10.1016/j.csite.2023.103069.

Wang, Shengnan, et al. ‘A Forced Gas Cooling Circle Packaging with Liquid Cooling Plate for the Thermal Management of Li-Ion Batteries under Space Environment’. Applied Thermal Engineering, vol. 123, Aug. 2017, pp. 929–39. ScienceDirect, https://doi.org/10.1016/j.applthermaleng.2017.05.159.

Wang, Tao, et al. ‘Thermal Investigation of Lithium-Ion Battery Module with Different Cell Arrangement Structures and Forced Air-Cooling Strategies’. Applied Energy, vol. 134, Dec. 2014, pp. 229–38. ScienceDirect, https://doi.org/10.1016/j.apenergy.2014.08.013.

Wang, Wei Dawid, et al. ‘Kirigami-Based Passively Shape-Morphing Wheel Covers with Self-Adaptive Aerodynamics’. Device, vol. 4, no. 3, Mar. 2026, p. 101046. ScienceDirect, https://doi.org/10.1016/j.device.2025.101046.

Wang, Xuehui, et al. ‘Experimental Investigation on End Winding Thermal Management with Oil Spray in Electric Vehicles’. Case Studies in Thermal Engineering, vol. 35, July 2022, p. 102082. ScienceDirect, https://doi.org/10.1016/j.csite.2022.102082.

Wang, Yan, et al. ‘Advances in Integrated Vehicle Thermal Management and Numerical Simulation’. Energies, vol. 10, no. 10, Oct. 2017, p. 1636. www.mdpi.com, https://doi.org/10.3390/en10101636.

Wen, Tianze, et al. ‘Advances and Challenges in the Battery Thermal Management Systems of Electric Vehicles’. Materials, vol. 18, no. 20, Oct. 2025, p. 4718. PubMed Central, https://doi.org/10.3390/ma18204718.

Wu, Pey-Shey, et al. ‘Heat Transfer and Thermal Management of Interior Permanent Magnet Synchronous Electric Motor’. Inventions, vol. 4, no. 4, Dec. 2019, p. 69. www.mdpi.com, https://doi.org/10.3390/inventions4040069.

Yang, Dazhang, et al. ‘Recent Advances on Air Heating System of Cabin for Pure Electric Vehicles: A Review’. Heliyon, vol. 8, no. 10, Oct. 2022. www.cell.com, https://doi.org/10.1016/j.heliyon.2022.e11032.

Yang, Yinye, et al. ‘Thermal Management of Electric Machines’. IET Electrical Systems in Transportation, vol. 7, no. 2, 2017, pp. 104–16. Wiley Online Library, https://doi.org/10.1049/iet-est.2015.0050.

Yuan, Zhiqun, and Yiping Wang. ‘(PDF) Effect of Underbody Structure on Aerodynamic Drag and Optimization’. ResearchGate, July 2026. www.researchgate.net, https://doi.org/10.21595/jme.2017.19210.

Zhang, Xing, et al. ‘A Switched Reluctance Motor Torque Ripple Reduction Strategy With Deadbeat Current Control and Active Thermal Management’. IEEE Transactions on Vehicular Technology, vol. 69, no. 1, Jan. 2020, pp. 317–27. IEEE Xplore, https://doi.org/10.1109/TVT.2019.2955218.

Zhang, Yingchao, et al. ‘Research on Intelligent Control Strategy of Active Grille Shutters of Hybrid Vehicles Based on Low Energy Consumption’. Case Studies in Thermal Engineering, vol. 81, May 2026, p. 108026. ScienceDirect, https://doi.org/10.1016/j.csite.2026.108026.

To What Extent Do Taxes on High Income and High Wealth Individuals Affect Their Financial Behaviour and Location Decisions, and Why do These Responses Vary Across Contexts? By Pranav Krishna Narayanan

Abstract

Taxation is a component of a country's economy, a valuable way to generate funding to put towards public services, infrastructure, and economic development to better the country for its citizens. Taxation systems vary across countries, particularly in their structures and rates, and depend on the wealth or income class individuals belong to. In some countries, high-income individuals face substantially higher tax rates than lower-income groups, while others adopt lower or even no income tax. This is an issue prevalent in many countries just like the United Kingdom, but not as much in other countries, such as the United Arab Emirates, where there is no income tax. These differences have led to debate over whether higher tax rates encourage high-net-worth individuals to relocate or adjust their financial behaviour. This review paper evaluates the extent to which taxes on high-income and high-wealth individuals affect their financial behaviour and location decisions and examines why these responses vary across contexts. This will be done by comparing studies from countries and examining evidence on migration, investment, tax planning and wealth management. The evidence in this paper suggests that taxation influences the behaviour of wealthy individuals, but the extent depends on the type of behavioural response. Taxation appears to have a stronger influence on financial decisions with relatively low adjustment costs, while relocation is more complex and influenced by taxation alongside non-tax factors such as family, career opportunities, lifestyle and environment. This research matters because governments must balance revenue objectives against behavioural responses, avoidance, and mobility.

Introduction

Taxation is one of the primary ways governments raise money to fund public services. Tax revenue supports economic development, finances healthcare, education, infrastructure, public safety, and other essential government services. However, tax systems vary significantly across countries in both their structure and rates, reflecting different economic priorities and policy objectives. This has led to ongoing debate over whether taxation influences the final decisions and location choices of high-net-worth individuals in the world. For the purposes of this paper, ultra-high-net-worth individuals (UHNWIs) are defined as those possessing at least US$30 million in net wealth, while high-net-worth individuals (HNWIs) are defined as those possessing at least US$1 million in net wealth, including their primary residence (The Wealth Report 2023, 2023). High-income individuals are distinguished from these wealth categories because income represents earnings over a period rather than accumulated net wealth. As direct research on UHNWIs is limited, this paper also examines studies involving billionaires, high-income earners and individuals in the top wealth percentiles, while acknowledging that these populations are not directly equivalent.

Despite sharing the common purpose of raising government revenue, taxation systems differ considerably across countries because they reflect different economic priorities and policy objectives. Each country has different goals and objectives they set to meet, which their tax systems reflect. Some countries implement a high-tax, high-benefit tax model with the focus of social safety nets and universal services to all residents regardless of their income range. In contrast, other countries implement lower tax rates which are designed to attract foreign capital, or prioritize individual financial freedom, and resource-rich nations may bypass heavy taxation entirely by leveraging state-owned commodities. These taxation methods are not universally the same but are tailored to meet the specific countries' domestic agendas.

These differences in taxation systems may influence whether ultra-wealthy individuals consider relocating. These people have greater flexibility than the average taxpayer, which allows them to relocate and even invest internationally. This means that they may have greater opportunities to relocate to lower-tax jurisdictions if they desire to, but also may not want to due to other factors such as political stability, business opportunity, and quality of life.

This paper examines the extent to which taxes on high-income and high-wealth individuals affect their financial behaviour and location decisions, and why these responses vary across different contexts. By reviewing existing literature on tax policies, migration and wealth management, this paper evaluates whether taxation is the primary driver of relocation decisions or whether other economic, political and social factors play a greater role. Research shows that taxes can affect behaviour through relocation, income timing or asset decisions. Other research suggests that relocation may be weaker because of personal factors as well as economic costs. Therefore, the response can depend on which tax and which behaviour is being examined. Based on the research examined, this paper argues that taxation influences the behaviour and relocation decisions of wealthy individuals to a certain extent, but it is rarely the only factor. Non-tax factors such as family, social life, career opportunities, lifestyle and the overall environment of a potential location can also strongly influence relocation decisions. The analysis is structured by examining relocation and migration responses, non-tax factors that may constrain relocation, income reporting and tax-planning responses, and investment and asset-realisation decisions. These different behavioural responses are then compared to assess the circumstances in which taxation appears to have the greatest influence.

This is a review paper based on existing literature consisting of sources such as peer-reviewed empirical studies, plus government sources for information on tax rates and national tax policies. This paper uses the source by Henley & Partners only as descriptive migration context, not as causal evidence, as this would depend on multiple other factors rather than just taxation. Direct research on UHNW individuals is limited, which means that evidence concerning billionaires, top-income earners, high-wealth individuals and other closely related studies and evidence about those populations have also been included. This creates a comparability limitation because the studies examine different populations, countries, taxes and behavioural outcomes, meaning that their findings cannot always be directly compared or generalised across contexts.

Literature Review

Tax Systems and Their Differences

Tax systems differ across countries due to various historical, economic, social, and political factors which reflect the country's planned trajectory, goals and priorities for the future (Luzkow, 2018). The main taxes relevant to wealthy people include progressive income tax, where higher incomes face higher marginal rates; wealth tax, which is levied on net wealth above certain thresholds; and inheritance tax, which affects transfer of wealth. For example, Luzkow (2018) argues that certain features of tax systems, including those of countries such as the United Kingdom, can disproportionately benefit the ultra-rich at the expense of lower- and middle-income groups. However, this is a contested interpretation, as the distributional impact of taxation depends on the specific taxes considered, their effective incidence and the wider system of government transfers. Taxation should not be treated as one uniform factor because different forms of taxation can create different incentives and behavioural responses. The studies examined in this paper consider several different forms of taxation, including income, estate, and capital gains tax. Therefore, the type of tax being examined is important when comparing how wealthy individuals respond to different taxation systems.

The tax treatment of wealthy individuals does not just differ within tax systems. They also differ drastically between countries. Even though countries like the UK have personal income taxes, other countries such as the UAE have zero personal income taxes (Taxation | the Official Platform of the UAE Government, 2025). The UK has a 45.0% additional marginal income-tax rate on income within the relevant highest band (Government Digital Service, 2024), which is significantly higher than 0% from the UAE. This difference may create a financial incentive for some wealthy individuals to consider lower-tax jurisdictions. But this is not the only factor individuals consider when it comes to relocating.

In 2025 alone, an estimated 142,000 millionaires were projected to relocate internationally, with about 16,500 expected to leave the UK. At the same time, the UAE was expected to receive around 9,800 millionaires (Henley & Partners, 2025). One relevant policy development was the UK's abolition of the previous non-domiciled tax regime from 6 April 2025, replacing its remittance-based treatment of foreign income and gains with a residence-based system. Under the new regime, qualifying new arrivals may receive relief on foreign income and gains for their first four years of UK tax residence (Reforming the Taxation of non-UK Domiciled Individuals, 2024). These changes may have influenced the financial incentives and relocation considerations of some wealthy individuals. However, the migration projections do not establish taxation as the cause of these movements or how many individuals intended to leave specifically because of the reforms. This raises the question of whether taxation is the primary motivation for relocation or whether other economic and non-economic factors also play a role.

Relocation and Migration Responses

Research by Moretti and Wilson (2023) provides evidence that wealthy individuals may respond to changes in taxation through relocation. In their paper, they studied the Forbes 400 and examined differences in estate taxation between US states, estimating that estate taxes had a significant effect on the geographical location of billionaires. However, this study compares states within one country and does not establish the same effects occurring internationally. International relocation involves many more factors that individuals would consider. The findings from one US interstate study cannot establish that taxation causes relocation similarly in an international context.

Similar research has examined whether the same occurs internationally. Denmark introduced a special tax scheme specifically targeted towards high-earning foreign workers that were coming to Denmark, so those that qualified could receive substantially lower taxation on their earnings compared to what they would face in Denmark's normal tax system. Research finds that the number of highly paid foreigners above the eligibility threshold doubled relative to those just below it after the scheme was introduced (Kleven et al., 2013), providing international evidence that tax incentives can influence where high earners choose to locate. However, it is important to mention that the study mentions highly paid foreigners, not necessarily ultra-high-net-worth individuals. Baselgia and Martínez (2024) studied super-rich foreigners in Switzerland and examined a special tax regime based on living expenses rather than actual income or wealth. Some cantons repealed the regime due to concerns over tax equity and public discontent. Following the repeal, the stock of super-rich foreigners in those cantons fell by 43%. The super-rich Swiss individuals who were unaffected by the reform did not show the same response, which provides international evidence that taxation can affect location decisions among very wealthy individuals. However, this evidence comes from a specific Swiss tax regime. It should not automatically be generalised to other taxes or countries.

Jakobsen et al. (2024) examined how wealth taxation affects international migration among wealthy taxpayers using Scandinavian administrative data. A 1 percentage-point increase in the top wealth-tax rate was associated with a 2% reduction in the stock of wealthy taxpayers. Although wealthy taxpayers responded through migration, the estimated overall effects on employment, investment and value added were relatively modest, which provides further evidence that taxation can influence location decisions, while also showing that behavioural responses among wealthy taxpayers do not necessarily produce equally large economy-wide effects.

Non-Tax Factors Constraining Relocation

Even though these studies demonstrate taxation influencing location decisions, international relocation decisions also depend on several non-tax factors, including political stability, business opportunity, quality of life, safety and social life. Research shows that wealthy individuals favoured investments in "place specific cultural capital above the relatively small gains in economic capital associated with tax migration" (Friedman et al., 2025) meaning that the potential financial benefits of tax migration can be outweighed by the cultural, social,

professional and personal benefits wealthy individuals receive from remaining in their existing location. Friedman et al.'s (2025) findings provide further evidence of this effect. Out of the 35 members of the UK's wealthiest 1%, none of them were actively considering relocation primarily for tax reasons (Friedman et al., 2025). Some of the reasons included family disruption, career and business considerations, cultural and lifestyle preferences, attachment to their existing location and even reputational concerns. However, Friedman et al.'s study may be affected by survivorship or selection bias, as its sample could underrepresent wealthy individuals who had already permanently relocated for tax reasons. Although some interviewees had experience living abroad, the findings may place greater emphasis on the views of individuals with continuing ties to the UK. Therefore, the study cannot establish that tax considerations are unimportant to wealthy individuals generally.

This contrasts with the evidence by Moretti and Wilson (2023) that people respond to taxation through changes in location decisions. Therefore, the literature so far suggests taxation can influence ultra-wealthy location decisions, but the extent of that influence depends on the individual circumstances and the non-tax benefits associated with their existing or potential location.

Income Reporting and Tax-Planning Responses

Relocation is not the only behavioural response to taxation. Wealthy/high-income taxpayers can also respond through financial decisions, including changing when income is reported or realised. In 2013, the US increased its top marginal tax rates, which raised top rates in capital income by about 9.5 percentage points and on labour income by about 6.5 percentage points (Saez, 2017). Saez examined IRS tax statistics to see how top earners responded. Findings by Saez show that income reported by the top 1% was unusually high in 2012 and unusually low in 2013, indicating that many top earners shifted income forward into 2012 before the higher 2013 tax rates were set. Saez estimates that the top 1% shifted about 11% of their 2013 income into 2012 (Saez, 2017), which was strong for capital gains, dividends, business income, and was strongest among people at the very top of the income distribution.

This evidence suggests that taxation can indeed influence top-income earners' behaviour even when individuals don't relocate. Saez's findings suggest that much of the response reflected short-term income shifting rather than a permanent behavioural change. As well as this, this study, much like the Moretti and Wilson study, is US-specific, which is a limitation, meaning the findings cannot automatically be generalised to the ultra-wealthy individuals operating different tax systems internationally.

Investment and Asset-Realisation Responses

Taxation can affect not only the timing of reported income but also the investment and asset-realisation decisions. Agersnap and Zidar (2021) study how capital-gains tax rates affect capital-gains realisations. When an investor owns an asset that has increased in value, the capital gain generally is realised when the asset is sold. They find that capital-gains realisations respond

negatively to tax rates, where higher rates are associated with fewer realisations, providing further evidence that taxation can change financial behaviour without requiring physical relocation. Again, this evidence is largely US-specific and concerns investors/capital gains rather than exclusively ultra-high-net-worth individuals.

Comparative Assessment: When Does Taxation Matter Most?

Taken together, literature suggests that taxation's influence on the ultra-wealthy is real but uneven, and the unevenness itself is informative. When the evidence is separated into studies of financial behaviour that stop short of relocation (Saez, 2017; Agersnap & Zidar, 2021) and studies of full international or interstate relocation (Moretti & Wilson, 2023; Kleven et al., 2013; Friedman et al., 2025), a consistent pattern can be identified: responses to taxation may be stronger when adjustment costs are lower, and weakest and most contested where the cost of adjusting is highest.

This suggests that taxation does not necessarily drive relocation on its own. The reviewed evidence suggests that taxation may influence relocation alongside other factors, rather than acting as the sole cause; it is more often seen as the amplifier or the tipping point, and the differences between these findings may partly be explained by the different populations and forms of taxation examined across the studies. For example, Moretti and Wilson studied billionaires in US states, Kleven et al. studied high-earning foreign workers, Friedman et al. studied the top 1% in the UK, Saez studied US top-income earners and Agersnap and Zidar studied capital-gains taxpayers and investors. The studies examine different populations and behavioural outcomes, which limits direct comparison between their findings. Because countries have very different tax systems and wealthy individuals have different personal circumstances, findings from one country cannot automatically be applied internationally.

Conclusion

The reviewed literature provides evidence that taxation influences certain behaviours of wealthy and high-income individuals, but the extent of its influence depends on the type of behaviour being considered. When it comes to financial behaviour, taxation produces observable changes in income reporting, asset realisation and other financial decisions, whereas relocation is more complex because it involves additional considerations. Moretti and Wilson (2023) and Kleven et al. (2013) provide evidence that taxation can influence location. Additionally, Saez (2017) and Agersnap and Zidar (2021) provide evidence that taxation changes financial behaviour without relocation. Friedman et al. (2025) provide evidence that tax savings can be outweighed by family, career, cultural and lifestyle considerations. The reviewed evidence suggests that taxation seems more capable of influencing behaviours that have relatively low adjustment costs, such as changing timing of income or capital-gains realisation, whereas moving to another country is a substantially larger decision. This evidence therefore suggests that taxation acts as one incentive or tipping factor rather than automatically causing relocation. However, this evidence has several limitations, such as several studies being US-specific, some

examining high-income earners rather than specifically ultra-high-net-worth individuals, and the fact that studies examine different forms of taxation. Governments face a trade-off between raising sufficient revenue for public services and designing taxation systems that account for the behavioural responses of highly mobile and financially flexible taxpayers. Future research should examine whether these behavioural responses also occur among genuinely ultra-high-net-worth individuals and across a wider range of countries and wealth-tax systems.

Works Cited

Luzkow, J. L. (2018). The Politics of Taxes (pp. 187–239). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-93994-0_6

Taxation | the official platform of the UAE government. (2025). https://u.ae/en/information-and-services/finance-and-investment/taxation

Henley & Partners, H. &. (2025, June 24). Millionaires on the move: UK braces for historic wealth flight as global migration peaks. Henley & Partners. https://www.henleyglobal.com/newsroom/press-releases/henley-private-wealth-migration-report-2025

Kleven, H. J., Landais, C., Saez, E., & Schultz, E. (2013). MIGRATION AND WAGE EFFECTS OF TAXING TOP EARNERS: EVIDENCE FROM THE FOREIGNERS' TAX SCHEME IN DENMARK. The Quarterly Journal of Economics. https://eml.berkeley.edu/~saez/kleven-landais-saez-schultzQJE13danishscheme.pdf

Government Digital Service. (2024, November 7). Income Tax rates and Personal Allowances. GOV.UK. https://www.gov.uk/income-tax-rates

Moretti, E., & Wilson, D. J. (2023). Taxing billionaires: estate taxes and the geographical location of the Ultra-Wealthy. American Economic Journal Economic Policy, 15(2), 424–466. https://doi.org/10.1257/pol.20200685

Friedman, S., Gronwald, V., Summers, A., & Taylor, E. (2025). 'But Switzerland's boring': tax migration and the pull of place-specific cultural capital. Socio-Economic Review, 23(3), 1091–1112. https://doi.org/10.1093/ser/mwaf002

Saez, E. (2017). Taxing the Rich More: Preliminary Evidence from the 2013 Tax Increase. Tax Policy and the Economy, 31(1), 71–120. https://doi.org/10.1086/691084

Agersnap, O., & Zidar, O. (2021). The tax elasticity of capital gains and Revenue-Maximizing rates. American Economic Review Insights, 3(4), 399–416. https://doi.org/10.1257/aeri.20200535

Stiglitz, J. E., & Rosengard, J. K. (2015). Economics of the public sector: Fourth International Student Edition. W. W. Norton & Company.

Jakobsen, K., Kleven, H., Kolsrud, J., Landais, C., & Muñoz, M. (2024). Taxing Top Wealth: Migration Responses and their Aggregate Economic Implications. In National Bureau of Economic Research. https://doi.org/10.3386/w32153

Baselgia, E., & Martínez, I. Z. (2024). Mobility Responses to Special Tax Regimes for the Super-Rich: Evidence from Switzerland. The Economic Journal, 135(668), 1388–1409. https://doi.org/10.1093/ej/ueae101

The Wealth Report 2023. (2023). https://content.knightfrank.com/resources/knightfrank.com/wealthreport/2023/the-wealth-report-2023.pdf

Reforming the taxation of non-UK domiciled individuals. (2024, October 30). GOV.UK. https://www.gov.uk/government/publications/tax-changes-for-non-uk-domiciled-individuals/reforming-the-taxation-of-non-uk-domiciled-individuals

Additive Manufacturing Technology in Civil Engineering and Its Application in Seismic Zones By Yerbol Janzakov

Abstract

The construction industry faces mounting pressure to reduce waste and build times while improving structural resilience, particularly in regions prone to earthquakes. Additive manufacturing (AM), commonly known as 3D printing, has emerged as a promising alternative to conventional construction methods, offering rapid layer-by-layer fabrication of structural elements with reduced material waste and labor requirements. This paper examines the current state of 3D printing implementation in civil engineering. It evaluates its structural, economic, and environmental advantages alongside its technical limitations, including job displacement, interlayer bonding weaknesses and structural instability, equipment and cybersecurity risks, and building services engineering problems. Particular attention is given to the application of 3D-printed structures in seismically active zones, where the technology's layered construction method raises unique concerns regarding tensile strength, reinforcement integration, and dynamic load resistance during ground motion events. Through a review of existing case studies, material testing data, and seismic performance models, this paper assesses whether 3D-printed construction can meet the structural demands of earthquake-prone regions, or whether current limitations necessitate hybrid approaches combining additive manufacturing with traditional construction techniques. The findings aim to clarify the practical viability of 3D printing as a mainstream construction method and identify key research gaps that must be addressed before widespread adoption in high-risk seismic areas.

Keywords additive manufacturing, 3D printing, seismic resilience, earthquake, engineering, applicability, limitations, earthquake-prone regions, civil engineering

1. Introduction

Throughout history, humankind has always faced the problem of efficient and quick manufacturing of buildings. Thus, civil engineers have come up with various techniques and materials that they use to improve the efficiency rate and structural resilience, as well as making a building look better (Bechthold and Weaver). This paper discusses a new construction method known as additive manufacturing (AM).

In recent years, the additive manufacturing technology has moved from niche to a commonly used strategy. Additive manufacturing is a technique of creating objects (from simple toys to complex aerospace parts) (Zhou et al.) in a layer-by-layer method directly from the 3D model that has been designed before. This methodology has already been implemented in various fields of human activity. 3D printing is widely used in aerospace, biomedicine, automotive, and energy industries. The objects are created using provided computer-aided design (CAD) models.(Zhang et al.)

Additive manufacturing technology is spreading across different engineering fields, and it has already been integrated into civil engineering and architectural design.

Initially, AM was introduced for several reasons (Besklubova et al.; El-Sayegh et al.), and this invention was a way to solve problems including:

- Expensive construction, including high formwork labor costs
- Bad environmental performance and high emission rate
- Slow construction due to large supply chains
- Low geometric freedom
- Unsafe sites

Despite AM's rapid adoption and clear economic benefits, the structural viability of 3D-printed buildings remains subject to critical limitations when exposed to extreme environmental forces. This question is particularly crucial since people from various parts of the world suffer from earthquakes and their consequences. Applying 3D concrete printing (3DCP) in seismically active regions introduces unique engineering challenges. Because the printer builds walls layer by layer, the joints between those layers are often weak. In addition, it is also very difficult to put standard steel rebar inside 3D-printed concrete to make it stronger. Since earthquakes push buildings hard from side to side, people need to understand how printed concrete handles these heavy vibrations before we can safely build homes with it in active earthquake zones. Therefore, this paper reviews the current situation of AM in civil engineering in seismic zones.

This review paper evaluates the structural, economic, and environmental performance of additive manufacturing in civil engineering, with a specific focus on high-risk seismic zones. 3D printing makes construction faster and allows for unique building designs. Although it has several benefits that may positively affect civil engineering, it also brings several drawbacks. The main drawbacks are more impactful than benefits. Given this, to implement AM in seismic zones, engineers need to take a different approach.

The following section details the historical evolution, core materials, and primary printing techniques used in construction AM. Section 3 analyzes the broader techno-economic and environmental advantages and limitations of 3DCP. Section 4 examines the specific structural vulnerabilities and potential performance benefits of printed structures under dynamic seismic loading.

2. Fundamentals and Background

Before diving into the benefits and drawbacks of AM, it is important to establish the fundamentals(materials and techniques) and the history of the technology. This information introduces the reader to this technology. At the end, this section will summarize the information and give data about the most commonly used and the most convenient material and technique to use.

2.1 History

AM technology emerged in the late 20th century, when many engineers tried to adopt 3D printing into civil engineering practice.(Wohlers et al.) At that time, people only started to come up with effective techniques and ways of implementing AM in civil engineering. During the same period, engineers tried to identify which materials are better to use for construction and what the benefits of using them are.

One of the first real examples and methodologies for AM in civil engineering was described in an empirical paper by Joseph Pegna ”Exploratory investigation of solid freeform construction” (Pegna). In this paper, J. Pegna revolutionised the field of civil engineering. Before him, engineers were tackling automation completely differently. They tried to keep building bigger and bigger machines to assemble bigger and bigger components, replacing human hands with robot hands doing the same complex tasks. Therefore, J. Pegna flipped the entire philosophy. Instead of constructing large-scale machines to build large structures, he built a structure from an enormous number of small-scale, simple operations. This is directly borrowed from additive manufacturing logic, where complexity in the final shape comes from simplicity in the repeated operation. This technology was named Contour Crafting (CC), and it allowed an engineer to layer sand and Portland cement and therefore make complex buildings. (Figure 1)(Pegna)

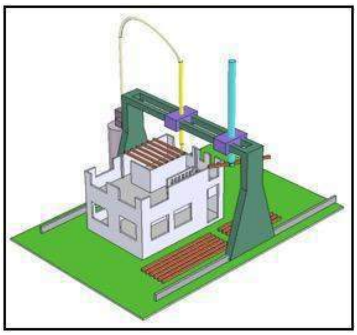


Figure 1: Construction of conventional buildings using CC

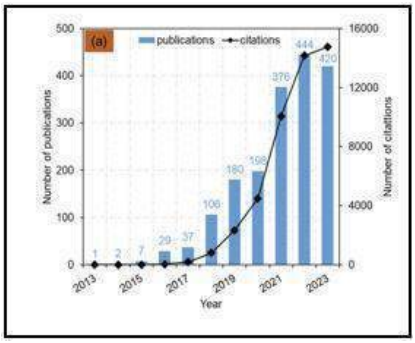


Figure 2: The number of citations and publications on 3D printing

In recent years, the popularity of this technology has drastically increased, which is reflected by the number of publications and citations that are related to the field of additive manufacturing. (Figure 2)(Zhuang et al.)

This illustrates that, although AM is a relatively new technology, the popularity of it has increased, and it can be used for large-scale buildings. However, to tackle the main benefits and drawbacks of this tech paper needs to address the main materials and techniques that are implemented for the manufacturing.

2.2 Materials

Overall, researchers and civil engineers divide materials into four main groups: binders, aggregates, fibers, and admixtures. While all of them are used for AM each of these groups differs from each other, providing valuable benefits and distinct drawbacks (Hutyra et al.), as shown in Table 1.

Binders. The most common binder in 3D concrete printing is Ordinary Portland Cement (OPC), which works by reacting with water to harden and bind the mix together. It is widely used due to its availability and well-understood behavior, though its production releases significant CO_2. Geopolymers serve as the main alternative, produced from industrial waste materials like fly ash or slag activated by an alkaline liquid, emitting around five times less CO_2 than OPC. For situations requiring faster hardening, Calcium Sulfoaluminate (CSA) cement is used, as it gains strength much more rapidly than OPC. Supplementary cementitious materials like silica fume, fly ash, GGBS, and limestone powder are commonly added to partially replace OPC, reducing cost and emissions while fine-tuning properties like flowability and setting time.

Aggregates. Aggregates act as the structural filler in 3DCP mixes, taking up volume and contributing to compressive strength. Fine sand with particles up to 2mm is the universal standard, as the narrow printer nozzle physically prevents the use of larger particles like gravel. Desert sand has been researched as a regional alternative, offering consistent grain size that improves mix predictability. Recycled aggregates from demolished buildings, including crushed concrete, glass, and rubber, can partially replace natural sand to reduce waste, though they tend to lower compressive strength as substitution rates increase due to higher porosity.

Fibers. Since traditional steel reinforcement bars cannot be integrated into the printing process, fibers are mixed directly into the concrete to compensate for its natural weakness under tension and bending. Steel fibers are the most widely used, bridging cracks and improving layer-to-layer bonding, while polypropylene fibers serve a more specific role of controlling early shrinkage cracking at lower cost. PE and PVA fibers are used in high-performance mixes designed for maximum ductility and crack resistance. Carbon fibers offer a superior strength-to-weight ratio but remain too expensive for widespread use, while basalt and glass fibers present challenges due to directional alignment during printing, which creates uneven strength distribution across the structure.

Chemical Admixtures. Admixtures are chemicals added to control the behavior of concrete during and after the printing process, making the balance between flowability and rapid

hardening achievable. Superplasticizers allow the mix to flow easily through pumps and nozzles without adding extra water, which would weaken the hardened structure. Retarders slow hardening during transport and early printing stages, while accelerators speed it back up at the nozzle so each deposited layer hardens quickly enough to support the next. Viscosity-modifying agents like nanoclay thicken the mix after deposition to prevent sagging, and air-entraining agents create small air bubbles that improve both flowability and freeze-thaw resistance of the hardened structure.

Table 1: Four main categories of materials for additive manufacturing

Category	Main options/material	Primary role	Main trade off
Binder	OPC, geopolymer, CSA, blended	Cohesion, structural integrity	releases some co2, high cost
Aggregates	Fine sand, desert sand, recycled	Filler, compressive strength	Coarse excluded — limits strength
Fibers	Steel, PP, PE, PVA, carbon, basalt	Tensile strength, crack control	Reduces flowability, expensive
Admixtures	Superplasticizers, retarders, accelerators, VMAs	Control workability and setting	Complex balancing act

For commercial buildings, the most widely adopted materials in construction AM are cement-based mixes, valued for their workability and structural properties, alongside sand-based composites used in binder jetting processes, and more recently fiber-reinforced and geopolymer alternatives.

2.3 Techniques

The techniques used for 3D printing vary depending on the paper that is being examined. However, experts in civil engineering most commonly highlight four techniques for building 3D printed structures. Each of these techniques is unique, and on their base methodology, the other various AM techniques are created. The main difference between them is the type of raw materials they use and the tools that are needed for successful construction completion.

Contour Crafting (CC). Contour Crafting is the most widely used technique in construction 3D printing, developed by Dr. Behrokh Khoshnevis at the University of Southern California in the mid-2000s. The process works by mounting a nozzle on a robotic arm or gantry system, which follows a programmed path from a CAD model and extrudes concrete layer by layer. What makes Contour Crafting distinct from simple extrusion is that the nozzle uses two trowels on its sides to smooth the outer surface of each printed layer as it moves, producing cleaner wall edges than other extrusion methods. The printer can operate on-site by moving directly to the construction location, or off-site in a factory setting producing prefabricated

elements that are then assembled. As Nadarajah (Nadarajah) describes, the technique comes in two main configurations: framed printers, which are large gantry systems fixed to a frame that must be bigger than the structure being printed, and non-framed printers, which are robotic arms mounted on mobile vehicles that can move freely on uneven ground. Most major real-world 3DCP projects use this approach, including Winsun's printed apartment blocks(Figure 3) and Apis Cor's Moscow house (Figure 4).

Figure 3: WinSun's printed apartment blocks

Figure 4: Apis Cor's Moscow house

Selective Laser Sintering (SLS). Selective laser sintering, a variant of powder bed fusion, is a process used to produce objects from powdered materials using one or more lasers to selectively fuse the particles at the surface, layer upon layer, in an enclosed chamber. The printing system consists of a laser source, a scanning system that directs the laser beam, a roller that spreads fresh powder, and two platforms — one that holds the powder supply and one that holds the object being built (Figure 5). At the start of each layer, the roller spreads a thin, even layer of powder across the build platform. The laser then traces the cross-section of that layer, partially melting and fusing the powder particles together in the desired pattern. After the first layer of powder is fused, the sintering platform moves a step downward, and the next layer of powder is fused. Crucially, the surrounding loose powder that has not been sintered acts as a natural support for the structure as it grows, which means overhanging shapes and complex internal geometries can be printed without needing extra support structures — a significant geometric advantage over extrusion methods. Once printing is complete, the un-sintered powder is removed, and the part is extracted. The energy density — determined by laser power, scan velocity, beam diameter, and hatch spacing — is the most vital parameter responsible for the overall process and part properties. In construction, SLS is not yet practical for full-scale printing due to the high energy cost and the difficulty of scaling up the laser system, but it is relevant for printing precise, smaller structural or architectural components.(Kafle et al.; Gueche et al.)

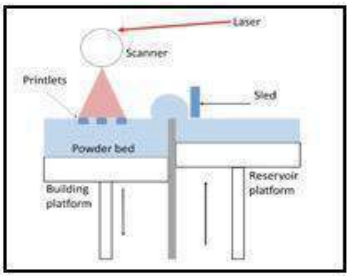


Figure 5: How SLS works

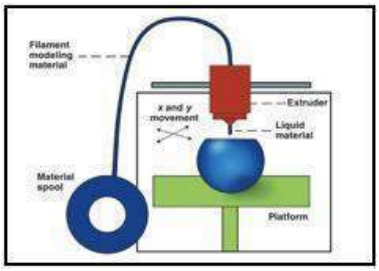


Figure 6: How FDM works

Fused Deposition Modelling (FDM). Fused deposition modelling, also known as fused filament fabrication, is a process of depositing thermoplastic filaments layer by layer on a build platform. The material (typically a plastic) starts as a solid filament wound on a spool. It is fed into a heated nozzle that melts it into a semi-solid state, then deposits it onto the build platform following the path generated by slicer software from the CAD model. As each line of material is deposited, it immediately cools and hardens, bonding to the layer below it. Once one layer is complete, the nozzle moves up by one layer thickness, and the process repeats until the full object is built. The nozzle temperature, bed temperature, and layer height are the responsible parameters for the mechanical behaviour of the 3D printed parts. In construction, FDM is not used to print structural concrete elements, but it is widely used to produce complex custom formwork molds, architectural models, and small structural components. It is also the most straightforward and accessible technique of all discussed.(Figure 6)(Kafle et al.)

Binder jetting (BJ). Binder jetting is an additive manufacturing technique that dispenses a liquid binding agent onto powder to form a two-dimensional pattern on a layer, with the layers stacked to build a physical article. The process starts with a thin layer of powder, such as sand, ceramic, metal, or a composite material, spread evenly across a build platform by a roller or blade. A printhead, working similarly to an inkjet printer, then moves across the powder bed and selectively deposits liquid binder only onto the areas that should become solid, based on the cross-section of the CAD model for that layer. The binder bonds the powder particles together wherever it lands. A fresh layer of powder is then spread on top, and the process repeats until the full three-dimensional object is built. The main advantages over extrusion methods are the geometric freedom and the ability to use locally available materials. However, developing

printing and post-processing methods that maximize part performance remains a challenge, and surface finish tends to be rougher than extrusion-based methods.(Figure 7)(Ziaee and Crane)

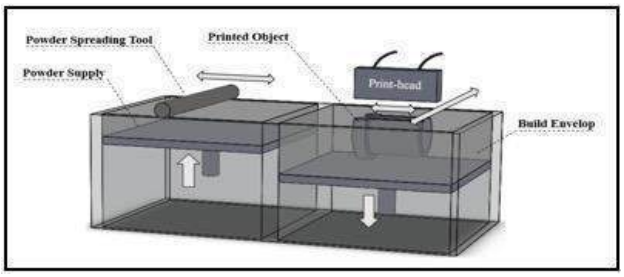


Figure 7: How BJ works

The main material categories- binders, fibers, admixtures, and aggregates provide unique benefits and various drawbacks of using them. The same thing goes for the techniques. Each method of AM provides various opportunities and challenges. Although each of these four main materials and techniques of 3D building is important to the field of AM, there is a combination of them that is more popular and more suitable for large-scale production of 3D-built structures. (Žujović et al.). This combination is Contour Crafting(CC) and OPC(Ordinary Portland Cement). CC is popular in civil engineering due to its flexibility because it can work both on-site and off-site. In addition, CC provides smooth contouring, which leads to better wall quality. OPC also has several advantages over other types of materials. Firstly, concrete is a cheap binder, which is suitable for mass construction of buildings. Secondly, it is predictable and well known, since it has already been used for several decades. With all that being said, this combination (CC plus OPC) is the most popular and convenient method of AM.(Perkins and Skitmore; Raphael et al.). The next chapter discusses the main benefits and drawbacks of AM.

3. Additive manufacturing benefits and drawbacks

The paper has established the core materials and methods that are used for AM. It has provided all the fundamentals needed for understanding what AM even is. In the previous chapter, the paper has established that Contour crafting using OPC is the most popular way for AM and that it is the best way for building structures. However, even though AM is rapidly evolving and increasing its popularity in the civil engineering industry, adopting it includes challenges, such as interlayer bonding weaknesses and structural instability, equipment and cybersecurity risks, job displacement, and building services engineering problems. Despite all these drawbacks, AM technology provides certain benefits that are highly valuable for the construction industry. Such benefits are environmentally friendly construction, structural freedom, low cost, high productivity and efficiency, and social benefits. Although the challenges facing AM are significant, the technology's long-term environmental, economic, and design advantages ultimately outweigh these drawbacks, making it a viable and transformative evolution for modern construction.

3.1 Benefits

This section's main aim is to address the main benefits of implementing 3D concrete printing into civil engineering. These pros of AM are compared in regard to traditional construction methods. Traditional civil engineering has many problems that can be solved (partially or completely) using AM. The advantages that are discussed later in this section can be divided into five main categories, which this paper reviews. As it has already been mentioned earlier, these main pros of AM are:

- Environmentally friendly construction
- Structural freedom
- Low construction cost
- High productivity
- Social benefits

3.1.1 Environmentally friendly construction

Implementing AM in the construction industry will be sustainable and ecologically friendly. This technology provides several benefits for the environment. These advantages are: the reduction of emissions, waste, and reduced formwork. The research that was conducted in the past decades has concluded that 3D printing in civil engineering significantly reduces the mass of emitted gases compared to traditional cement construction. (Gangotra et al.). The researchers conducted an experiment that illustrates that the optimized 3DCP achieves up to 53.1 % reduction in life cycle carbon emissions compared to traditional methods.(Mu et al.). Another research study demonstrates that buildings that were constructed using AM have less effect on global warming, and these constructions use less water(Abdalla et al.).

In addition to the atmospheric benefits of implementing 3DCP, there are benefits in regard to the amount of waste that is being thrown away. There are many studies that actually address the harmful effects that conventional construction has on the environment. Conventional construction severely impacts the environment by generating high carbon emissions, consuming vast resources, and producing massive amounts of material waste. (Mohajan; Tafesse et al.). Thus, AM is a good solution that can reduce the amount of waste that is released into the environment. Moreover, 3DCP significantly enhances resource efficiency compared to conventional construction by eliminating temporary formwork and other wasted materials. By depositing material exclusively where structural design requires it, 3D printing minimizes on-site material waste and cuts operational costs. Furthermore, its capacity to fabricate intricate, topology-optimized geometries enables lightweight structural components that further reduce overall material consumption. (Campbell et al.). Moreover, if implemented, AM will significantly decrease the formwork that is being used on construction sites. This reduction will positively impact the amount of waste that is thrown away. (Delgado Camacho et al.). Thus, the integration of AM into civil engineering is an efficient way to reduce the amount of emitted gases as well as wasted materials and formwork.

3.1.2 Structural freedom

The other major advantage of AM over traditional construction is the opportunity to create complex geometrical shapes of structures. Since AM is a layering technique, it allows engineers to construct complicated shapes that cannot be produced by any means using basic construction methods.(Campbell et al.). As can be seen in Figure 8 below, AM is used for making curved constructions; it is very easy to build curved buildings using AM that are hard to build using traditional manufacturing techniques. At the same time, this variety of shapes allows buildings to be aesthetic, which helps structures to become art installations and cultural heritage (Feng).

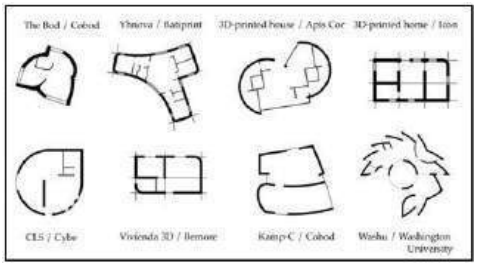


Figure 8: Plans of some 3D-printed constructions

3.1.3 Low construction cost

Another benefit of AM over conventional construction is the cost: 3D printing is much cheaper. Firstly, as it was mentioned before, AM helps reduce waste, which is crucial, since many unused materials are later thrown away. The material waste is 30% of all cost overrun(Tafesse et al.). Furthermore, in traditional manufacturing, 9.06% of all cement is wasted. Thus, the operational costs decrease (Delgado Camacho et al.). In addition to the reduced wastage, AM offers a reduction in the cost of transportation and storage (Hager et al.). For instance, Chinese company WinSun has built a mansion and an apartment building using AM. (Figure 3, Figure 9) 3D printing saves between 30 and 60 percent of construction waste, and decreases labour costs by between 50 and 80 percent. (WinSun). AM positively affects construction costs.

Figure 9: WinSun's cottage

3.1.4 High productivity

The productivity and efficiency of construction are drastically increased in comparison with traditional manufacturing. In the example of WinSun, the production time is decreased by between 50 and 70 percent.(WinSun). Furthermore, in 2014 WinSun claimed to have printed ten houses in a single day in Shanghai, and in 2016 its Dubai Office of the Future needed 17 days of printing and two days of installation on site (ASME). Another company that has significant improvements in manufacturing speed is CyBe. They state that their materials enable a fast printing process as well as improvements in stability and resilience of their building(CyBe Construction).

In addition to the improved printing work, AM has a shorter supply chain, which is important for emergency situations (Koç et al.) and beneficial for mass manufacturing. For mass manufacturing, the aim of AM is to significantly diminish human intervention and complex formworks.(Ghaffar et al.)

In conclusion, AM is a highly productive way of building complex structures. It not only decreases the supply chain but also significantly lowers the construction time.

3.1.5 Social benefits

Many people tend to overlook the social advantages of AM. One of the benefits is the safety of sites. According to the International Labour Organization, at least 108,000 workers are killed on construction sites every year, which is about 30 percent of all fatal occupational injuries (International Labour Organization). The implementation of 3D printing may reduce this number. By automating nearly every part of the construction process, AM helps to save people's lives and provides safer working conditions (El-Sayegh et al.). Furthermore, AM is able to change social structure in general, because it changes the way people work on worksites. In the past, the working site was harsh and demanding, and it required men to work on sites and lift heavy weights. Today, as the work becomes more automated, it may become more accessible to women, who hold only a small share of construction trade jobs (about 4% in the United States) partly because of physically demanding tasks; however, the evidence on how new technologies affect women in male-dominated jobs is still mixed (Rodgers et al.). In addition, the number of workers is reduced. This is beneficial for some companies since they have to pay less (Delgado Camacho et al.).

3.2 Drawbacks

After reviewing the advantages of AM, it is clear that it is a beneficial technology. It has several advantages that significantly affect civil engineering in general. Nevertheless, it is worth mentioning the drawbacks of this system, so that readers can be aware of the limitations this technology brings. Some of the disadvantages may fall into categories that were already mentioned in the advantages; that is because some benefits create the drawbacks at the same time. For instance, AM brings social job displacement, while providing social benefits, such as greater accessibility for women and worker safety. This section addresses the primary drawbacks

of implementing AM in civil engineering, drawing on the same comparison to traditional methods used above. The main cons fall into four main categories:

- Job displacement
- Building services engineering problems
- Equipment and cybersecurity risks
- Interlayering bonding weaknesses and structural instability

3.2.1 Job displacement

One of the main threats of AM is job displacement. According to the OECD, approximately 220 million people, about 7% of global employment, work in the construction sector (OECD). There is a major risk that 3D printing will reduce the number of workers needed to construct a building. In the Netherlands, 3D printed residential buildings have reduced the number of on-site workers by approximately 70% (Banihashemi et al.). The World Economic Forum's 2016 Future of Jobs report projected a net loss of more than 5.1 million jobs across 15 major economies between 2015 and 2020 because of disruptive labour market changes such as automation and robotization, with about two-thirds of the total losses concentrated in office and administrative roles (World Economic Forum). However, despite the rapid adoption of AM in civil engineering, the threat of job displacement is not yet significant. Only 356 3D-printed buildings have been constructed from 2013 to 2024 (University of Florida Warrington College of Business). This trend is confirmed by Gerrit Placzek et al (Placzek and Schwerdtner) and visualized by Washington College of Business (University of Florida Warrington College of Business).(Figure 10)

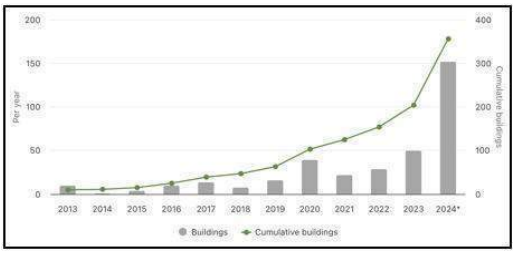


Figure 10: 3D printed buildings worldwide

Nevertheless, the threat of job displacement is still a major concern in the civil engineering field. Because, based on the same Figure 10, it can be seen that the number of constructed buildings using AM has rocketed in 2024. It means that the number will certainly increase; the risks of job displacement among builders may rise in the next decades.

3.2.2 Building services engineering problems

Despite AM offering fast construction time and unique building shapes, it also has one major internal issue — building services engineering problems. Adapting MEP to a 3D printed house is a challenge. Plumbing systems, electrical sockets, and windows have not been

integrated into the design process yet (El-Sayegh et al.). In WinSun projects, one of the issues was the integration of MEP into a building. The 3D printing process does not include building systems. Thus, the implementation of these systems required additional work and caused problems with the structural resilience and integrity (Wu et al.).

Even if the problem of integration is resolved, there will still be a major unresolved challenge and drawback. This disadvantage is the long-term sustainability and serviceability of MEP connections. Since the long-term performance of printed buildings, including their building services, has received little study, it remains an open problem (Wu et al.). Building services also have a much shorter service life than the structure: on average about 20 years, compared with a design life of 50 years for many buildings (NBS). These are the main concerns and drawbacks of AM in MEP and building services.

3.2.3 Equipment and Cybersecurity risks

This part combines two large categories of AM drawbacks: equipment risks and cybersecurity problems. Starting with the equipment risks category, the adoption of AM into civil engineering comes with several harmful equipment drawbacks. Because 3D printing equipment costs a lot of money, it creates pressure on companies. If the equipment is broken, the repair of it may cost a lot of money, while traditional manufacturing relies on relatively cheap tools and machinery.(El-Sayegh et al.). In addition, it also entails a malfunction and hazard. The electrical, mechanical, and chemical systems might not be fully verified. For instance, Robotiq states that electronic components might sometimes be counterfeit.(Robotiq) That is why it is crucial to mitigate these malfunctions and electrical hazards. Mechanical and electrical malfunctions pose a major threat to the safety of on-site workers. (Rojas et al.). That being said, it is crucial to develop safety standards so no equipment failure occurs. Nevertheless, nowadays, the equipment risks are still a huge drawback of AM implementation.

In addition to the equipment-related challenges, AM has cybersecurity risks. To clarify, cybersecurity risks are the risks that are related to the breach of computer networks by hackers(design theft, network breach, digital sabotage)(El-Sayegh et al.). The hacking can be divided into three categories:

- Printer hacking
- Raw materials attack
- Design hacking

Each category can be attacked. Thus, the cybersecurity of AM is low(Gupta et al.). Since AM largely relies on CAD models (Zhang et al.), hacking or stealing them is a huge threat for AM companies. The CAD design can be stolen and damaged. However, this drawback can be solved using special supply chains and digital threads that would mitigate the risks related to cyberhacking(Gupta et al.).

These two issues pose a big threat to the AM industry, and although there are some improvements in regard to these two problems, the equipment and cybersecurity risks have to be mentioned and analyzed.

3.2.4 Interlayer bonding weaknesses and structural instability

The next major drawback of AM is the weakness of interlayer bonding and the instability of a structure. The main problem is that in conventional cast concrete, the whole mass sets simultaneously; in comparison, in 3D printing, layers are deposited one at a time with a time gap between them. During that gap, the lower layer starts hardening, drying, and hydrating, while the layer about to be printed on top of it is still fresh. That mismatch means the two layers never fully fuse into one continuous material. Instead, a weaker "transition zone" forms at the interface, marked by extra pores, gaps in the hydration process, and localized soft spots. This main problem causes walls to be less resilient, stable, and not hard enough, compared to traditional construction(Zhang et al.). One study found that extending the delay between layers from 3.5 to 7 hours dropped relative bond strength from 72.7% down to 25.6%, while porosity nearly doubled (36.8% → 62.8%) and stiffness dropped by more than half (26.2 → 10.7 GPa)(Geng et al.). The main contributing factors to this are printing speed, layer height, nozzle distance, and overlap between filaments. Another study has found that the compressive strength of a 3D printed cube was 32.3~42.6% lower than that of the monolithic one. In addition, in the same paper, they found that the tensile strength was 11.1~39.7% lower than that of the monolithic one(Joh et al.).

After reviewing the benefits and drawbacks of this technology, it can be stated that the advantages of AM are promising and definitely viable in the future. It is important to develop this technology and try to implement it more widely all around the world. That being said, the performance of AM technology under extreme conditions is unclear. That is why the next chapter is dedicated to the application of AM in seismic regions. This is an important topic, since AM should guarantee the best performance in earthquake-prone regions. It is still unclear whether it can be implemented in seismic areas or not.

4. Implementation of AM in seismic zones

One of the largest concerns regarding AM is its performance in earthquake-prone regions. During earthquake loading, construction experiences substantial stress, so that its dynamic response can maintain resilience and structural integrity while providing human serviceability. The four main criteria of a seismic-resilient building are: ductility, damping, stiffness, and energy dissipation. Ductility is the ability of a building to experience large inelastic deformation without breaking; damping characterizes how well a building can dissipate vibrational energy under seismic load; stiffness refers to the resistance to the applied seismic loads; energy dissipation is the ability to absorb the seismic energy through inelastic deformation and damage mechanisms. (Figure 11)(Berwal et al.). In addition to these criteria, it is crucial to take into account the ramifications of earthquakes and also consider fast post-earthquake recovery speed as another

important criterion. These five criteria are essential for the successful seismic resilience of a building.

Much of the past research about seismic resilience of 3D-printed buildings was conducted using seismic simulators and was tested only in laboratories. That changed when Heidelberg materials constructed the first earthquake-resistant building in Naples, Italy.(Heidelberg Materials) They claim that it can withstand seismic loading in the region where it is located. However, the building itself has not yet been tested, since there have been no earthquakes since its construction; this is why this paper analyses the broader picture by collecting data from tested environments to assess whether AM can be implemented in seismic areas. Later, the paper discusses the potential benefits and vulnerabilities of AM in earthquake-prone regions, then concludes with an assessment of AM's application in earthquake-prone zones.

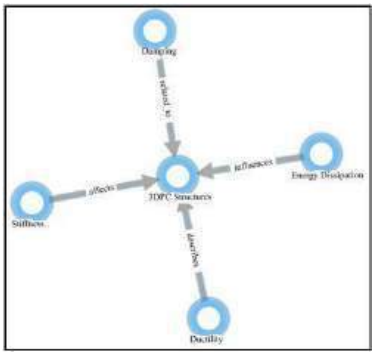


Figure 11: Illustration of the main earthquake-resistant parameters

4.1 Potential benefits

AM has been tested many times in laboratories, so based on this research, the paper identifies four main future benefits for 3D printing implementation. The benefits are:

- Optimized geometry of a construction for better lateral strength and energy dissipation
- Reduced mass for lower inertial force on a building
- Topology optimization to increase resilience
- Potential high recovery speed after an earthquake

Firstly, for earthquake-prone regions, the shape and topology of the building play a vital role, because they directly affect the structure’s ability to withstand physical loading(Hernández Vargas et al.). This is supported by Berwal et al., whose review of optimization case studies reports that, under nonlinear cyclic loading, shape-optimized walls showed up to a 25% increment in lateral strength and an 18% increment in energy absorption compared to conventional rectangular sections. In addition, the geometric tailoring of the walls improved the crack propagation pattern and reduced stiffness degradation. The same review describes a case study of curvilinear shell elements for 3D printing that were shape-optimized so that their natural frequency avoids the dominant frequencies of earthquake ground motion; in simulations, this

produced a roughly 15% higher fundamental frequency and a more uniform stress distribution (Berwal et al.).

Furthermore, AM-printed structures weigh less, which is crucial for seismic zones. Based on the formula F=ma, the decrease in mass significantly impacts the overall force that is generated. A case study summarized by Berwal et al. found that lattice-structured columns made about 30% lighter dissipated substantially more energy than solid columns in pseudo-dynamic simulations, which suggests that the reduced weight can improve the damping abilities of a structure (Berwal et al.).

One key factor is the topology of a structure. Alexandros Chortis et al. have found that the smaller the ratio of height to length, the better the resilience; the most slender wall fails significantly earlier due to high flexural demand(Chortis et al.). Hernández Vargas et al. tested three unreinforced 3DCP beams under three-point bending and found that beams with internally optimized material distribution reached a strength-to-weight ratio that was on average 47% and 63% higher than that of a conventionally printed beam (Hernández Vargas et al.). This suggests that internal optimization could make printed members more efficient. However, these were static tests on unreinforced beams, so their effect on seismic performance remains to be demonstrated.

Lastly, minimizing the effect of an earthquake's vibration is crucial, but for earthquake-prone regions it is also important to increase the recovery speed of buildings after the disaster — that is, how fast a building can be reconstructed. In this aspect, AM technology provides the best results. Firstly, during the emergency, mobile on-site printing systems can produce load-bearing structures quickly. For example, Liang et al. printed a 6 m polyurethane foam bridge on site in 20 minutes, using drone-based scanning to model the site (Liang et al.). What is more, 3D printing technology offers fast post-disaster management. Structures can be constructed within several days or even hours. That is crucial because it allows for faster recovery from disaster and provides housing for people affected by an earthquake(Iyoha et al.).

Ultimately, it is important to state that most of the previously discussed experiments used computer modelling or small-scale models.

4.2 Vulnerabilities

Although there are several promising advantages of AM, it is important to consider the drawbacks of this technology in seismic areas. This is important to analyze since it will directly illustrate the extent to which AM can be used in seismic regions and where it cannot. The main drawback of AM in seismic zones is its low interlayer bonding strength and structural instability. These vulnerabilities of AM are interrelated and each depends on the previous one.

Although this paper has already stated the benefit of geometry freedom of AM, there is still much room for development, because this technology is still worse in some aspects than traditional manufacturing. Reducing mass lowers inertial force, but if the structural members lack ductility and continuous tensile capacity, even reduced forces will cause brittle shear failure along the print layers. This means that even though the mass of AM-constructed structures is

lowered, if the main materials and boundaries are inefficient, the wall would buckle or fail under large forces. The first and main drawback is directly caused by the AM technique itself. Since AM is a layer-by-layer deposition construction method, the different layers are built in different time periods. Thus, the layer that has already been deposited has already started losing surface moisture and stiffening, and the wall that has been created is not monolithic. Many papers about interlayer bonding weakness were discussed previously, in other sections. Overall, it can be concluded that the structural integrity and resilience of a building decline as the time period between layering increases(Zhang et al.; Geng et al.; Joh et al.). Results of another experiment showed clear degradation in mechanical properties with increasing interlayer time, particularly perpendicular to the layers: flexural strength dropped over 25% and direct tensile strength up to 40% at a 25-minute gap(Maroszek et al.). This illustrates that 3D printed buildings are not ready for seismically active regions, since houses in these areas should withstand high forces.

Because AM walls by themselves are not as stable as traditionally constructed walls, there is another problem. Traditional construction has a built-in step where reinforcement fits naturally: formwork is built, the rebar cage is placed inside it, and concrete is then poured around it. 3D printing eliminates the formwork step entirely. The printed wall is the formwork, or the final wall, so there is no natural moment in the process where a cage can be inserted before the material sets. Every reinforcement strategy for 3DCP is essentially a workaround adapted onto a process that was not originally designed to need one (Kloft et al.). Printing the wall as a hollow shell and inserting a rebar cage afterward is currently the most practical reinforcement method. However, it lowers the practicality of AM since it requires the involvement of humans(Nan et al.). This system may benefit AM in the future if more automated methods are introduced. Another method of reinforcement that does not include labor is called U-nail. In this process, rebars are inserted during the AM process. Rebars inserted in the printing direction act as mechanical dowels across the weak interlayer boundaries. This bridging and dowel action drastically increases cross-layer shear and tensile resistance (often by over 140–220%), preventing horizontal sliding between printed layers during cyclic seismic forces(Wang et al.).

Nevertheless, integrating cast-in-place frame reinforcement significantly enhances the lateral load capacity and energy dissipation of 3D-printed concrete walls compared to unreinforced printed elements. However, experimental evaluations show that these hybrid systems remain structurally inferior to traditional monolithic reinforced concrete. Mechanical discontinuities at the wall-to-frame interface lower overall ductility and trigger premature joint cracking under cyclic loading, preventing the seamless energy dissipation typical of conventional reinforced concrete(RC) structures(Fu, Zhang, and Wu). Based on this evidence, it can be stated that reinforcement methods for AM remain underdeveloped; each type has drawbacks.

4.3 Suitability of AM in Seismic Environments

This chapter of the paper provides a critical synthesis of the extent to which AM can be implemented in seismic zones. In the last two parts, the paper discussed the benefits and vulnerabilities of AM in seismic zones. Based on the five previous categories: ductility, stiffness,

damping, energy dissipation, and reconstruction speed, the effectiveness of AM in earthquake zones can be evaluated.

To begin with, the answer to whether AM is ready for seismic zones is not simple; there are some points for and against its implementation. Firstly, there are several benefits of AM for seismic-resilient structures. These benefits are the topology optimization, tunable dynamic response of buildings, mass reduction, slenderness control, and fast recovery. The experiments, which illustrate the benefits of AM, were conducted with one-block systems, which contain only the wall or other part tested. This may overstate the benefits. Most research did not test benefits in a real-world environment. On the other hand, vulnerabilities of AM in seismic zones were mostly demonstrated using real large-scale models of buildings. The main drawbacks of AM are: interlayer bonding weakness, underdeveloped reinforcement methods, and a lack of real-world validation. Based on these findings, it can be seen that AM benefits should be tested in a real-world environment for more accurate data.

Each benefit and vulnerability depends on another. Overall, it can be seen that most of the drawbacks overlap with the advantages. Firstly, even though AM reduces the mass of a building, the interlayer bonding is not resilient enough to make that reduction meaningful. Secondly, the same applies to topology optimization. In the real world, the topology, slenderness control, and responsiveness benefits are not as critical when the reinforcements are underdeveloped and lack the ability to withstand high forces. The only advantage not offset by any disadvantage is the quick post-earthquake recovery.

Based on this evidence, at least for now, this paper states that the primary viable application of AM in seismic zones is supporting the construction of buildings and other infrastructure. Currently, the main drawbacks of AM are still too significant. Using AM for permanent construction in seismic zones is risky. Thus, it is not yet viable for construction in seismic zones. Nevertheless, AM can be used to reconstruct buildings quickly. The speed is the main advantage of AM over traditional manufacturing. Future researchers ought to focus more on this aspect of AM and, in addition, try to find techniques that could be used to increase the resilience of 3D printed buildings.

5. Conclusion

Structural drawbacks currently outweigh advantages for load-bearing members; standalone 3D concrete printing is not yet viable for primary seismic-resistant structures, leaving rapid post-disaster shelter and infrastructure recovery as its most immediate, practical application.

This paper conducted a comprehensive review evaluating the structural, economic, and environmental viability of additive manufacturing in civil engineering, with a specific focus on its performance, limitations, and safety within earthquake-prone regions. The research systematically investigated the foundational materials, printing techniques, operational trade-offs, and dynamic response characteristics of 3D concrete printing to determine whether

layer-by-layer extrusion can meet modern structural demands or if current technical boundaries necessitate alternative approaches.

This paper examined the historical background and material fundamentals, detailing Joseph Pegna's 1997 shift toward Contour Crafting, categorizing primary constituent materials into binders, aggregates, fibers, and chemical admixtures, and comparing four key printing techniques including Contour Crafting, Selective Laser Sintering, Fused Deposition Modelling, and Binder Jetting. From this evaluation, the combination of Contour Crafting and Ordinary Portland Cement was identified as the most practical and widely adopted method for mass-scale construction.

Section 3 evaluated the broader techno-economic benefits and drawbacks of additive manufacturing relative to conventional construction methods. It weighed five major advantages—environmentally friendly construction, geometric design freedom, lower construction costs, high productivity, and social site safety benefits—against four critical drawbacks: labor job displacement, building services (MEP) integration difficulties, equipment and cybersecurity risks, and interlayer bonding weaknesses causing structural instability.

Section 4 focused specifically on the implementation of printed structures in seismic zones, evaluating performance across five key criteria: ductility, damping, stiffness, energy dissipation, and post-earthquake recovery speed. It contrasted theoretical laboratory benefits like topology optimization, slenderness control, and mass reduction (F=ma) with severe real-world vulnerabilities such as anisotropic interlayer cold joints, brittle shear failure, and reinforcement integration friction. Thus, AM-printed structures are not yet strong enough for seismic zones.

In the future, engineers should use AM where it offers clear benefits, such as speed, reduced waste, and design freedom. However, this review shows that standalone 3D concrete printing is not yet suitable for primary load-bearing structures in seismically active zones, because of weak interlayer bonding, underdeveloped reinforcement methods, and the lack of full-scale validation. Future research should therefore focus on full-scale seismic testing of reinforced printed structures and on evaluating AM for post-earthquake restoration, which is currently its most practical application in these regions.

Works Cited

Abdalla, H., Fattah, K. P., Abdallah, M., Tamimi, A. K. Environmental Footprint and Economics of a Full-Scale 3D-Printed House. Sustainability, vol. 13, no. 21, 2021, 11978. https://doi.org/10.3390/su132111978.

ASME. The 3D Printed Office of the Future. American Society of Mechanical Engineers, https://cdn.asme.org/topics-resources/content/3d-printed-office-the-future.

Banihashemi, S., Akbarnezhad, A., Sheikhkhoshkar, M., Bril El Haouzi, H., Rolfe, B. 3D Printing in Construction: Sustainable Technology for Building Industry. Prog. Addit. Manuf., vol. 10, 2025, 11729–11762. https://doi.org/10.1007/s40964-025-01314-y.

Bechthold, M., Weaver, J. C. Materials Science and Architecture. Nat. Rev. Mater., vol. 2, no. 12, 2017, 17082. https://doi.org/10.1038/natrevmats.2017.82.

Berwal, P., Alamiery, A. A. H., Smerat, A., Abed, M. H., Majdi, A. Seismic Performance of 3D-Printed Concrete Structures: A Comprehensive Review of Material Behaviour, Geometry, and Optimization. Int. J. Civ. Eng., vol. 13, no. 1, 2026, 76–92. https://doi.org/10.14445/23488352/IJCEV13I1P107.

Besklubova, S., Raza, M. H., Zhong, R., Skibniewski, M. 3D Printing vs. Traditional Construction: Cost Comparisons from Design to Waste Disposal Stages. 2024. https://doi.org/10.22260/ISARC2024/0161.

Campbell, T., Williams, C., Ivanova, O., Garrett, B. Could 3D Printing Change the World? Technologies, Potential, and Implications of Additive Manufacturing. Strategic Foresight Initiative Report, Atlantic Council, 2011. https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/could-3d-printing-change-the-world/

Chortis, A., Gkountas, C., Melidis, L., Katakalos, K. Seismic Performance Evaluation of 3D-Printed Concrete Walls Through Numerical Methods. Buildings, vol. 15, no. 17, 2025, 3205. https://doi.org/10.3390/buildings15173205.

CyBe Construction. “CyBe Mortar.” https://cybe.eu/3d-concrete-printing/cybe-mortar/eco-friendly-material/.

Delgado Camacho, D., Clayton, P., O’Brien, W. J., Seepersad, C., Juenger, M., Ferron, R., Salamone, S. Applications of Additive Manufacturing in the Construction Industry—A Forward-Looking Review. Autom. Constr., vol. 89, 2018, 110–119. https://doi.org/10.1016/j.autcon.2017.12.031.

El-Sayegh, S., Romdhane, L., Manjikian, S. A Critical Review of 3D Printing in Construction: Benefits, Challenges, and Risks. Arch. Civ. Mech. Eng., vol. 20, no. 2, 2020, 34. https://doi.org/10.1007/s43452-020-00038-w.

Feng, L. Study on the Status Quo and Problems of 3D Printed Buildings in China. Global J. Hum.-Soc. Sci., vol. 14, no. H5, 2014, 1–4. https://socialscienceresearch.org/index.php/GJHSS/article/view/1317

Fu, Y., Zhang, Y., Wu, L. Experimental Study on Structural Performance of Cast-in-Place Frame Printed Concrete Wall. Int. J. Concr. Struct. Mater., vol. 18, no. 1, 2024, 77–91. https://doi.org/10.1186/s40069-023-00620-5.

Gangotra, A., Del Gado, E., Lewis, J. I. 3D Printing Has Untapped Potential for Climate Mitigation in the Cement Sector. Commun. Eng., vol. 2, 2023, 6. https://doi.org/10.1038/s44172-023-00054-7.

Geng, Z., et al. Layer-Interface Properties in 3D Printed Concrete: Dual Hierarchical Structure and Micromechanical Characterization. Cem. Concr. Res., vol. 138, 2020, 106220. https://doi.org/10.1016/j.cemconres.2020.106220.

Ghaffar, S. H., Corker, J., Fan, M. Additive Manufacturing Technology and Its Implementation in Construction as an Eco-Innovative Solution. Autom. Constr., vol. 93, 2018, 1–11. https://doi.org/10.1016/j.autcon.2018.05.005.

Gueche, Y. A., Sanchez-Ballester, N. M., Cailleaux, S., Bataille, B., Soulairol, I. Selective Laser Sintering (SLS), a New Chapter in the Production of Solid Oral Forms (SOFs) by 3D Printing. Pharmaceutics, vol. 13, no. 8, 2021, 1212. https://doi.org/10.3390/pharmaceutics13081212.

Gupta, N., Tiwari, A., Bukkapatnam, S. T. S., Karri, R. Additive Manufacturing Cyber-Physical System: Supply Chain Cybersecurity and Risks. IEEE Access, vol. 8, 2020, 47322–47333. https://doi.org/10.1109/ACCESS.2020.2978815.

Hager, I., Golonka, A., Putanowicz, R. 3D Printing of Buildings and Building Components as the Future of Sustainable Construction? Procedia Eng., vol. 151, 2016, 292–299. https://doi.org/10.1016/j.proeng.2016.07.357.

Heidelberg Materials. “3D-Printed Earthquake Resistance: First Earthquake-Resistant Building in Italy Made with 3D Concrete Printing.” https://www.heidelbergmaterials.com/en/sustainability/we-decarbonize-the-construction-industry/3d-concrete-printing/earthquake-resistance.

Hernández Vargas, J., Sjölander, A., Westerlind, H., Silfwerbrand, J. Internal Topology Optimisation of 3D Printed Concrete Structures: A Method for Enhanced Performance and Material Efficiency. Virtual Phys. Prototyp. 2024, 19 (1). https://doi.org/10.1080/17452759.2024.2346290.

Hutyra, A., Maroszek, M., Rudziewicz, M., Góra, M., Tyliszczak, B. Review of Material Processing Technology for 3D Concrete Printing. Materials, vol. 19, 2026, 564. https://doi.org/10.3390/ma19030564.

International Labour Organization. Construction: A Hazardous Work. ILO, 23 Mar. 2015, https://www.ilo.org/resource/construction-hazardous-work.

Iyoha, P., Ali, N., Mostafavi, S. Advancing Sustainable and Disaster-Resilient Housing through 3D Printing Technology: A Systematic Review. Sustain. Cities Soc.: Advances, vol. 2, no. 2, 2026, 100070. https://doi.org/10.1016/j.scsadv.2026.100070.

Joh, C., Lee, J., Bui, T. Q., Park, J., Yang, I.-H. Buildability and Mechanical Properties of 3D Printed Concrete. Materials, vol. 13, no. 21, 2020, 4919. https://doi.org/10.3390/ma13214919.

Kafle, A., Luis, E., Silwal, R. 3D/4D Printing of Polymers: Fused Deposition Modelling (FDM), Selective Laser Sintering (SLS), and Stereolithography (SLA). Polymers, vol. 13, no. 18, 2021, 3101. https://doi.org/10.3390/polym13183101.

Kloft, H., Empelmann, M., Hack, N., Herrmann, E., Lowke, D. Reinforcement Strategies for 3D-Concrete-Printing. Civ. Eng. Des., vol. 2, no. 4, 2020, 131–139. https://doi.org/10.1002/cend.202000022.

Koç, M., Khan, S. A., İlcan, H., Şahmaran, M. Conceptual Design of Autonomous Rapid Printing System for Emergency and Humanitarian Needs. Mater. Today Proc., vol. 70, 2022, 1–5.

Liang, J., Zeng, Z., Wang, P., Wang, Z., Wen, S., Zhou, Y., Zhou, C. Emergency Response Mobile 3D Printing System: A Case Study in Rapid Manufacturing of Polyurethane Foam Bridges. Case Stud. Constr. Mater. 2025. https://doi.org/10.1016/j.cscm.2025.e05051.

Maroszek, M., Hager, I., Mróz, K., Sitarz, M., Hebda, M. Anisotropy of Mechanical Properties of 3D-Printed Materials—Influence of Application Time of Subsequent Layers. Materials, vol. 18, no. 16, 2025, 3845. https://doi.org/10.3390/ma18163845.

Mohajan, H. K. Impacts of Construction and Demolition Waste on Environment: An Overview. Stud. Soc. Sci. Hum., vol. 4, no. 3, 2025, 1–5. https://www.paradigmpress.org/SSSH/article/view/1618

Mu, X., Zhang, X., He, G., Liew, K. M. Toward Low-Carbon 3D Concrete Printing through Circular Construction and Waste Valorization. J. Build. Eng., vol. 119, 2026, 115308. https://doi.org/10.1016/j.jobe.2026.115308.

Nadarajah, N. Development of Concrete 3D Printing. Master's Thesis, Aalto University, 2018. https://aaltodoc.aalto.fi/items/5bd5d1ab-5b2e-4f29-8560-2a36949e5d87.

Nan, B., Qiao, Y., Leng, J., Bai, Y. Advancing Structural Reinforcement in 3D-Printed Concrete: Current Methods, Challenges, and Innovations. Materials, vol. 18, no. 2, 2025, 252. https://doi.org/10.3390/ma18020252.

NBS. How Long Will Components and Systems Last? NBS, https://www.thenbs.com/knowledge/how-long-will-components-and-systems-last.

OECD. Due Diligence Essentials for Responsible Construction. OECD Publishing, 2026, https://www.oecd.org/en/publications/responsible-business-conduct-spotlights_03a75bf9-en/due-diligence-essentials-for-responsible-construction_6d82bd67-en.html.

Pegna, J. Exploratory Investigation of Solid Freeform Construction. Autom. Constr., vol. 5, no. 5, 1997, 427–437. https://doi.org/10.1016/S0926-5805(96)00166-5.

Perkins, I., Skitmore, M. Three-Dimensional Printing in the Construction Industry: A Review. Int. J. Constr. Manag., vol. 15, no. 1, 2015, 1–9. https://doi.org/10.1080/15623599.2015.1012136.

Placzek, G., Schwerdtner, P. A Global Snapshot of 3D-Printed Buildings: Uncovering Robotic-Oriented Fabrication Strategies. Buildings, vol. 14, 2024, 3410. https://doi.org/10.3390/buildings14113410.

Raphael, B., Senthilnathan, S., Patel, A., Bhat, S. A Review of Concrete 3D Printed Structural Members. Front. Built Environ., vol. 8, 2023, 1034020. https://doi.org/10.3389/fbuil.2022.1034020.

Rodgers, Y. van der M., Liu, X., Yi, J., Zhang, L. Women in the Construction Industry: Can Assistive Robotic Technologies Help to Close Employment and Pay Gaps? AEA Conference Paper, 2025, https://www.aeaweb.org/conference/2025/program/paper/KkHntFeh.

Robotiq. "Avoid Common Robotics Hazards by Following These 6 Rules." https://blog.robotiq.com/avoid-these-common-robotics-hazards-by-following-these-6-rules.

Rojas, J., Trout, D., Earnest, G. S., Hasanzadeh, S. Occupational Safety and Health Considerations in 3D-Printing Construction. J. Manag. Eng. 2026. https://doi.org/10.1061/JMENEA.MEENG-7383.

Tafesse, S., Girma, Y. E., Dessalegn, E. Analysis of the Socio-Economic and Environmental Impacts of Construction Waste and Management Practices. Heliyon, vol. 8, no. 3, 2022, e09169. https://doi.org/10.1016/j.heliyon.2022.e09169.

University of Florida Warrington College of Business. "Big Print Job: 356 3D-Printed Buildings Constructed Worldwide Since 2013." 2025, https://warrington.ufl.edu/news/big-print-job/.

Wang, L., Ma, G., Liu, T., Buswell, R., Li, Z. Interlayer Reinforcement of 3D Printed Concrete by the In-Process Deposition of U-Nails. Cem. Concr. Res., vol. 148, 2021, 106535. https://doi.org/10.1016/j.cemconres.2021.106535.

WinSun. "3D-Printed Construction Data and Company Information." CNET, https://www.cnet.com/culture/worlds-first-3d-printed-apartment-building-constructed-in-china/.

Wohlers, T., et al. History of Additive Manufacturing. Wohlers Associates, Inc., 2016. https://doi.org/10.2139/ssrn.4474824.

World Economic Forum. The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution. World Economic Forum, 2016.

Wu, P., Wang, J., Wang, X. A Critical Review of the Use of 3-D Printing in the Construction Industry. Autom. Constr., vol. 68, 2016, 21–31. https://doi.org/10.1016/j.autcon.2016.04.005.

Zhang, B., Goel, A., Ghalsasi, O., Anand, S. CAD-Based Design and Pre-Processing Tools for Additive Manufacturing. J. Manuf. Syst., vol. 52, 2019, 227–241. https://doi.org/10.1016/j.jmsy.2019.03.005.

Zhang, R., Liu, C., Duan, Z., Lv, Z., Zhang, W., Liu, H. Weak Interlayer Interfaces in 3D-Printed Concrete: Formation Mechanisms, Cross-Scale Consequences, and Control Strategies. Coatings, vol. 16, no. 6, 2026, 660. https://doi.org/10.3390/coatings16060660.

Zhou, L., Miller, J., Vezza, J., Mayster, M., Raffay, M., Justice, Q., Al Tamimi, Z., Hansotte, G., Sunkara, L. D., Bernat, J. Additive Manufacturing: A Comprehensive Review. Sensors, vol. 24, no. 9, 2024, 2668. https://doi.org/10.3390/s24092668.

Zhuang, Z., Xu, F., Ye, J., Hu, N., Jiang, L., Weng, Y. A Comprehensive Review of Sustainable Materials and Toolpath Optimization in 3D Concrete Printing. npj Mater. Sustain., vol. 2, 2024, 12. https://doi.org/10.1038/s44296-024-00017-9.

Ziaee, M., Crane, N. B. Binder Jetting: A Review of Process, Materials, and Methods. Addit. Manuf., vol. 28, 2019, 781–801. https://doi.org/10.1016/j.addma.2019.05.031.

Žujović, M., Obradović, R., Rakonjac, I., Milošević, J. 3D Printing Technologies in Architectural Design and Construction: A Systematic Literature Review. Buildings, vol. 12, no. 9, 2022, 1319. https://doi.org/10.3390/buildings12091319.

A Study of Wing Aspect Ratio for Aerodynamic Efficiency Across Different Mission Profiles By Vatsal Singhal

Abstract

Aircraft wing aspect ratio is an important design parameter for aerodynamic efficiency and fuel consumption. It is defined as the ratio of wingspan to average chord. It affects lift generation and drag. This relationship matters because even small geometric changes can affect fuel consumption and operational costs.

This study investigates how aspect ratio affects the aerodynamic efficiency of subsonic aircraft and how the optimum aspect ratio varies with mission profile. This study shows through simulation and literature analysis that the optimum aspect ratio depends on how a given mission balances aerodynamic efficiency against structural cost. This is important as aspect ratio also affects the operating costs of a flight. It is hypothesized that the L/D ratio increases as aspect ratio increases, since induced drag drops. However, aspect ratio cannot increase infinitely, as structural weight becomes a problem, so the best aspect ratio depends on the mission's requirements.

To test this, we use XFLR5 to model wings of varying aspect ratio (2–16) using the NACA 2412 airfoil, supported by secondary literature on mission-specific design trends.

Keywords

Aspect ratio, Vortex Lattice Method, Aerodynamic efficiency, Induced drag.

1. Introduction

A small change in the geometry of an aircraft's wing can influence how much fuel it burns over its journey, and so it is important to understand how changes affect aerodynamic efficiency and what changes should be implemented to maximize efficiency. The growing concerns about how greenhouse gases affect our environment, and the increasing prices of fuel, have led to more research being published on efficient aircraft. Since fuel burn is closely linked to aerodynamic drag, improving aerodynamic efficiency has become one of the main priorities in modern aircraft design.[7] For example, the Boeing 787-8 has an aspect ratio of 10, which is 15% greater than the earlier model 777-200 (AR 8.68).[6]

Aerodynamic efficiency is given by the lift-to-drag ratio (L/D):

$$L/D = \frac{C_l}{C_d} \quad (1)$$

where Cl is the lift coefficient and Cd is the drag coefficient. It is also said that the maximum ratio (L/Dmax) is given when the derivative of this equation equals zero.

One of the key design parameters that affect aerodynamic efficiency is the aspect ratio, given by the formula:

$$AR = \frac{b^2}{S} \quad (2)$$

An increase in the lift-to-drag ratio generally leads to a reduction in an aircraft's fuel consumption. This occurs because, during cruise conditions, thrust balances drag, and consequently lower drag minimizes the required thrust, which is generated by burning fuel. Chemical energy converts to heat, causing gases to expand and accelerate backward through a nozzle; by Newton's third law, this creates an equal and opposite forward reaction force.[1]

Aspect ratio directly influences the induced drag, and this is given by the equation:

$$C_{Di} = \frac{C_l^2}{\pi \cdot e \cdot AR} \quad (3)$$

Therefore, the induced drag coefficient CDi is dependent on the lift coefficient, aspect ratio, and Oswald's efficiency factor e.

Oswald's efficiency factor takes into account the shape of the lift distribution. It is observed clearly that the induced drag coefficient has an inverse relation with the AR, showing that an increase in AR would reduce the induced drag.[2] To implement these benefits, aircraft such as UAVs and propeller aircraft use high aspect ratios, as done by Goetzendorf-Grabowski et al.,[5] and it has been found that a high aspect ratio significantly increases the lift-to-drag ratio.[4]

Having said that, to reduce the induced drag, the AR cannot simply be maximized, as that increases structural costs and structural complexities due to an increase in bending moments and shear stresses; this is because a longer span shifts the centre of lift further away from the fuselage. Usually, an aircraft that prioritises minimum Direct Operating Costs (DOC) is expected to have a lower AR than an aircraft optimized for minimum fuel burn.[7,8]

This study investigates how wing aspect ratio influences the aerodynamic efficiency of subsonic aircraft, and examines how the optimum aerodynamic aspect ratio changes under different mission profiles, while maintaining constant wing area and airfoil geometry.

We hypothesise that increasing the aspect ratio will increase aerodynamic efficiency for an aircraft operating at an optimum angle of attack. This happens because there is a decrease in the induced drag, which results in an increase in the lift-to-drag ratio. However, the optimum aspect ratio varies across different mission profiles. This is because different missions require different balances between aerodynamic efficiency and structural considerations, meaning that an aspect ratio that is optimal for one application may not provide the best overall performance for another.

2. Aspect ratio to maximize aerodynamic efficiency

In this section, we try to answer the first question of our research and create a relationship between aerodynamic efficiency and aspect ratio.

a) Experimental Design

The objective of this experiment was to investigate aspect ratio as the only independent variable and quantify its effect on L/D while eliminating influence from airfoil geometry, planform shape and airflow. This isolation is essential in establishing a controlled cause-effect relationship between AR and aerodynamic efficiency.

The independent variable in our experiment was aspect ratio, which we varied across different values ranging from 2 to 16. This range was selected as it goes from low aspect ratio to high aspect ratio, equating to a range of aircraft types. Investigating a range of aspect ratios allows the resulting L/D trend to be evaluated not only for its general shape but also for identifying whether diminishing or adverse returns emerge at the extremes of the range, because when we combine equations (1) and (3), we notice that the maximum L/D ratio scales with $\sqrt{AR}$ instead of AR, hence giving us diminishing returns.

To ensure that any observed variation in aerodynamic efficiency could be attributed to aspect ratio alone, all other geometric and flow-related parameters were maintained constant across the full set of simulations.

Table 1 shows each parameter and its controlled condition.

Parameter	Controlled Condition
Airfoil section	NACA 2412
Wing planform area (S)	Held constant
Wing planform shape	Held constant
Freestream velocity	30 m/s
Reynolds number	Varies with chord
Angle of attack range	-5° to 15°
Analysis scope	Wing-only (isolated)

The airfoil chosen was NACA 2412 because it gives a stable balance of lift and drag coefficients, making the airfoil suitable for flight operations at low to high speeds. Moreover, this airfoil is commonly used in various general aviation aircraft because of its moderate 2% camber at 40% chord and a practical 12% thickness ratio; together, they produce stable and well-behaved flight conditions. Moreover, this configuration also does not lead to abrupt stalls.

The aerodynamic simulations were performed using software called XFLR5, which is an open-source design tool for low-speed subsonic aircraft at low Reynolds numbers, applying the Vortex Lattice Method (VLM). VLM is a simplified but effective way to model how a 3D wing generates lift and drag. By breaking the wing into a grid of vortex segments and solving for their strengths, it tends to be cheaper and faster than full computational fluid dynamics, while still capturing the key 3D aerodynamic behaviour.

It should be noted that VLM-based analysis assumes inviscid, incompressible, attached flow (so it is an ideal model with zero friction and constant density), and its accuracy becomes very low near stall angles of attack, where flow separation is high. Additionally, we simulate the wings as the only factor being tested, testing no other factors, so fuselage and tail drag are not captured, meaning the values of L/D are different from what they might be if a complete aircraft

were simulated in real-life scenarios; however, this limitation does not affect the validity of the conclusions drawn.

b) Methods

This section explains, in detail, how the simulations described in the experimental design were carried out in XFLR5, step by step, so that the process could be reproduced by someone else using the same software.

Before we could begin with the wing simulation, we had to create a 2D polar for the NACA 2412 aerofoil used in the simulation [Table 1]. For this, XFLR5 has an inbuilt module for direct foil analysis (Xfoil). This can be done using the NACA 4-digit input rather than importing coordinates. The polar was run at a Mach number of 0.088 and an air density of 1.225 kg/m³ across an angle of attack from -5° to 15°.

Each of the wings was designed using the Plane Design module in XFLR5, with the NACA 2412 aerofoil throughout the full span. The aspect ratio was varied from 2 to 16; span and chords were recalculated using equation (2), keeping wing area constant at around 1 m². The resulting dimensions for each configuration are shown in Table 2.

Table 2 shows how the span and the chord varied across different aspect ratios.

Aspect Ratio (AR)	**Span (m)**	**Chord (m)**	**Area (m²)**
2	1.41	0.71	1
4	2.00	0.50	1
6	2.45	0.41	1
8	2.83	0.35	1
10	3.16	0.32	1
12	3.46	0.29	1
14	3.74	0.27	1
16	4.00	0.25	1

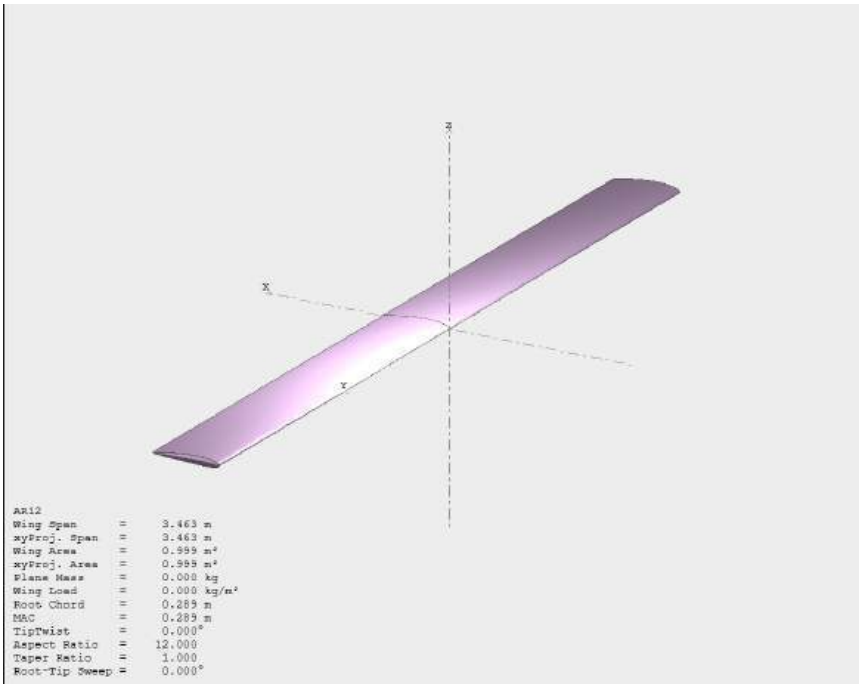


Fig 1 shows a wing with AR 12 modeled in XFLR5.

Once each of the wings was created, a Type 1 VLM analysis was run at a fixed velocity of 30 m/s. The angle of attack was swept from -5° to 15° in increments of 0.5°. It is to be noted that when the AoA was around 11.5–14° for wings with high aspect ratio, convergence was not achieved; therefore, these points have been excluded from the datasheet. This is likely because the local lift exceeded the range of the 2D polar near stall.

For each combination of angle of attack and aspect ratio, XFLR5 mainly generated two important values: the lift coefficient and the drag coefficient. Using equation (1), we then calculated the lift-to-drag ratio, the measure of aerodynamic efficiency. XFLR5 also generated other values which are not useful for our project. The data from XFLR5 were exported to a .txt file. This was done using the inbuilt export function within the software. This resulted in less time being consumed and reduced transcription error.

From this data, two sets of plots were created: L/D against angle of attack, and the drag polar (Cl against Cd). Only two graphs were created, and each of the graphs has different coloured lines which each correspond to a different aspect ratio; this allowed us to visualize the impact of aspect ratio on aerodynamic efficiency and how it varied across different angles of attack.

c) Results

This section presents the results obtained from the XFLR5 simulation as described in the methods section above. Results are presented in two parts: first, the variation of lift-to-drag ratio (L/D) with angle of attack for each aspect ratio configuration, and second, the drag polars (Cl vs Cd) generated for the same configurations.

Figure 2 shows the lift-to-drag ratio (y-axis) plotted against angle of attack (x-axis) for each of the aspect ratios (2 to 16), which are denoted by different colours. When we study the shape of the graph, we observe that L/D increases from a negative value and then peaks at a low positive angle of attack. The value then decreases gradually as the angle of attack increases. This is expected because drag continues to increase as the angle of attack increases while lift growth slows, so the L/D value decreases after the optimum point.

As the aspect ratio increased, the entire curve shifted upwards along the y-axis, and the peak L/D value increased substantially. The wing with an aspect ratio of 2 reached a maximum L/D of only 16.80, while the wing with an aspect ratio of 16 reached a maximum L/D ratio of 40.01, which is more than double despite the same planform area. Moreover, the angle of attack at which the peak L/D occurred shifted from 2.5° (for AR 2–10) to 3.0° (for AR 12) and 3.5° (for AR 14 and 16).

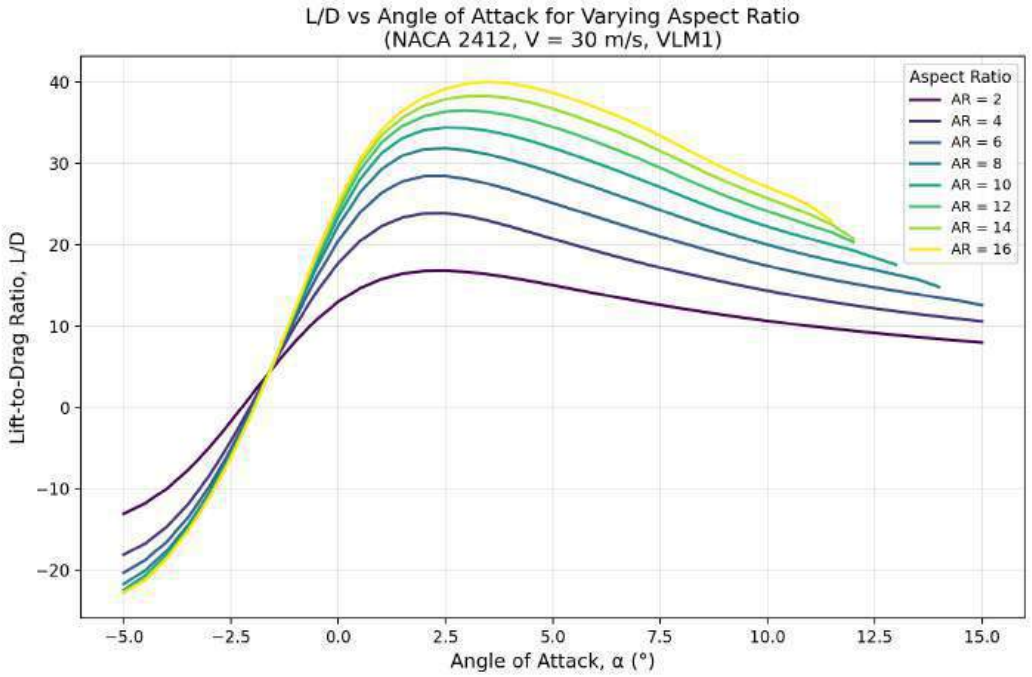


Fig 2 shows how L/D changes with angle of attack for all aspect ratios.

This trend can be explained using equation (3): increasing AR reduces induced drag for any given lift coefficient. This is why the AR 16 curve lies above the AR 2 curve on the y-axis. We can also explain the shift in the peak angle of attack by rearranging (3) for the condition Cd_0 = Cdi, which gives Cl_opt = $\sqrt{(\pi \cdot e \cdot AR \cdot Cd_0)}$. Since Cl increases roughly linearly with angle of attack, a higher Cl_opt (the lift coefficient for which L/D reaches its maximum value) at higher aspect ratios requires a higher angle of attack to reach it, explaining why the peak L/D angle shifted from 2.5° at low AR to 3.5° at AR 14 and 16.

Table 3. Maximum Lift-to-Drag ratio and corresponding angle of attack for each aspect ratio.

Aspect Ratio (AR)	Max L/D	AoA at Max L/D (°)
2	16.80	2.5
4	23.86	2.5
6	28.45	2.5
8	31.86	2.5
10	34.40	2.5
12	36.48	3.0
14	38.30	3.5
16	40.01	3.5

It is worth noting one limitation of this dataset: the angle of attack range simulated was not the same for all configurations. When simulating aspect ratios 2, 4 and 6, the range was from -5° to 15°, but wings with high aspect ratios diverged at the upper end of this range; wings with an aspect ratio of 16 only produced results up to 11.5°, after which the graph finishes. This is clearly visible in Fig 2, which shows that high aspect ratios (typically 8–16) all decline near the end, giving no results at the upper end of our range.

Fig 3 shows the drag polar (CL vs CD) for each of the aspect ratio configurations. We noticed that when the aspect ratio increased, the curves shifted towards the left, meaning that for a given lift coefficient a wing with a high aspect ratio tends to produce less drag than a wing with a low aspect ratio would. Since most of the conditions, like aerofoil selection and wing planform shape, were kept constant, it is clear this leftward shift is mainly due to reduced induced drag as the aspect ratio increased rather than viscous drag changes.

We can quantify our data to observe trends: at a fixed lift coefficient of 0.5, it is observed that induced drag accounted for 86.8% of the total drag for aspect ratio 2; however, this value dropped significantly when aspect ratio 16 was analysed, as induced drag accounted for only 42.5% of the total drag (Fig 4).

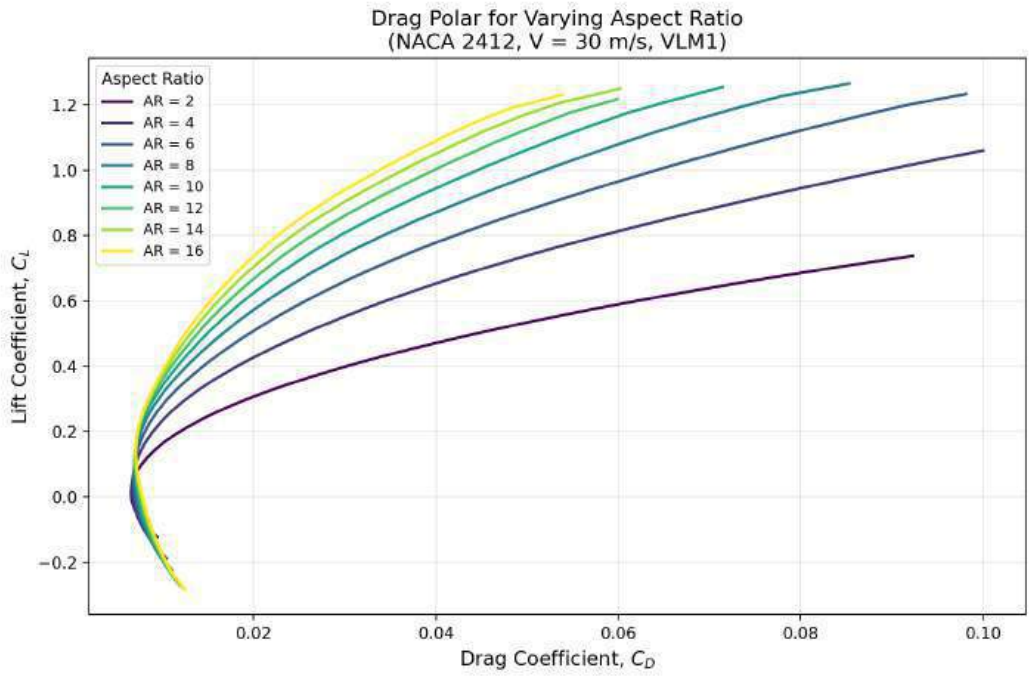


Fig 3 shows the drag polar (lift coefficient against drag coefficient for all aspect ratios).

The gap between the lines of adjacent aspect ratios is not constant: the difference between AR 2 and AR 4 is much greater than it is between AR 12 and AR 14. This tells us that the benefit of increasing aspect ratio, and the change in induced drag, decrease significantly at high aspect ratios, even though they never vanish completely.

A second mechanism behind these diminishing returns, alongside the √AR scaling, is the Reynolds number. Since the wing area was held constant, the chord decreased as the aspect ratio increased, so the chord-based Reynolds number at 30 m/s fell from about 1.4×10^6 at AR 2 to about 5.1×10^5 at AR 16. A lower Reynolds number generally increases the skin friction and profile drag of the airfoil, so the narrower high-AR wings pay a larger viscous drag penalty, which partly offsets their reduction in induced drag and further reduces the gain in L/D at high aspect ratios. If the 2D polar is generated at a single Reynolds number, this effect is not captured by the simulation, so the real diminishing returns are likely to be slightly stronger than those shown in Fig 5.

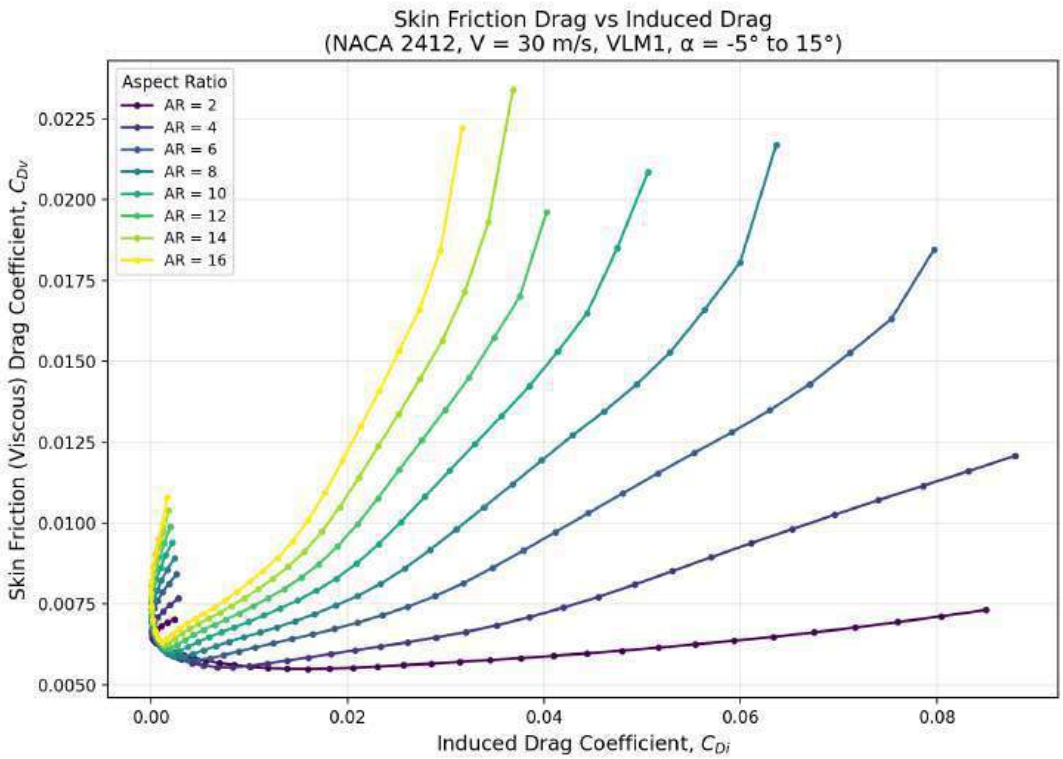


Fig 4 shows viscous drag vs induced drag.

This confirms that the improvement in aerodynamic efficiency at higher aspect ratios is driven almost entirely by the reduction in induced drag, exactly as predicted by equation (3), rather than by any change in the viscous drag characteristics of the airfoil itself.

Fig 5 shows the aspect ratio against maximum L/D. This graph visualizes the values in Table 3. The relationship is clearly increasing but is not linear, as the gradient of the line is not constant throughout but rather decreases as the aspect ratio increases. This means the rate of improvement slowly decreases as the aspect ratio increases: between aspect ratios 2 and 4, the maximum L/D increased by 7.06, but between aspect ratios 14 and 16, the increase was only 1.71 despite the same increment in aspect ratio. This suggests that while a higher aspect ratio increases the maximum L/D, the increase tends to diminish. This trend becomes relevant when weighed against the structural weight penalties of high-AR wings discussed in the next section.

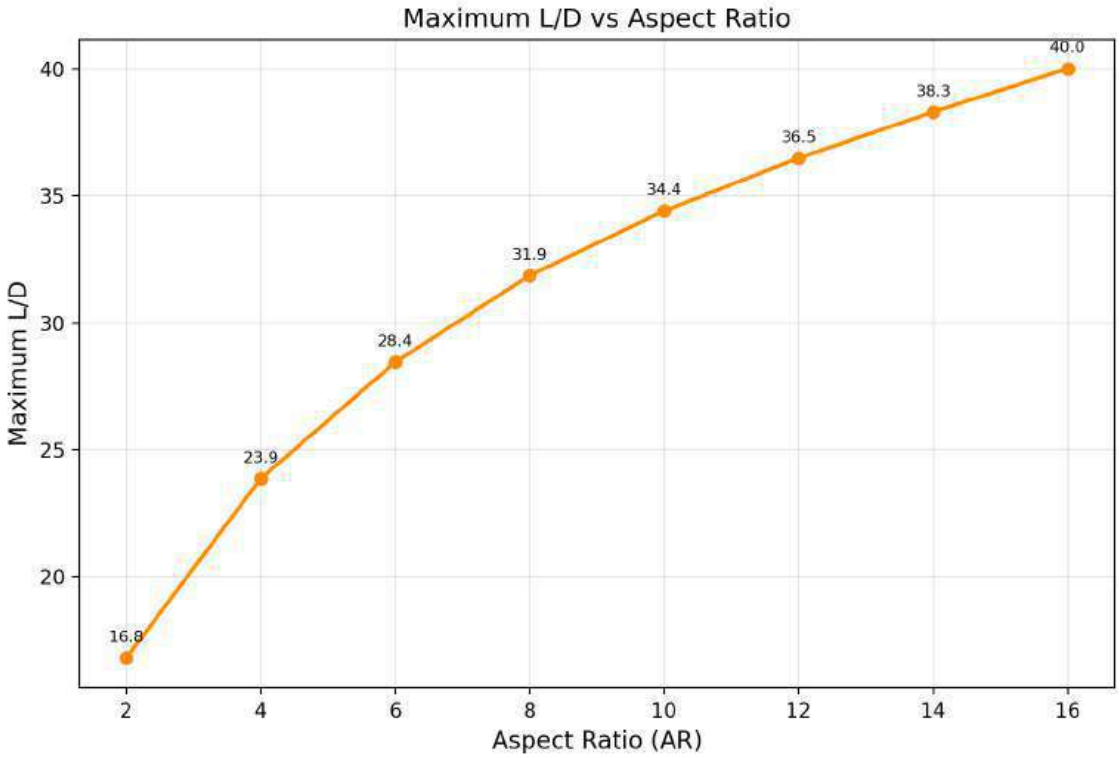


Fig 5 shows aspect ratio against maximum lift-to-drag ratio.

In summary, increasing the aspect ratio across all the configurations increased the lift-to-drag ratio, proving our hypothesis to be correct. This improvement is due to reduced induced drag at high values of aspect ratio, as visualised in Fig 4. However, the rate of improvement diminishes at high aspect ratios, indicating that beyond a certain point, further aspect ratio increases yield comparatively small aerodynamic changes. This factor must be weighed against the structural costs of high-AR wings when identifying an optimum aspect ratio for a given application.

3. Optimum aspect ratio based on mission profile

In this section, we talk about how the optimum aspect ratio varies across different subsonic mission profiles. This section mainly talks about four types of mission profiles:

- Glider wing
- Fixed wing UAV
- Very high altitude aircraft wing (e.g. solar HALE UAVs)
- Trainer aircraft wing

This section has been further divided into four subsections, each talking about a specific mission profile.

a) Experimental design

In Section 2 of the paper, we investigated the relationship between aspect ratio and aerodynamic efficiency using simulation software called XFLR5. While this method was good for isolating aspect ratio and quantifying its impact on aerodynamic efficiency, in reality aspect ratio selection considers many factors outside of aerodynamic efficiency alone: structural weight, manufacturing cost, manoeuvrability, mission duration, and operational environment all play a role.[12] In this section, we shift from our previous approach of simulations to a secondary data analysis; we analyse published research to understand how the optimum aspect ratio changes based on the mission profile.

We selected the four mission profiles mentioned above; these were chosen so that a wide spectrum of data could be analysed, because glider wings and high-altitude aircraft focus on minimizing drag and maximizing endurance,[13] whereas trainer aircraft prioritize forgiving and stable flight, rather than aerodynamic efficiency, since they are designed to help student pilots learn to fly. Fixed-wing UAVs sit in between, and actually span a wide range of aspect ratios depending on the specific sub-role of the UAV in question, from small short-range platforms to long-endurance surveillance aircraft.[17,18]

For each mission profile, this study gathered aspect ratio values. Priority was given wherever possible to sources that did not just report a numerical aspect ratio, but explicitly explained why that value was chosen for that particular application, since the goal of this section is not simply to catalogue AR values but to connect them back to the aerodynamic theory established in Section 2.

It is worth noting that, unlike Section 2, this analysis does not hold variables like wing area, airfoil section, or Reynolds number constant across the comparison, since these vary between real aircraft of different sizes. This is a limitation of using secondary, real-world data rather than a simulation.

b) Methods

This section is further divided into four subsections, which are different mission profiles; we talk about how each of these requires a different optimum aspect ratio.

I) Gliders

Gliders are one of the most relevant examples that we have for our study, as they push aspect ratio to its limit. Gliders have no engine, and so they generate no thrust and rely only on minimizing drag to sustain flight. Because of this, gliders provide some of the most relevant literature for our study.

Zhang (2024) investigated the relationship between aerodynamic efficiency and aspect ratio in gliders; this is closely related to our study in Section 2 but is narrowed down only to the concept of gliders.[15] Another recent study, “Impact of Wing Aspect Ratio on Aerodynamic Performance in Gliding Flight” (2024),[10] suggests that an increase in aspect ratio results in a reduction in the glide angle, generally because of the equation:

$$tan(\gamma) = \frac{D}{L} (4)$$

Here, γ is the glide angle; a reduction in the glide angle results in a low D/L ratio, which therefore results in a high L/D ratio, the general measure of aerodynamic efficiency.

The same paper also states that the relationships are more complex than a simple linear trend of high aspect ratio.

To see how high the aspect ratio is pushed in real competition gliders, we identify two sailplanes, the Eta and the Concordia, which have wingspans of about 30 m; this was deliberately done to ensure maximum aerodynamic efficiency within structural limits. The Eta project has a wingspan of about 30 m and an aspect ratio of 51.33, and the Concordia design has a wingspan of about 28 m and an aspect ratio of 57. Both of these designs exceed the range which we tested (2–16).

II) Fixed-wing UAVs

Fixed-wing unmanned aerial vehicles are one of the most promising and futuristic types of subsonic aircraft. UAV is a broad term that covers many different sorts of aircraft, each with different roles and endurance requirements. A high aspect ratio was deliberately chosen for a medium altitude long endurance unmanned aerial vehicle (MALE UAV) specifically to reduce drag and extend the duration of the flight, consequently increasing endurance.[17] A recent study used CFD to study the impact of aspect ratio on aerodynamic performance in a fixed-wing UAV and to examine the trade-offs between aerodynamic benefit and structural costs.[25] This structural cost is demonstrated in an aeroelastic evaluation of a flexible high-aspect-ratio wing UAV, which found that a UAV wing with an aspect ratio of 13 underwent large structural deformation even under low subsonic flight conditions.[26]

However, as discussed earlier, different sorts of UAVs have different roles and endurance requirements. On the lower end of the UAV category lies a small quadrotor biplane VTOL UAV, which only had an aspect ratio of around 5–8 because it prioritized maneuverability rather than only focusing on aerodynamic efficiency.[18]

III) High-altitude aircraft

This section mainly talks about high-altitude long-endurance (HALE) aircraft, since sustainable flight routines shift designers to favour high aspect ratio wings. Olivier Montagnier and Laurent Bovet developed a straightforward mathematical relationship between lift coefficient and aspect ratio during HALE UAV flight routines.[19] Another gradient-based optimization study from Brigham Young University (BYU) analyses a solar-regenerative high-altitude long-endurance (SR-HALE) flying-wing aircraft designed to fly at an altitude of 18 km above sea level; the aircraft has an aspect ratio of 54.52, a surface area of 73.56 m² and a mass of 349.5 kg. This can be compared to an already operational aircraft, the EAV-3 solar-powered HALE UAV designed by the Korea Aerospace Research Institute, which has a 19.5 m wingspan and an aspect ratio of 17.4.[20,21]

The structural cost of maintaining such a high aspect ratio is discussed in a lot of aeroelastic literature. A study from the CEAS Aeronautical Journal talks about this very clearly; it states that wings with a high aspect ratio do reduce induced drag and improve aerodynamic efficiency, but they also increase weight and flexibility, thus giving rise to stabilization challenges.[22] Another CEAS Aeronautical Journal study examines active flutter control for high aspect ratio wings. Flutter is an aeroelastic phenomenon in which the wing's structural vibrations interact with the surrounding airflow; a high aspect ratio means longer and more slender wings, so they twist more under load, and in response to this a wing requires a heavier structure as the aspect ratio increases. This helps explain the gap between the 17.4 aspect ratio of the EAV-3 and the 54.52 aspect ratio of BYU's theoretical study.[23]

IV) Trainer Aircraft

To cover trainer aircraft, we went for a rather simple and different approach, since studies based on trainer aircraft are almost negligible compared to the three other categories discussed above. To analyse trainer aircraft, we analyse general aviation trainer aircraft and compare them to pre-existing literature on low aspect ratio wings, since most trainer aircraft have a low aspect ratio.

The Piper J-3 Cub has an aspect ratio of 6.96, the Cessna 172 has an aspect ratio of 7.32, the Beechcraft Bonanza has an aspect ratio of approximately 6.20 and the Piper Cherokee has an aspect ratio of approximately 5.6. All these values are calculated with equation (2), using wingspan and wing area, which are widely available online. We notice that most of the aspect ratios are in a low-to-moderate range. To explain this, we drew on a wind tunnel study from the Journal of Mechanics, which studies low aspect ratio wings.[24] It was found that wings with a low aspect ratio typically have higher induced drag and lower aerodynamic efficiency but a delayed stall; low aspect ratio wings reach a stall angle of about 20 degrees, indicating that a low aspect ratio increases the margin before reaching the stall angle. This directly impacts maneuverability and increases forgiveness, a factor that is very important as trainer aircraft are meant for pilots in training and require a high level of forgiveness.

c) Results

Table 4 covers all mission profiles and their examples, aspect ratios, primary design drivers and structural trade-offs for all examples discussed in the methods.

Aircraft type	Example Aircraft	Aspect Ratio	Primary Design Driver	Structural Trade-off Noted
Glider	Eta	51.33	Maximise L/D for endurance soaring	Wingspan near structural material limits
Glider	Concordia	57	Maximise L/D, minimum drag	Wing loading constrained to control weight

Aircraft type	Example Aircraft	Aspect Ratio	Primary Design Driver	Structural Trade-off Noted
Fixed-wing UAV (small)	Quadrotor biplane UAV	5–8	Manoeuvrability, compact size	Minimal — low structural risk
Fixed-wing UAV (MALE/HALE)	MALE ISR UAV	up to 25	Long endurance, minimum drag	Increased bending moment, large deformation at AR 13
High altitude aircraft	EAV-3 solar HALE UAV	17.4	Sustained lift, low air density	Real-world AR limited by flutter/weight
High altitude aircraft (theoretical)	BYU SR-HALE flying wing	54.52	Year-round stratospheric endurance	Idealised optimum, before full structural constraint
Trainer aircraft	Cessna 172 / Piper Cub / Bonanza / Cherokee	5.6–7.3	Manoeuvrability, stall tolerance	Minimal — low aspect ratio avoids structural penalty

I) Glider wing

The glider wing has one of the largest values of aspect ratio within our study. Table 4 covers this by highlighting the aspect ratio of the Eta and the Concordia, both of the aspect ratios of these wings exceed the limit we tested by almost three times. This was confirmed by Zhang (2024),[15] who found the L/D to be increasing as the aspect ratio was increased, and another study on the aerodynamic characteristics of gliders by Liang (2024)[14] also confirmed this using another methodology.

Importantly, the sailplane wing sectioning study (2023)[16] makes clear that this extreme aspect ratio is not achieved without structural compromise. Eta and Concordia are built almost exclusively from carbon fibre and glass-reinforced composites, since a metallic structure could not achieve the required stiffness-to-weight ratio at these spans, and their wings must be sectioned into multiple detachable parts purely to remain transportable given their extreme length. This clearly shows us a real-world structural trade-off; the main reason gliders can achieve such a high aspect ratio is that they are unpowered and do not carry as much load as a powered aircraft would. This means the structural weight and bending moments are significant enough to require composite materials, but are much more manageable than in a heavier, engine-carrying aircraft of the same span.

II) Fixed-wing UAVs

The fixed-wing UAV results show a wide spread driven by different sub-mission profiles: small UAVs have a low aspect ratio of 5–8, whereas long-endurance UAVs push aspect ratio up to 25, as mentioned by Adeleke et al. (2022) for a medium altitude long endurance (MALE) UAV wing designed specifically to reduce drag and extend flight duration.[17]

This upper end of an aspect ratio of 25 comes with structural penalties. The 2025 CFD/FEA study on aspect ratio's impact on UAV structural integrity found that aerodynamic performance improves with aspect ratio, but bending stress and structural weight also increase; this was not just a theoretical relationship but a directly quantified one.[25] This is supported by the aeroelastic evaluation of a flexible high-aspect-ratio wing UAV, which found that the wing with aspect ratio 13 was deliberately made to undergo large structural deformation; this shows that structural flexibility becomes an important factor well before the AR of 25 seen in the MALE UAV.[26] This shows that fixed-wing UAV designers do not select aspect ratio based on aerodynamic efficiency alone, but rather based on maximising aerodynamic efficiency while minimizing structural penalties.

III) High-altitude aircraft

HALE aircraft once again produce the largest and most extreme aspect ratios, but also clearly explain the gap between aerodynamic efficiency and a structurally realistic design. The Korea Aerospace Research Institute's EAV-3, a real solar-powered HALE UAV, uses an aspect ratio of 17.4, whereas the gradient-based optimization study by Brigham Young University, which modeled a solar-regenerative HALE flying wing intended to fly at an altitude of 18 km, had an aspect ratio of 54.52.[20,21]

A CEAS Aeronautical Journal study on multidisciplinary aeroelastic optimisation of high-aspect-ratio composite wings found that these wings do reduce induced drag, but they also increase both weight and structural flexibility. A related CEAS Aeronautical Journal study on active flutter control confirms this directly, stating that flutter constraints force a wing to be built stiffer and therefore heavier unless active flutter suppression systems are incorporated. Taking everything together, this implies that BYU's 54.52 aspect ratio wing is likely an aerodynamic efficiency optimized wing that does not fully take into consideration the flutter and structural weight constraints, indicating why a real flyable HALE plane such as the EAV-3 might have such a low aspect ratio. This is one of the clearest examples in this study's results of the theoretical aerodynamic optimum aspect ratio diverging substantially from the practically achievable one once structural trade-offs are properly accounted for.[22,23]

IV) Trainer Aircraft

In contrast to the previous three aircraft types, trainer aircraft clustered in a narrow, low to moderate aspect ratio range of AR 5.6 to AR 7.3. Unlike for gliders, UAVs and HALE aircraft, there is almost no established literature which talks about optimum aspect ratio and structural trade-offs for trainer aircraft. This is because trainer aircraft have a low aspect ratio, so there is not a major concern for flutter and bending moments, as there is for high aspect ratios of 17–54.

Instead, the concern is aerodynamic efficiency rather than structural penalties. Our simulations in Section 2 show that low aspect ratios produce higher induced drag and significantly decrease aerodynamic efficiency. This is also discussed in the Journal of Mechanics study, which found that low aspect ratio wings produce high drag but reach high stall angles; a

wing of AR 1.0 reached a stall angle of 20 degrees.[24] The trainer aircraft in our study have moderate aspect ratios rather than a very low value such as AR 1.0. However, the same principle still applies: a lower aspect ratio delays stall onset and improves the forgiving, predictable handling characteristics that flight training specifically requires, at the cost of higher induced drag and reduced cruise efficiency compared to a higher-AR wing of the same area.

Discussion

This study brings together two distinct investigations: an aerodynamic simulation followed by a literature review to study optimum aspect ratios across different mission profiles.

The XFLR5 simulations confirm the theoretical relationship between aspect ratio and induced drag, as they adhere to the formula stated in the introduction.

The maximum lift-to-drag ratio more than doubled when we increased the aspect ratio from 2 to 16, while the induced drag dropped significantly, aligning strongly with established aerodynamic theory.[2] Reviews of aircraft design already confirm that lift-induced drag can account for up to 80% of total aerodynamic drag during climb and around 40% during cruise, meaning that a reduction in induced drag through a higher aspect ratio can reduce fuel burn and consequently decrease the cost of operations.[11]

The study's own data in Section 2 show that the gains in lift-to-drag ratio are not linear; the increases in the L/D ratio decrease significantly at the high range of aspect ratio, giving diminishing returns. This follows from the maximum L/D scaling with $\sqrt{AR}$ and from the lower Reynolds number of the narrower high-AR wings, and it matters because of the structural costs that come with increasing aspect ratio. Published research shows that higher aspect ratio wings do increase aerodynamic efficiency, but also increase the overall mass, so even though fuel mass decreases, the aircraft's overall weight still increases.[28]

This structural trade-off was clearly visible in Section 3 of this study. Gliders and high-altitude aircraft, which prioritize aerodynamic efficiency, have a high aspect ratio, some even greater than 50. Transport aircraft prefer a moderate aspect ratio despite caring about fuel burn; this is generally because structural trade-offs become very significant in real-life operations. This is reflected in a study evaluating high aspect ratio composite wings for future Boeing 737 and A320-class aircraft, which found that while increasing wingspan reduces induced drag and improves efficiency, it also complicates airport gate compatibility and requires solutions such as folding wingtips to remain operationally practical.[9]

Our results support the view that aspect ratio depends on multiple factors and is not only an aerodynamic problem. This conclusion is present in many aerodynamic studies, which say that aerodynamic, aeroelastic, and aerostructural considerations should be considered very early in the design of a high aspect ratio wing.[11]

A limitation of this study worth noting is that the XFLR5 simulations in Section 2 modelled the wing in isolation, without a fuselage, tail, or structural model, meaning the results capture the aerodynamic side of this trade-off but not the structural cost directly. The mission profile comparison in Section 3 helps us identify the gap in real-world scenarios. A better future

study could combine an aerodynamic solver like XFLR5 with a structural weight model, allowing both sides to be analysed in a consistent manner.

Conclusion

The motive of this study was to answer how aspect ratio affects aerodynamic efficiency, and to what extent aerodynamic efficiency changes based on the mission profile.

The simulation results from XFLR5 in Section 2 confirm the first hypothesis. When the aspect ratio increased from 2 to 16, the maximum L/D ratio also increased from 17 to 40, while induced drag decreased drastically, in line with published research.[11]

However, the gain in L/D ratio is not linear; there are diminishing returns, with every increment in aspect ratio giving a smaller efficiency gain.

The comparison between different mission profiles in Section 3 also confirms our second hypothesis. Aspect ratio in real-world scenarios depends a lot on the aircraft type; this disparity is clearly visible in that competition gliders and HALE aircraft had a very high aspect ratio, while trainer aircraft had a very low aspect ratio. This variation in aspect ratio is mainly because of factors such as structural weight, manoeuvrability, and operational trade-offs, which vary across different aircraft types; this aligns well with current aeroelastic and structural design studies.[23,26,27]

Both of these arguments work hand in hand, supporting the central argument of this paper: aspect ratio is one of the most crucial factors that we can change to alter the performance of an aircraft. The aerodynamically ideal and optimal aspect ratio is not one singular value but, rather, multiple values which are affected by structural constraints, operating costs, etc. Future work related to this study could perform aerodynamic and structural analysis together rather than treating them as two different sections; this would allow the true optimum aspect ratio for a given mission to be analysed within one model.

Acknowledgements

I express my sincere gratitude to Dr. Devin Carroll, a professor at the University of Pennsylvania, for his help in reviewing the paper and providing valuable feedback; through his help, the paper turned out to be what it is now. I would also like to thank Mr. Kieran Tait, who is doing his PhD at the University of Bristol. His support and guidance helped me navigate through tough moments where I was confused or felt lost.

Authors

Vatsal Singhal is currently a grade 11 high school student at Sacred Heart Convent International School, Ludhiana, Punjab. He has always been interested in aerospace engineering and wants to pursue a career in that field.

Works Cited

[1] National Aeronautics and Space Administration. “Lift-to-Drag Ratio.” NASA Glenn Research Center, https://www.grc.nasa.gov/www/k-12/VirtualAero/BottleRocket/airplane/ldrat.html.

[2] Wunderlich, T. F. “Optimal Wing Aspect Ratio of Highly Efficient Long-Range Aircraft.” Proceedings of the 34th Congress of the International Council of the Aeronautical Sciences (ICAS 2024), Florence, Italy, 2024. German Aerospace Center (DLR).

[3] National Aeronautics and Space Administration. “Aircraft Geometry.” NASA Glenn Research Center, https://www.grc.nasa.gov/www/k-12/VirtualAero/BottleRocket/airplane/geom.html.

[4] Park, K., et al. “Optimal Design of Airfoil with High Aspect Ratio in Unmanned Aerial Vehicles.” International Journal of Aerospace and Mechanical Engineering, vol. 2, no. 4, 2008, pp. 171–177. https://doi.org/10.5281/zenodo.1077457.

[5] Goetzendorf-Grabowski, T., A. Frydrychewicz, Z. Goraj, and Suchodolski. “MALE UAV Design of an Increased Reliability Level.” Aircraft Engineering and Aerospace Technology: An International Journal, vol. 78, no. 3, 2006, pp. 226–235.

[6] Norris, G. “‘Shark Tail’ Shed as Carriers Wake Up to New 787 Shape.” FlightGlobal, https://flightglobal.com/reality-bites-for-dreamliner-fin/60080.article.

[7] Karpuk, S., R. Radespiel, and A. Elham. “Assessment of Future Airframe and Propulsion Technologies on Sustainability of Next-Generation Mid-Range Aircraft.” Aerospace, vol. 9, 2022, article 279. https://doi.org/10.3390/aerospace9050279.

[8] Fu, B., Y. Yang, H. Qi, J. Xu, and S. Wang. “Modeling and Analysis of High Aspect Ratio Wing Considering Random Structural Parameters.” Scientific Reports, vol. 11, 2021, article 15609. https://doi.org/10.1038/s41598-021-95187-0.

[9] Kretov, Anatolii, and Dmytro Tiniakov. “Evaluation of the Mass and Aerodynamic Efficiency of a High Aspect Ratio Wing for Prospective Passenger Aircraft.” Aerospace, vol. 9, 2022, article 497. https://doi.org/10.3390/aerospace9090497.

[10] Ding, Z. “Impact of Wing Aspect Ratio on Aerodynamic Performance in Gliding Flight: A Comprehensive Review.” Highlights in Science, Engineering and Technology, vol. 119, 2024, pp. 867–871. https://doi.org/10.54097/8p1hg461.

[11] Ma, Yiyuan, and Ali Elham. “Designing High Aspect Ratio Wings: A Review of Concepts and Approaches.” Progress in Aerospace Sciences, vol. 145, 2024, article 100983. https://doi.org/10.1016/j.paerosci.2024.100983.

[12] Jenkinson, Lloyd R., and Joseph F. Marchman III. “Trade-Off Studies.” Aircraft Design Projects, 2003.

[13] Nikolaev, N. V. “Optimization of Airfoils along High-Aspect-Ratio Wing of Long-Endurance Aircraft in Trimmed Flight.” Journal of Aerospace Engineering, vol. 32, no. 6, 2019, article 04019090. https://doi.org/10.1061/(ASCE)AS.1943-5525.0001086.

[14] Liang, B. “Research on the Aerodynamic Characteristics of the Relationship between Lift-to-Drag Ratio and Wing Shape of Gliders.” Highlights in Science, Engineering and Technology, vol. 93, 2024, pp. 185–192. https://doi.org/10.54097/ee2pgs04.
[15] Zhang, F. “The Relationship Between Aspect Ratio and Lift/Drag Ratio.” Highlights in Science, Engineering and Technology, vol. 81, 2024, pp. 411–417. https://doi.org/10.54097/r70n2m51.
[16] Dubnický, L., J. Šplíchal, and J. Juračka. “Sailplane Wing Sectioning Based on Structural Weight.” Engineering Mechanics 2023: 29th International Conference, 2023, pp. 59–62. https://doi.org/10.21495/em2023-59.
[17] Adeleke, O. A., G. E. Abbe, P. O. Jemitola, and S. Thomas. “Design of the Wing of a Medium Altitude Long Endurance UAV.” International Journal of Engineering and Manufacturing, vol. 12, no. 1, 2022, pp. 37–47. https://doi.org/10.5815/ijem.2022.01.04.
[18] Hrishikeshavan, V., C. Bogdanowicz, and I. Chopra. “Design, Performance and Testing of a Quad Rotor Biplane Micro Air Vehicle for Multi Role Missions.” International Journal of Micro Air Vehicles, vol. 6, no. 3, 2014, pp. 155–166. https://doi.org/10.1260/1756-8293.6.3.155.
[19] Montagnier, O., and L. Bovet. “Optimisation of a Solar-Powered High Altitude Long Endurance UAV with Composite Wings.” Proceedings of the 3rd European Conference for Aerospace Sciences, 2009.
[20] McDonnell, T., and A. Ning. “Gradient-Based Optimization of Solar-Regenerative High-Altitude Long-Endurance Aircraft.” Journal of Aircraft, vol. 57, no. 6, 2020, pp. 1189–1201. https://doi.org/10.2514/1.C035566.
[21] Hwang, S., S. Kim, and Y.-G. Lee. “Developing High Altitude Long Endurance (HALE) Solar-Powered Unmanned Aerial Vehicle (UAV).” Journal of Aerospace System Engineering, vol. 10, no. 1, 2016, pp. 59–65. https://doi.org/10.20910/JASE.2016.10.1.59.
[22] Ziakos, N., S. Kilimtzidis, and V. Kostopoulos. “Multidisciplinary Aeroelastic Optimization of High Aspect-Ratio Composite Aircraft Wings via Tow-Steering.” CEAS Aeronautical Journal, vol. 17, 2026, pp. 399–419. https://doi.org/10.1007/s13272-025-00851-0.
[23] Sahyoun, D., Y. Boose, R. Hein, N. Liebers, and R. Prussak. “Design and Manufacturing of an Aileron for a High-Aspect Ratio Wing with Active Flutter Control.” CEAS Aeronautical Journal, 2026. https://doi.org/10.1007/s13272-026-00954-2.
[24] Hsiao, F.-B., C.-Y. Lin, Y.-C. Liu, D.-B. Wang, C.-C. Hsu, and C.-H. Chiang. “Thickness Effect on Low-Aspect-Ratio Wing Aerodynamic Characteristics at a Low Reynolds Number.” Journal of Mechanics, vol. 24, no. 3, 2008, pp. 223–228. https://doi.org/10.1017/S1727719100002288.
[25] Uzun, M. “Impact of Aspect Ratio on Structural Integrity and Aerodynamic Performance in Fixed-Wing UAV.” Journal of Aviation, vol. 9, no. 1, 2025, pp. 19–27. https://doi.org/10.30518/jav.1583881.

[26] Bras, M., S. Warwick, and A. Suleman. “Aeroelastic Evaluation of a Flexible High Aspect Ratio Wing UAV: Numerical Simulation and Experimental Flight Validation.” Aerospace Science and Technology, vol. 122, 2022, article 107400. https://doi.org/10.1016/j.ast.2022.107400.
[27] Kilimtzidis, S., and V. Kostopoulos. “Static Aeroelastic Optimization of High-Aspect-Ratio Composite Aircraft Wings via Surrogate Modeling.” Aerospace, vol. 10, no. 3, 2023, article 251. https://doi.org/10.3390/aerospace10030251.
[28] Wahler, N. F. M., and A. Elham. “Towards Optimal Wing Design for Novel Airframe and Propulsion Opportunities.” Aerospace, vol. 12, no. 6, 2025, article 459. https://doi.org/10.3390/aerospace12060459.

The BDNF Val66Met Polymorphism As A Potential Mechanism of Individual Differences in Cognitive Vulnerability to Sleep Deprivation: A Review

By Hania Nabeel

Abstract

Sleep deprivation is known for causing impairments in attention, working memory, learning, and many other cognitive abilities, making it a crucial concern for health and cognitive performance. The negative effects of sleep deprivation are well documented, however, individuals vary in their susceptibilities to these impairments, suggesting the involvement of other biological factors that can contribute to these differences. Researchers have used functional magnetic resonance imaging (fMRI) to investigate the regions of the brain that are affected by sleep deprivation. By measuring brain activity through changes in blood oxygenation and blood flow, fMRI studies have shown that sleep deprivation can affect prefrontal cortex function, a prominent brain region involved in working memory, attention, and decision making. Despite the well established cognitive effects of sleep deprivation, the biological mechanisms explaining individual differences in vulnerability remain unclear. This review investigates whether genetic variation in BDNF and its effects on synaptic plasticity can help explain why some individuals experience greater cognitive impairment than others due to sleep deprivation. Understanding this relationship may provide insights into the neurological and genetic factors associated with cognitive vulnerability and contribute to a more complete understanding of individual differences in the cognitive effects of sleep deprivation.

Keywords

Sleep deprivation, Synaptic plasticity, Individual differences, Brain-derived neurotrophic factor (BDNF), cognitive vulnerability.

Introduction

Sleep is a fundamentally important recovery and biological process that allows the body to enter a deep state of rest to maintain physical, mental, and cognitive function. It is characterized by behavioral and neurophysiological changes that can include reduced responsiveness to external stimuli (Chokroverty). Beyond being a state of rest, sleep is essential for supporting cognitive processes such as learning, memory, and decision-making. Despite its importance, many individuals are unable to get the required amount of sleep for efficient cognitive performance. Insufficient sleep may result from lifestyle factors and can range from partial sleep restriction to complete sleep deprivation (Orzeł-Gryglewska). Sleep deprivation has been shown to impair many cognitive functions, including attention, learning, memory consolidation, alertness, and other physiological processes such as maintaining homeostasis (Khan and Al-Jahdali). However, despite a lot of evidence documenting these effects, the reasons why some individuals appear to be more resilient or more vulnerable to cognitive impairment following sleep deprivation remain incompletely understood. One potential explanation for these

individual differences is variation in biological mechanisms involved in cognitive function, including brain-derived neurotrophic factor (BDNF). BDNF is a neurotrophin involved in important processes such as synaptic plasticity, a key neurological process involved in learning and memory (Kowiański et al.). It may contribute to differences in vulnerability to cognitive impairments following sleep deprivation. This review evaluates evidence on the cognitive consequences of sleep deprivation, the role of the prefrontal cortex, BDNF-regulated synaptic plasticity, and BDNF signaling. Furthermore, it evaluates evidence for the genetic variants of BDNF, particularly Val66Met, and how they may represent a potential biological mechanism that contributes to inter-individual differences in cognitive performance following sleep deprivation.

Methods

This narrative review examines whether brain-derived neurotrophic factor (BDNF) may help explain individual differences in cognitive vulnerability to sleep deprivation. Relevant literature was identified through Google Scholar between 6 July and 26 August 2026, using search terms including "BDNF Individual differences Sleep deprivation," "Sleep deprivation memory," "Sleep deprivation decision-making," "Sleep deprivation attention," "Sleep deprivation prefrontal cortex," and "How does lack of sleep affect synaptic plasticity."

The search initially focused on studies published between 2015 and 2026 to prioritize recent evidence. Relevant reviews were used to understand the main findings in each area and their reference lists were followed to locate the original studies. This also led to older foundational studies on the Val66Met polymorphism and individual differences in vulnerability following sleep loss. Human and animal studies were included, with animal studies providing mechanistic evidence when human evidence was limited.

As this was a narrative rather than a systematic review, the search was not designed to identify every relevant publication. Reliance on Google Scholar and reference lists from existing reviews may have resulted in some relevant studies being missed.

The cognitive consequences of sleep deprivation

Sleep plays a vital role in cognitive performance during wakefulness, allowing individuals to carry out daily activities such as learning and working. Insufficient sleep, referred to as sleep deprivation, is an increasingly common public health concern that has been shown to impair performance on cognitive tasks involving response speed, memory, verbal comprehension, and arithmetic efficiency. These impairments are caused by disruption in three core cognitive processes, which are attention, working memory, and executive functions (García et al.). Among these processes, memory is a key process that is consistently affected by sleep deprivation.

Sleep plays a vital role in memory consolidation. Evidence suggests that different stages of sleep contribute to memory processing. REM sleep contributes to the refinement of memories, including procedural and emotional memory, whereas NREM sleep is primarily associated with the strengthening and stabilisation of newly acquired memories, particularly declarative memories (Adeniji et al.). Disruption of these sleep stages may therefore interfere with memory

consolidation, leaving newly encoded memory traces more vulnerable to interference (Gais et al.).

A theory exploring the role of sleep in memory is the active system consolidation theory (ASC). According to this theory, newly formed memories initially depend strongly on the hippocampus and are gradually reorganised within distributed neocortical networks, particularly during NREM sleep. During sleep, these memories undergo repeated reactivation, which supports their consolidation into longer-term memory representation through interactions between the hippocampus and neocortex (**Figure 1**). However, not all memories are strengthened equally, as future relevance of the memory can influence consolidation (Adeniji et al.; Born and Wilhelm).

Research investigating memory reactivation has shown that the timing of reactivation can influence whether memories become strengthened or disrupted. During slow wave sleep (SWS), reexposure to an odour cue associated with previously learned card locations stabilized these memories, making them more resistant to interference. However, similar reactivation during wakefulness can disrupt memories making them more vulnerable to interference (Born and Wilhelm; Diekelmann et al.). These findings support the ASC theory by demonstrating that memory reactivation during NREM sleep contributes to the stabilization of previously learned memories. Therefore, disruption of NREM sleep through sleep deprivation may interfere with memory reactivation and consolidation, potentially contributing to impaired memory.

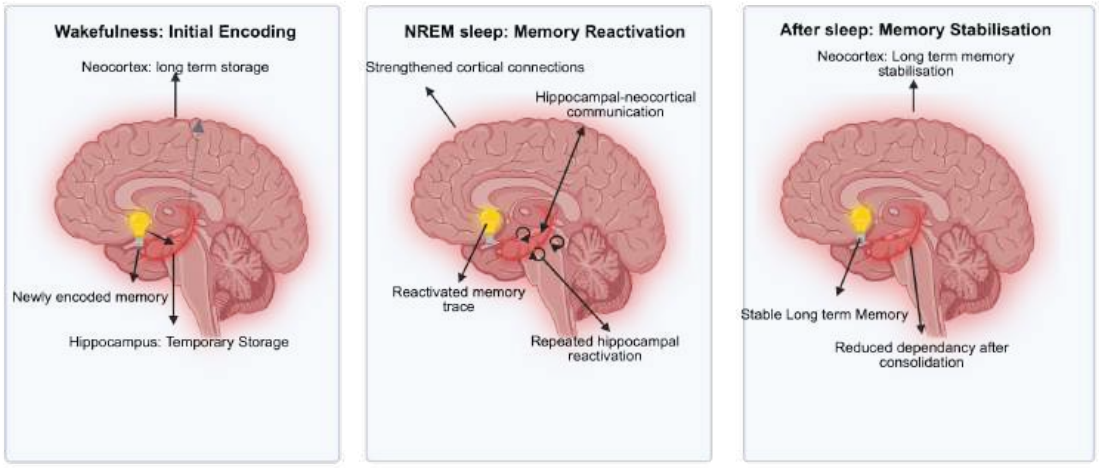


Figure 1. Active system consolidation (ASC) during sleep-dependent memory consolidation. Newly encoded memories are initially associated with the hippocampus during wakefulness. During NREM sleep, the memory trace is repeatedly reactivated and communicated between the hippocampus and neocortex, allowing the memory to become more stable and consolidated over time (Born and Wilhelm, 2012).

In a controlled laboratory study of 39 adults, Chai et al (2020) found that one night of total sleep deprivation (TSD) significantly impaired episodic memory performance. Compared with baseline performance, participants showed lower hit rates, more false alarms, and reduced discrimination accuracy, indicating that participants found it difficult to recognise previously learned information. This suggests that even a single night of sleep deprivation is sufficient to impair episodic memory. Additionally, memory performance remained significantly low compared to baseline despite two consecutive nights of recovery sleep, suggesting that two nights of recovery sleep were insufficient to fully restore episodic memory function following one night of sleep deprivation (Chai et al.). This indicates that the cognitive effects of sleep loss may persist beyond the initial period of sleep deprivation (Zamore and Veasey).

Decision-making is a complex cognitive process that requires individuals to evaluate different choices, consider potential consequences, and use previous experiences to select the most appropriate response to a scenario that has multiple alternatives and outcomes (Lim et al.). Because decision-making depends on multiple cognitive processes, disruption of these processes may impair judgment and the evaluation of potential consequences. Therefore, sleep deprivation may impair decision-making by disrupting the cognitive processes required to evaluate given information and make informed choices (Hyndych et al.).

The Iowa Gambling Task measures whether individuals can adjust their decision-making strategies based on implicit learning of punishment and reward contingencies (Killgore et al.). This makes the task useful for investigating real-world decision making because effective decisions require individuals to consider short-term benefits and long-term consequences.

In a study where participants underwent 49.5 hours of sleep deprivation, Killgore et al (2006) found that under rested conditions participants eventually avoided disadvantageous card decks and selected more advantageous card decks. However, after 49.5 hours of sleep deprivation, participants were less able to maintain an advantageous decision-making strategy, which led to selecting more cards from the disadvantageous decks. This suggests that sleep deprivation impairs the ability to learn and adjust decisions based on previous patterns and outcomes. Furthermore, sleep deprivation reduced participants' ability to balance the immediate benefits of short-term rewards against greater long-term consequences. This may negatively influence judgement by disrupting the cognitive processes involved in evaluating risk and making appropriate choices (Killgore et al.).

Alertness and attention are crucial cognitive processes that support higher cognitive processes. Without attention, individuals are unable to engage in complex cognitive activities such as reasoning and decision-making. In particular, alertness and attention are two main processes consistently impaired by insufficient sleep (Killgore). Because higher-order cognitive processes depend on sustained attention, impairments in alertness may also compromise reasoning and decision making.

Vigilant attention, also known as sustained attention, is the ability to maintain stable and focused attention across a specific period of time. The psychomotor vigilance task (PVT) is the most common metric used to evaluate vigilance and alertness. The standard PVT is a 10-minute reaction-time task in which participants respond as quickly as possible to visual stimuli that are

presented at random intervals (Hudson et al.). Using the PVT, Mao et al (2021) found that a total night of sleep deprivation (TSD) significantly slowed down reaction times and increased the number of attentional lapses compared to baseline performance. These findings demonstrate that vigilant attention is particularly vulnerable to sleep loss (Mao et al.).

Furthermore, similar findings have been reported by Whitney et al (2015), who found that vigilant attention deteriorated over 62 hours of sleep deprivation. Participants demonstrated an increasing number of attentional lapses on the 10-minute PVT as sleep deprivation continued, with impairments being the greatest during morning hours because of circadian influences. However, performance improved greatly following recovery sleep, suggesting that some cognitive impairments after sufficient recovery are partially reversible. The progressive increase in PVT lapses across the 62-hour-sleep deprived period suggests that vigilant attention is progressively impaired as sleep deprivation continues (Hudson et al.; Whitney et al.).

Despite consistent evidence that sleep deprivation impairs vigilant attention, the vulnerability of these impairments varies considerably among individuals. This variability has been observed under total sleep deprivation (TSD), suggesting that vulnerability to sleep loss is not uniform across the population. A study investigating interindividual differences in vigilant attention following 36 hours of total sleep deprivation found that, although prior sleep history influenced performance, its effects were relatively small compared with stable trait-like differences (Van Dongen et al.). In particular, trait-like differences in vulnerability accounted for 67.5% of the variance in vigilant attention impairments (Hudson et al.). This indicates that stable individual characteristics may contribute to an individual's susceptibility to cognitive decline following sleep loss. These findings suggest that while environmental factors such as previous sleep duration can influence cognitive performance, they do not fully explain why some individuals remain relatively resilient while others experience attentional impairments to a great extent (Kuna et al.). This suggests that biological differences between individuals may contribute towards different sleep responses.

Since vigilant attention relies heavily on the prefrontal cortex, a brain region responsible for maintaining attention, cognitive control and decision-making, disruptions within this region may represent a key mechanism underlying cognitive decline following sleep deprivation. Therefore, understanding how sleep deprivation affects the prefrontal cortex provides an important insight into the basis of impaired cognition due to sleep loss.

The Prefrontal cortex. A neural target for sleep deprivation

The prefrontal cortex (PFC) is a crucial brain region involved in higher-order cognitive function, including reasoning, planning, and decision-making. Unlike regions responsible for basic sensory or motor processing, the PFC integrates information from multiple brain regions to support complex cognitive processes that are essential for daily life (Funahashi). This demonstrates that the PFC plays a central role in regulating executive functions that allow individuals to maintain and manipulate information and make informed decisions.

Sleep deprivation produces variable effects amongst individuals. However, neuroimaging evidence suggests that the prefrontal cortex (PFC) and its connection with other subcortical

structures, including the amygdala and hippocampus, are particularly prone to disruption following sleep loss (Frau et al.). This suggests that the PFC relies on communication with other brain regions involved in processes such as memory and emotional regulation, meaning that disruption of these networks may have broader cognitive consequences (Harrington et al.). Furthermore, the PFC is involved in both attention and working memory with shared neural mechanisms supporting both processes (Bahmani et al.). This indicates that disruptions in PFC activity may impair higher-order cognitive abilities that rely on sustained attention and the maintenance and manipulation of information.

To investigate how sleep deprivation alters PFC function, neuroimaging studies using fMRI have examined changes in PFC activity during cognitively demanding tasks. Using functional magnetic resonance imaging (fMRI), Chee and Choo (2004) investigated how sleep deprivation alters brain activity during verbal working memory. Participants were scanned twice, once during rested wakefulness and once after approximately 23 hours of sleep deprivation. There were two memory tasks, a memory maintenance task (LTR) and a memory manipulation task (PLUS), which required more advanced cognitive functioning. The LTR task was a memory maintenance task requiring participants to memorise four letters that were briefly displayed and determine whether a letter probe matched the original information. The PLUS task assessed memory manipulation, requiring participants to mentally transform letters before comparing them with a later probe. While participants completed the tasks inside the MRI scanner, fMRI measured changes in blood oxygenation level dependent (BOLD) signals, which provide an indirect measure of changes in neural activity. Researchers conducted both whole-brain voxel-by-voxel analysis and a region of interest (ROI) analysis to examine changes in brain activity following sleep deprivation, including activity within the frontal, parietal, and thalamic regions. Researchers examined the relationship between the BOLD signal and task performance and found that after sleep deprivation the left dorsolateral prefrontal cortex (dlPFC) showed greater activation during the working memory task compared with when participants were rested. This greater activation suggested increased recruitment of the dlPFC during the more demanding PLUS task following sleep deprivation. Despite this increased activation, participants demonstrated slower reaction times on both working memory tasks following sleep deprivation, and the researchers interpreted this pattern as compensatory neural effort rather than enhanced cognitive performance. Furthermore, increased prefrontal activation was not observed during the simpler LTR maintenance, suggesting that sleep deprivation may affect higher cognitive processes requiring manipulation of information rather than simpler maintenance processes. Additionally, reduced deactivation was observed in the anterior medial frontal cortex and posterior cingulate cortex in both voxel-by-voxel contrast and ROI-based analysis, suggesting that regulation of activity in these regions following sleep deprivation was altered (Chee and Choo). Because fMRI measures blood oxygenation rather than neural activity directly, these findings should be interpreted cautiously, and they do not explain why some individuals compensate for sleep loss better than others. This raises the possibility that differences in the

mechanisms that support PFC function, such as synaptic plasticity and BDNF signaling, may contribute to individual differences in cognitive vulnerability.

Importantly, the cognitive effects of sleep deprivation vary between individuals with some individuals being more vulnerable or resistant to sleep loss than others. Chuah et al (2006) investigated this interindividual variability using a go/no-go task during fMRI and classified participants into low vulnerable, moderately vulnerable, and highly vulnerable groups according to the extent of their decline in inhibitory efficiency following sleep deprivation. Inhibitory efficiency did not significantly differ between the three groups during rested wakefulness, indicating that they performed similarly before sleep deprivation. However, following sleep deprivation, variability in go/no-go performance increased significantly, alongside reductions in hit and stop rates, indicating an overall decline in inhibitory control. Furthermore, the extent of this decline differed between individuals, with some participants maintaining inhibitory efficiency better than others. Individuals who best maintained inhibitory efficiency showed lower right ventrolateral PFC (vlPFC) activation during rested wakefulness compared with more vulnerable individuals, but were able to increase activation in this region following sleep deprivation, whereas the highly vulnerable group did not. Additionally, greater right vlPFC activation during rested wakefulness was strongly associated with a greater decline in inhibitory efficiency, whereas a greater relative increase in activation following sleep deprivation was associated with a smaller decline in inhibitory efficiency. The authors proposed that this transient increase in right vlPFC activation may represent a compensatory response to reduced sustained task-related PFC activation following sleep deprivation. These findings may indicate that cognitive vulnerability to sleep deprivation involves disruption of PFC activity and differs according to an individual's ability to recruit PFC activation in response to this disruption (Chuah et al.). Therefore, differences in the biological mechanisms supporting this neural response may help explain why some individuals vary in the extent of cognitive impairment following sleep deprivation.

These differences in PFC responses may help explain why some individuals are more resilient or vulnerable to sleep loss than others. One process that may contribute towards this difference is synaptic plasticity, which allows neural connections to change in response to activity and experiences. Because sleep deprivation has been shown to alter the capacity for further synaptic plasticity, differences in the mechanisms affecting synaptic plasticity may help explain why individuals respond differently to sleep loss.

Synaptic Plasticity and BDNF

Learning is associated with experience-driven changes in the brain that allow neural circuits and pathways to adapt to daily experiences. One mechanism underlying this adaptation is synaptic plasticity, which is defined as activity-dependent changes in the strength of connections between neurons, and learning and experience can produce lasting changes in these connections. Furthermore, synaptic plasticity is considered an important underlying mechanism for learning

and memory (Magee and Grienberger). Because synaptic plasticity is involved in learning and memory, sleep may also play an important role in regulating these synaptic changes.

The synaptic homeostasis hypothesis (SHY) proposes that synaptic strength increases during wakefulness as a result of ongoing experience and learning, while slow wave NREM sleep counteracts the accumulation of these experiences through the downscaling of synapses. This downscaling process is thought to explain the benefits of sleep in learning and memory by preventing excessive synaptic activity and energy use. However, sleep deprivation may prevent the normalisation of synaptic strength, potentially disrupting learning ability and neural networks. Therefore, continued sleep loss may interfere with the restoration of cellular and synaptic homeostasis, potentially affecting the brain's capacity for further plasticity during wakefulness. This synaptic renormalisation may also support learning and memory by reducing energy demands and cellular stress. Even mild sleep deprivation can impair the encoding of new information. Findings on long-term sleep deprivation and declarative memory have been mixed, with some studies finding no effect, including in adolescents. Sleep deprivation before training tasks can also reduce episodic memory retention despite reaction time during training remaining unaffected; this suggests that sleep deprivation may affect the encoding of new memories (Khan and Al-Jahdali). However the mixed findings on memory indicate that the synaptic homeostasis hypothesis may not fully account for the cognitive effects of sleep deprivation. This suggests that SHY provides a useful mechanistic framework, but other processes may also contribute to the effects of sleep loss on memory.

A study of 20 healthy participants by Kuhn and colleagues (2016) investigated how sleep deprivation affects synaptic plasticity by comparing participants after a normal night of sleep with one night of sleep deprivation. The researchers found that sleep deprivation increased cortical excitability and overall synaptic strength, which were measured using TMS and EEG. However, when LTP-like plasticity was induced using paired associative stimulation (PAS), participants showed a reduced ability to undergo further synaptic potentiation. After normal sleep, 70% of participants showed an LTP-like increase, compared with only 25% following sleep deprivation, while 75% instead showed an LTD-like response. Sleep deprivation was also associated with poorer word-pair learning, suggesting that reduced capacity for synaptic plasticity has consequences for memory formation. These findings support the synaptic homeostasis hypothesis, as prolonged wakefulness appears to increase synaptic strength while reducing the capacity for further potentiation, which may contribute to the cognitive impairments observed following sleep deprivation (Kuhn et al., 2016). However, the small sample size limits generalisability, opening the question whether differences in regulation of synaptic plasticity could explain why individuals vary in their cognitive responses following sleep deprivation.

Because BDNF is an important regulator of synaptic plasticity and LTP, it is a plausible candidate for explaining why plasticity under sleep loss differs between individuals. BDNF is one of the most common and well-researched neurotrophins found throughout the brain, and it contributes to the growth of neurons and glial cells, while also protecting nerve cells from damage. BDNF plays a key role in learning and memory by influencing how synaptic

connections behave over different periods of time. The effects of BDNF are mediated through interactions with different receptors and signaling pathways, allowing BDNF to influence the strength and function of neuronal connections. Therefore, BDNF may be a link between synaptic function and cognitive processes (Kowiański et al.). Because synaptic plasticity is important for maintaining cognitive function, changes in BDNF signaling could potentially contribute to how well the brain adapts to sleep deprivation.

A review by Wang et al (2022) summarised evidence that BDNF is involved in synaptic plasticity through both structural and molecular changes at the synapse. BDNF can promote synapse formation by increasing axon and dendrite branching, creating more boutons along them and stabilizing existing synapses. Furthermore, BDNF is also important for long-term potentiation (LTP). When BDNF is removed, both the induction and the strength of LTP are reduced, while adding recombinant BDNF can help rescue these impairments. Further research suggests that BDNF-TrkB signaling can both enable LTP and regulate how it develops during different phases. BDNF binds to TrkB receptors, which activates downstream pathways including PLCγ and Fyn. PLCγ signaling is involved in both early- and late-phase LTP, while Fyn increases NMDA receptor activity and calcium influx, promoting synaptic potentiation. TrkB signaling also has different roles depending on its location with postsynaptic TrkB involved in LTP induction and both presynaptic and postsynaptic TrkB involved in LTP maintenance. Stimulation can also cause BDNF to be released from the postsynaptic spine, where it activates nearby TrkB receptors and contributes to the structural and functional changes involved in LTP (Wang et al.). Overall, this evidence demonstrates that BDNF-TrkB signaling can influence changes in synaptic strength that occur during synaptic plasticity.

Researchers have also found that BDNF plays an important role in long-term memory (LTM). Lu et al (2014) reviewed evidence showing late-phase LTP, a form of synaptic plasticity that is considered to be the cellular foundation for long-term memory, and that BDNF has been shown to contribute to this process. The review also found that hippocampal BDNF mRNA levels increase following learning-related activities, suggesting that BDNF expression responds to neural activity during learning. In addition, disrupting BDNF signaling impairs long-term memory formation, supporting its role in memory processes. BDNF also appears to be involved at two separate time windows of memory consolidation, the first occurring 1-4 hours after encoding, which is important for memories lasting 1-2 days, and the second occurring around 12 hours after memory formation, which is important for memories lasting up to 7 days (Lu et al.). These findings support the role of BDNF as an important regulator of activity-dependent changes in synapses that support long-term memory. This makes BDNF particularly relevant to sleep deprivation, as changes in BDNF signaling could influence synaptic plasticity involved in the cognitive effects of sleep loss.

BDNF plays a role in both the regulation of sleep drive and the encoding of memories across the sleep-wake cycle (Schmitt et al.). Human studies also suggest a link between sleep deprivation and BDNF. Giacobbo et al (2016) found that acute total sleep deprivation increased serum BDNF levels in healthy young adults compared with the control group. However, the

sleep-deprived participants also performed worse on immediate and delayed verbal recall, suggesting that the increase in BDNF did not prevent memory impairments. The authors suggested that the increase in BDNF may partly be related to activity-dependent effects, as participants remained awake and engaged in activities throughout the night. This suggests that the changes in BDNF following sleep deprivation may be influenced by both sleep loss and the activities performed during prolonged wakefulness, although the authors note that this needs to be tested directly (Giacobbo et al.).

BDNF and Individual differences

The effects of sleep deprivation on neurobehavioral functioning vary substantially between individuals, and consistent predictors of vulnerability and resilience to sleep loss have yet to be established. Casale and Goel (2021) note that this vulnerability appears to be relatively stable and heritable. Twin studies have found greater similarity in responses to acute total sleep deprivation in monozygotic twins compared to dizygotic twins (Kuna et al.), demonstrating genetic contribution. The distribution of neurobehavioral responses to sleep loss suggests that multiple genes may contribute to individual differences, although candidate gene studies have identified specific polymorphisms associated with differences in response to sleep loss (Casale and Goel). Given this evidence for a genetic basis of sleep loss vulnerability, genes involved in neural plasticity, such as BDNF, represent potential candidates for further investigation.

A common single nucleotide polymorphism in the BDNF gene, Val66Met (rs6265), results in a valine-to-methionine substitution at codon 66. Although the variant occurs in the pro-BDNF region rather than the mature protein, it affects the intracellular trafficking and activity-dependent secretion of BDNF. Egan et al (2003) found that neurons expressing the Met variant had lower depolarisation-induced BDNF secretion, while constitutive secretion was unchanged. In human participants, the Met allele was also associated with poorer episodic memory, altered hippocampal activation and lower hippocampal N-acetyl aspartate (NAA), providing evidence that this variant may affect hippocampal function, and memory through altered BDNF signaling (Egan et al.). These findings suggest that genetic variation in BDNF could alter activity-dependent BDNF signaling and hippocampal function, providing a potential biological basis for differences in cognitive performance. Given the role of BDNF in synaptic plasticity and memory, this raises the possibility that variation in BDNF may also contribute to differences in cognitive vulnerability following sleep deprivation.

Altered BDNF secretion associated with Val66Met may have consequences for synaptic plasticity. Ninan et al (2010) found that BDNF Met/Met mice showed reduced long-term potentiation (LTP) at hippocampal CA3-CA1 synapses compared with wild-type mice. In young mice, TBS-induced LTP was almost completely absent in BDNF Met/Met mice, while in adult mice, initial potentiation declined towards baseline, resulting in significantly lower levels of late LTP. As activity-dependent BDNF release contributes to TBS-induced LTP, these findings indicate that altered BDNF signaling associated with the Met variant may impair both the induction and maintenance of hippocampal synaptic potentiation (Ninan et al.).

Sleep deprivation affects cognition at the group level, but not everyone is affected equally; some individuals are more vulnerable than others, and this variability could have a genetic basis behind it. One candidate gene is BDNF, Val66Met polymorphism; the Met substitution disrupts activity-dependent secretion, which is a mechanism important for synaptic plasticity. Grant et al (2018) tested whether this genotype affects how people respond to sleep loss and found that Val/Met carriers performed worse on some measures. Val/Met carriers made 3.54 percentage points more errors than the Val/Val participants on incongruent Stroop trials but only during the last 10 hours of extended wakefulness, suggesting greater vulnerability to cognitive impairment following prolonged sleep loss. Val/Met carriers also showed a greater overall cost of inhibition. This did not depend on time awake, although a follow-up analysis suggested the difference was concentrated during the biological night. Overall, these results suggest that BDNF genotype may contribute to differences in vulnerability to sleep loss, although the underlying mechanism has not yet been established (Grant et al.).

There is an important gap between the hippocampal and prefrontal evidence. The mechanistic evidence for the effect of Val66Met on synaptic plasticity primarily comes from the hippocampus. Egan et al (2003) linked the Met allele to differences in hippocampal function in humans, while Ninan et al (2010) found reduced LTP at hippocampal CA3-CA1 synapses in Met/Met mice. In contrast, much of the evidence for the cognitive vulnerability to sleep deprivation discussed in this review concerns prefrontal functions, including working memory and inhibitory efficiency (Chee and Choo; Chuah et al.). The finding from Grant et al (2018) offers a partial bridge, as described above, Val/Met carriers made more errors on incongruent Stroop trials in the third time-awake tertile. This is consistent with greater inhibitory control vulnerability in Met carriers, but Grant et al measured task performance rather than brain activity, and Chuah et al linked inhibitory efficiency to vlPFC activation using a different task (go/no-go). The connection between BDNF variation, prefrontal function and sleep deprivation vulnerability therefore remains inferential. Future studies combining BDNF genotyping with fMRI during sleep deprivation could test whether Met carriers show differences in prefrontal recruitment that accompany differences in inhibitory performance.

A further limitation is that Grant et al included no Met/Met participants, limiting the finding to heterozygous carriers and preventing conclusions about whether two Met alleles would produce a greater effect. A similar limitation applies to Bachmann et al (2012) who examined the effect of Val66Met on sleep EEG; one Met/Met individual was identified during genotyping, but the experimental analysis only included Val/Met and Val/Val participants, so the very small number of Met/Met individuals makes it difficult to investigate whether two Met alleles have a greater effect than one (Bachmann et al.).

Candidate-gene findings, including those on Val66Met, should be treated with caution, because associations found in small samples often fail to replicate in larger ones. A large study of 18 depression candidate genes, including BDNF rs6265 found no clear support for them (Border et al.). This was about depression, not sleep loss, but it shows how hard it is to link single variants to complex traits. The Val66Met literature shows a similar pattern. A meta-analysis

combining new data from 643 healthy adults with published studies, 5298 participants in total, found only a weak link between the Met allele and smaller hippocampal volume (Harrisberger et al.). The effect came mainly from manual-tracing studies, which showed publication bias, and studies using automated segmentation found no significant difference. Hippocampal volume is not a direct measure of BDNF secretion or plasticity, so this does not undermine the functional evidence from Egan et al (2003) and Ninan et al (2010). The sleep deprivation evidence is also limited. A study of only 30 participants found more incongruent Stroop errors in Val/Met carriers only late in extended wakefulness, so replication is needed (Grant et al.). Since vulnerability to sleep loss probably involves many genes (Casale and Goel), Val66Met is best seen as one possible contributor, not a full explanation.

Conclusion

This review examined how sleep deprivation affects cognitive functioning and explored genetic variation in BDNF as a potential mechanism underlying individual differences in vulnerability to the effects of sleep deprivation. The empirical evidence demonstrates that sleep deprivation impairs multiple cognitive domains, including memory consolidation, decision-making, and vigilant attention, and that these impairments are closely linked to changes in the prefrontal cortex. However, the degree of impairment varies considerably between individuals, and this variability appears to be stable and partly heritable, suggesting that genetic factors may contribute to these differences. BDNF is a potential contributor to this variability because of its established role in synaptic plasticity, specifically long-term potentiation (LTP), which is important for learning and memory. The Val66Met polymorphism, which affects activity-dependent secretion, can be a possible link between genetic variation and differences in cognitive vulnerability. Grant et al. (2018) provide preliminary support: Val/Met carriers made more errors than Val/Val homozygotes on incongruent Stroop trials late in extended wakefulness, and showed a greater cost of inhibition that did not depend on time awake.. Individuals carrying the Met allele may have a reduced ability to maintain the synaptic processes needed for normal cognitive performance during extended wakefulness, although the exact mechanisms behind this relationship appear to be unclear.

Future research should use larger and more genetically diverse samples which include Met/Met carriers if possible to determine whether vulnerability to sleep deprivation increases with the number of Met alleles. Recruiting from populations where the Met allele is more common would make the Met/Met participants easier to find, because its frequency ranges from 0 to 72% across 58 global populations and is higher in Asian rather than European populations (Petryshen et al.). This matters because the sample in Gran et al was entirely of European ancestry, where Met/Met individuals are less common. Further research is also needed to investigate how BDNF interacts with other biological factors that can contribute to individual differences in cognitive vulnerability following sleep deprivation. Overall, the evidence reviewed suggests that genetic variation in BDNF may contribute to interindividual differences in the cognitive effects of sleep deprivation, but BDNF is unlikely to provide a complete explanation.

Strengths and Limitations

A strength of this review is that it brings together evidence from different areas to look at why people may respond differently to sleep deprivation. The review considered behavioral studies, brain imaging, studies of synaptic plasticity and genetic research. Looking at these different types of evidence helps connect changes in cognitive performance with possible changes in brain function and BDNF signaling. The review also used both human and animal studies, which provides evidence and helps explain how BDNF Val66Met could affect synaptic plasticity.

However, there are some limitations to the current evidence. There is still limited research directly connecting BDNF Val66Met to changes in BDNF signaling, synaptic plasticity, and cognitive performance after sleep deprivation. Instead, the proposed link is based on findings from different studies that examine different parts of the process. The studies also use different sleep deprivation periods, cognitive tasks, and methods, which makes it difficult to directly compare their results. The human studies linking BDNF genotype to vulnerability to sleep loss also used small samples, which limits generalisability. In Grant et al (2018), only the error differences late in extended wakefulness clearly depended on sleep loss, while the other genotype differences did not depend on time awake. The replication record of candidate-gene studies, including those on Val66Met, is also limited, as discussed above. Another limitation is the lack of Met/Met participants in human sleep deprivation studies, meaning that it is not clear whether having more Met alleles leads to greater cognitive impairment. Some of the evidence for the effect of Val66Met comes from animal studies, which may not fully represent the complexity of the human brain. Therefore, BDNF Val66Met is a possible contributor to individual differences in response to sleep deprivation, but it cannot currently be considered the main or only explanation.

Works Cited

Adeniji, Kehinde Oluwaseyi, et al. “Sleep Deprivation and Memory: A Neurobiological Perspective.” IBRO Neuroscience Reports, 2026.

Bachmann, Valerie, et al. “The BDNF Val66Met Polymorphism Modulates Sleep Intensity: EEG Frequency-and State-Specificity.” Sleep, vol. 35, no. 3, 2012, pp. 335–44.

Bahmani, Zahra, et al. “Prefrontal Contributions to Attention and Working Memory.” Processes of Visuospatial Attention and Working Memory, 2019, pp. 129–53.

Border, Richard, et al. “No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression across Multiple Large Samples.” American Journal of Psychiatry, vol. 176, no. 5, 2019, pp. 376–87.

Born, Jan, and Ines Wilhelm. “System Consolidation of Memory during Sleep.” Psychological Research, vol. 76, no. 2, 2012, pp. 192–203.

Casale, Courtney E., and Namni Goel. “Genetic Markers of Differential Vulnerability to Sleep Loss in Adults.” Genes, vol. 12, no. 9, 2021, p. 1317.

Chai, Ya, et al. “Two Nights of Recovery Sleep Restores Hippocampal Connectivity but Not Episodic Memory after Total Sleep Deprivation.” Scientific Reports, vol. 10, no. 1, 2020, p. 8774.

Chee, Michael WL, and Wei Chieh Choo. “Functional Imaging of Working Memory after 24 Hr of Total Sleep Deprivation.” The Journal of Neuroscience, vol. 24, no. 19, 2004, pp. 4560–67.

Chokroverty, Sudhansu. “Overview of Sleep & Sleep Disorders.” Indian Journal of Medical Research, vol. 131, no. 2, 2010, pp. 126–40.

Chuah, YM Lisa, et al. “The Neural Basis of Interindividual Variability in Inhibitory Efficiency after Sleep Deprivation.” The Journal of Neuroscience, vol. 26, no. 27, 2006, pp. 7156–62.

Diekelmann, Susanne, et al. “Labile or Stable: Opposing Consequences for Memory When Reactivated during Waking and Sleep.” Nature Neuroscience, vol. 14, no. 3, 2011, pp. 381–86.

Egan, Michael F., et al. “The BDNF Val66met Polymorphism Affects Activity-Dependent Secretion of BDNF and Human Memory and Hippocampal Function.” Cell, vol. 112, no. 2, 2003, pp. 257–69.

Frau, Roberto, et al. “Neurobehavioural Complications of Sleep Deprivation: Shedding Light on the Emerging Role of Neuroactive Steroids.” Journal of Neuroendocrinology, vol. 32, no. 1, 2020, p. e12792.

Funahashi, Shintaro. “Working Memory in the Prefrontal Cortex.” Brain Sciences, vol. 7, no. 5, 2017, p. 49.

Gais, Steffen, et al. “Sleep Transforms the Cerebral Trace of Declarative Memories.” Proceedings of the National Academy of Sciences, vol. 104, no. 47, 2007, pp. 18778–83.

García, Aída, et al. “Sleep Deprivation Effects on Basic Cognitive Processes: Which Components of Attention, Working Memory, and Executive Functions Are More Susceptible to the Lack of Sleep?” Sleep Science, vol. 14, no. 2, 2021, p. 107.

Giacobbo, Bruno Lima, et al. “Could BDNF Be Involved in Compensatory Mechanisms to Maintain Cognitive Performance despite Acute Sleep Deprivation? An Exploratory Study.” International Journal of Psychophysiology, vol. 99, 2016, pp. 96–102.

Grant, Leilah K., et al. “Impaired Cognitive Flexibility during Sleep Deprivation among Carriers of the Brain Derived Neurotrophic Factor (BDNF) Val66Met Allele.” Behavioural Brain Research, vol. 338, 2018, pp. 51–55.

Harrington, Marcus O., et al. “Memory Control Deficits in the Sleep-Deprived Human Brain.” Proceedings of the National Academy of Sciences, vol. 122, no. 1, 2025, p. e2400743122.

Harrisberger, F., et al. “The Association of the BDNF Val66Met Polymorphism and the Hippocampal Volumes in Healthy Humans: A Joint Meta-Analysis of Published and New Data.” Neuroscience & Biobehavioral Reviews, vol. 42, 2014, pp. 267–78.

Hudson, Amanda N., et al. “Sleep Deprivation, Vigilant Attention, and Brain Function: A Review.” Neuropsychopharmacology, vol. 45, no. 1, 2020, pp. 21–30.

Hyndych, Anna, et al. “The Role of Sleep and the Effects of Sleep Loss on Cognitive, Affective, and Behavioral Processes.” Cureus, vol. 17, no. 5, 2025, pp. e84232–e84232.

Khan, Mohammad A., and Hamdan Al-Jahdali. “The Consequences of Sleep Deprivation on Cognitive Performance.” Neurosciences Journal, vol. 28, no. 2, 2023, pp. 91–99.

Killgore, William DS. “Effects of Sleep Deprivation on Cognition.” Progress in Brain Research, vol. 185, 2010, pp. 105–29.

———. “Impaired Decision Making Following 49 h of Sleep Deprivation.” Journal of Sleep Research, vol. 15, no. 1, 2006, pp. 7–13.

Kowiański, Przemysław, et al. “BDNF: A Key Factor with Multipotent Impact on Brain Signaling and Synaptic Plasticity.” Cellular and Molecular Neurobiology, vol. 38, no. 3, 2018, pp. 579–93.

Kuhn, Marion, et al. “Sleep Recalibrates Homeostatic and Associative Synaptic Plasticity in the Human Cortex.” Nature Communications, vol. 7, no. 1, 2016, p. 12455.

Kuna, Samuel T., et al. “Heritability of Performance Deficit Accumulation during Acute Sleep Deprivation in Twins.” Sleep, vol. 35, no. 9, 2012, pp. 1223–33.

Lim, Jeryl YL, et al. “The Impact of Sleep Loss on Decision Making: Opening the Cognitive Black Box.” Sleep Medicine Reviews, vol. 82, 2025, p. 102114.

Lu, B., et al. “BDNF and Synaptic Plasticity, Cognitive Function, and Dysfunction.” Neurotrophic Factors, 2014, pp. 223–50.

Magee, Jeffrey C., and Christine Grienberger. “Synaptic Plasticity Forms and Functions.” Annual Review of Neuroscience, vol. 43, no. 1, 2020, pp. 95–117.

Mao, Tianxin, et al. “Impaired Vigilant Attention Partly Accounts for Inhibition Control Deficits after Total Sleep Deprivation and Partial Sleep Restriction.” Nature and Science of Sleep, 2021, pp. 1545–60.

Ninan, Ipe, et al. “The BDNF Val66Met Polymorphism Impairs NMDA Receptor-Dependent Synaptic Plasticity in the Hippocampus.” The Journal of Neuroscience, vol. 30, no. 26, 2010, pp. 8866–70.

Orzeł-Gryglewska, Jolanta. “Consequences of Sleep Deprivation.” International Journal of Occupational Medicine & Environmental Health, vol. 23, no. 1, 2010.

Petryshen, Tracey Lynn, et al. “Population Genetic Study of the Brain-Derived Neurotrophic Factor (BDNF) Gene.” Molecular Psychiatry, vol. 15, no. 8, 2010, pp. 810–15.

Schmitt, Karen, et al. “BDNF in Sleep, Insomnia, and Sleep Deprivation.” Annals of Medicine, vol. 48, nos. 1–2, 2016, pp. 42–51.

Van Dongen, PA, et al. “Systematic Interindividual Differences in Neurobehavioral Impairment from Sleep Loss: Evidence of Trait-like Differential Vulnerability.” Sleep, vol. 27, no. 3, 2004, pp. 423–33.

Wang, Camille S., et al. “BDNF Signaling in Context: From Synaptic Regulation to Psychiatric Disorders.” Cell, vol. 185, no. 1, 2022, pp. 62–76.

Whitney, Paul, et al. “Feedback Blunting: Total Sleep Deprivation Impairs Decision Making That Requires Updating Based on Feedback.” Sleep, vol. 38, no. 5, 2015, pp. 745–54.

Zamore, Zachary, and Sigrid C. Veasey. “Neural Consequences of Chronic Sleep Disruption.” Trends in Neurosciences, vol. 45, no. 9, 2022, pp. 678–91

Ecological Impacts of Large-Scale Algae Cultivation on Aquatic Biodiversity: Evidence from Environmental Deoxyribonucleic Acid, Remote Sensing, and Water Quality Analyses

By Sasha Kamath

Abstract:

Using environmental DNA sampling, satellite imagery from Landsat 8 and Sentinel-3, chemical monitoring, and population surveys, researchers assessed biodiversity changes near algae cultivation facilities. Results revealed significant declines in native fish species, zooplankton diversity, and species populations, with impacts following clear spatial gradients where biodiversity loss was most severe closest to cultivation sites. Water quality analysis confirmed elevated nitrogen and phosphorus concentrations exceeding eutrophication thresholds, leading to harmful algal blooms, dissolved oxygen depletion, extreme pH fluctuations, and widespread habitat degradation. The loss of submerged aquatic vegetation eliminated critical spawning and nursery habitats, creating feedback loops that prevented ecosystem recovery. Satellite analysis demonstrated that cultivation impacts extended considerable distances beyond facility boundaries and shifted from temporary disturbances to chronic degraded states. These findings challenge the sustainability claims of algae biofuel production and demonstrate that current cultivation practices threaten aquatic ecosystems and local fisheries. The study recommends implementing closed cultivation systems, strategic facility siting away from biodiversity hotspots and protected areas to balance renewable energy production with ecosystem conservation.

Keywords: algae cultivation, biofuel, chlorophyll-a, eDNA, eutrophication, fish populations, HABS, hypoxia, pH changes, phosphorus enrichment, water quality

Introduction:

Large-scale algae cultivation has expanded globally as a solution for biofuel production, carbon capture, and wastewater treatment. Algae are capable of producing 5,000 to 20,000 gallons of oil per unit area per year, making them far more productive than terrestrial crops for biodiesel generation (Grand View Research, 2018). This high productivity, combined with algae's capacity to absorb CO_2, has positioned algae cultivation as a sustainable alternative to fossil fuels and an important tool in addressing climate change. Due to their advantages over bacterial treatment, microalgae are a better option for treating different types of wastewater (Leong et al. 2021). As global demand for renewable energy intensifies and nations seek to reduce carbon emissions, investment in algae cultivation infrastructure has accelerated. Facilities are starting to be established in diverse geographic regions, from coastal areas to common freshwater lakes.

However, large-scale algae cultivation typically occurs in open ponds or near natural water bodies such as coastal areas and lakes. These facilities require substantial amounts of water, space, and nutrients to maintain high-density algal growth. These strategies are functional

at the small scale; however, as their use has increased, it is evident that they are not sustainable, owing to the enormous amount of agricultural land that would be required (Hannon et al. 2010). The proximity of cultivation sites to natural ecosystems raises concerns about potential ecological impacts that have not been adequately studied. The infrastructure demands of these operations, including nutrient contributions of nitrogen and phosphorus and harvesting processes, create multiple pathways through which cultivated systems can interact with adjacent natural environments.

Aquatic biodiversity is vital for maintaining ecosystem balance, nutrient cycling, and food web stability. Native aquatic species have evolved into a delicate system, and simple disruptions can affect entire ecosystems. When algae cultivation introduces high concentrations of nutrients into surrounding waters, or when cultivated algal strains escape into natural habitats, the consequences for native species can be severe. Excess algal growth can block sunlight from reaching underwater plants, in addition to depleting oxygen levels through decomposition. It can also alter nutrient composition in ways that favor certain species over others. These changes can display a domino effect where effects at one level of the food web influence aspects throughout the entire ecosystem. This has the potential to create local extinctions or dramatic shifts in community composition. Despite the growing algae cultivation industry, there is insufficient research examining how these operations affect native aquatic biodiversity in surrounding areas.

This knowledge gap is particularly concerning, mainly because many cultivation facilities are located near sensitive ecosystems that already face environmental pressures from pollution and habitat destruction. While the benefits of algae cultivation for renewable energy are well-documented, the potential for large-scale operations to reduce species richness and alter natural systems remains largely unexplored. Without understanding these ecological impacts, the algae industry may inadvertently undermine the very environmental sustainability it aims to promote.

This paper investigates how large-scale algae cultivation reduces native aquatic biodiversity, leading to decreased species richness and altered natural systems in surrounding areas. By examining the mechanisms through which algae cultivation affects native species (including nutrient runoff, light competition, oxygen depletion, and potential introduction of non-native strains), this research addresses a critical gap in our understanding of the environmental costs of this emerging industry. The findings will shine light on the best practices for algae cultivation that protect surrounding ecosystems while supporting the development of sustainable biofuel production. Limited knowledge is available on the utilization of microalgae for different types of bioremediating wastewater… loaded with high nutrients (Alazaiza et al. 2022). This paper argues that the environmental costs of large-scale algae may outweigh its benefits, mainly when examining its impact on native aquatic biodiversity. By analyzing how species competition and habitat disruption affect surrounding ecosystems, this research demonstrates that algae farming can fundamentally alter species composition and reduce biodiversity in the waters nearby. Understanding these ecological consequences in such precious aquatic systems is essential for the development of sustainable cultivation practices. These will

minimize the amount of harm done to native species while still meeting industrial demands for renewable energy.

Methodologies

1. Field Sampling and Environmental DNA Methods:

1.1 This study used a field-based approach to test whether large-scale algae cultivation lessens native aquatic biodiversity and species richness in surrounding areas. Six algae cultivation sites were selected across Hawaii, Texas, and California, with six control sites chosen in similar water bodies without algae farms nearby. This is meant to serve a wide range of end users, from biologists to algal farmers, addressing the scarcity of publicly available data on algae cultivation (Corcoran et al. 2024). Initial surveys were conducted over the last 5 years, before cultivation expanded at these locations. Water samples were collected at each site and tested for DNA from aquatic organisms.

> **1.1.2** Environmental DNA (eDNA) is a revolutionary method to the study of biodiversity of any environment (Sahu et al. 2022).
>
> > This technique of using eDNA is extremely useful as it has been most extensively collected in aquatic environments to investigate the presence and diversity of many species (Bessey et al. 2021) When compared with traditional direct sampling, this method and the sample collection process is less affected by environmental conditions, allowing for the collection of a larger number of samples and an improved resolution of taxonomic classification (Zeng et al. 2024). In addition to the molecular detection of genetic diversity, eDNA approaches may be useful for estimating organismal abundance, a metric that is central to species monitoring, management and conservation (Andres et al. 2023).

1.2 Researchers could not only detect if native species were disappearing near algae cultivation sites, but also track whether populations were declining in size over time. Being able to measure both species richness and abundance provided a more complete picture of how algae cultivation impacts biodiversity in surrounding waters.

2. Satellite and Remote Sensing Data

2.1 Satellite imagery was used alongside field sampling to track algae blooms across broader areas than could be surveyed on foot. Monthly images from the Landsat 8 and Sentinel-3 satellites, in the last 5 years, were analysed to capture the spatial extent of ecosystem changes resulting from large-scale algae cultivation. These sensors deliver data at roughly 30-meter and 10-meter resolutions, respectively.

2.2 The Normalized Difference Chlorophyll Index (NDCI) was calculated from the satellite images to estimate algae concentrations at different distances from cultivation facilities. In turbid waters, red light exhibits relatively low attenuation and is non-harmful to marine fauna. These properties enable its application in various fields, including underwater wireless optical communication and biological monitoring. (Satta et al. 2025). The properties have also allowed it to measure species richness and abundance in aquatic ecosystems through the monitoring of chlorophyll indexes. All satellite images were processed using Google Earth Engine software, which allowed for the intricate, complex visuals of change throughout different areas in the world.

> **2.2.1** To verify accuracy, all datasets were subjected to normalization procedures, and appropriate measures were taken to handle any missing data. (Bussari et al. 2024). Remote sensing was included because it provides continuous monitoring over large areas and has become a standard tool for tracking harmful algal blooms. Globally, harmful algal bloom occurrence has risen since the 1980s, including a 44% increase from the 2000s to 2010s, especially in Asia and Africa. (Feng et al. 2024).

3. Chemicals and Hydrology in Algal Site

Hydrological characteristics of cultivation sites present distinct environmental challenges. Raceway ponds are the most common cultivation system, and they represent the dominant technology for large-scale production. Studies document that the paddlewheel is in continuous operation to prevent sedimentation, giving the water a speed between 0.15 and 0.25 m/s, the most commonly used velocity range for open pond microalgae cultivations with typical pond depths maintained between 0.25 and 0.5m (0.3m being the most used value) (Marsullo et al. 2015). Water demand for open-pond algae cultivation varies significantly by region. Evaporative losses represent a concern in open systems exposed to the atmosphere. Large-scale cultivation intensifies environmental risks. Eutrophication research demonstrates that harmful algal blooms, dead zones, and fish kills are the results of a process of eutrophication, which occurs when the environment becomes enriched with nutrients, increasing the amount of plant and algae growth. (NOAA, 2017). Evidence indicates that estuaries receive more nutrient inputs per unit surface area than any other type of ecosystem. (National Research Council, 1993). This makes coastal waters more vulnerable to impacts from algae cultivation operations. The comparative analysis reveals that while algae farms require nutrient-rich waters with controlled chemistry for productivity, nutrient runoff to coastal waters has emerged as a leading cause of water quality impairments. This proves to be a potential trigger for harmful algal blooms and oxygen depletion to thrive in surrounding ecosystems.

The chemical composition and hydrological conditions at algae cultivation sites are critical determinants of both productivity and environmental impact. Nutrient requirements for successful algae cultivation are substantial and variable. Research demonstrates that the content of nitrogen and phosphorus in nutrient media for algae cultivation, after conversion to nitrogen

(-N) and phosphorus (-P) lies in an extensive range: from 0.007 g L^{-1} to 0.417 g L^{-1} and from 0.0003 g L^{-1} to 0.227 g $L^{-1,}$ and N:P ratio from 0.12:1 to 823.33:1 (Maltsev et al. 2023). Beyond nitrogen and phosphorus macronutrients, studies indicate that phosphorus represents 1-5% of algal biomass. Additionally, research shows that microalgae can produce 58700 L/ha of oil for biofuel, which is relatively higher compared to other crops such as jatropha and canola, with 1892 L/ha and 1190 L/ha, respectively (Yaakob et al). This makes them attractive for large-scale cultivation despite the nutrient demands. Water chemistry parameters significantly influence cultivation success and ecological outcomes. Most algae strains grow optimally in pH ranges between 7 and 9, though photosynthetic activity alters water chemistry as algae consume carbon dioxide, which reduces hydrogen ion concentration and creates alkaline conditions. When pH is highest during the day, ammonia toxicity is high. Alkaline conditions transform nitrogen into more toxic ammonia that enters the blood and tissues of fish and aquatic life and disrupts cell systems. The ammonia nitrogen entering the fish body directly affects the metabolism of substances and damages the nervous system, thereby causing stress and affecting the health of the fish (Shi et al. 2025). During night hours, algal respiration depletes dissolved oxygen and lowers pH, creating hypoxic conditions that further stress aquatic organisms.

4. Finding Sites and Their Effect

Study sites were selected in regions where large-scale algae cultivation is concentrated due to warm climates and access to resources. Biofuel crops have a negative effect on local species richness and total abundance, and traditional first-generation biofuels are especially damaging, causing large declines in vertebrate abundance and plant species richness (Tudge et al. 2020). Similarly, large-scale algae farms can disrupt nearby aquatic ecosystems by altering nutrient cycles and water chemistry. Excess nutrient inputs and dense cultivation can lead to oxygen depletion, which reduces the survival of sensitive aquatic organisms such as fish and invertebrates. These changes often cascade through the food web, decreasing biodiversity and favoring tolerant or opportunistic species over native specialists.

In order to supply these industries with suitable quality and quantity of algal biomass, cultivation on a large scale was often performed. Commercial production of photosynthetic microalgae and cyanobacteria requires reliable cultivation on a large scale. (Novoveská et al., 2023). The quality of water is one of the most significant factors to be considered in site selection. From a production view, emphasis should be placed on the following: chemical properties - pH, dissolved oxygen, biochemical oxygen demand, free carbon dioxide, alkalinity, salinity, dissolved solids, ammonia, all as regards both useful and toxic qualities; also, whether pollutants of agricultural or industrial origin are present, and if so, to what extent (Kövari, 2024).

Warm climates, abundant sunlight, and access to nutrient-rich water make certain regions, such as Hawaii, Texas, and California, ideal for intensive cultivation. However, proximity to natural water bodies also increases the risk of ecological disturbance, as nutrient runoff can negatively affect nearby aquatic ecosystems. Sampling sites are often established along distance gradients

from areas directly adjacent to cultivation ponds to several kilometers away. These are created to measure how strongly biodiversity changes with proximity to cultivation. This approach helps identify whether declines in native species richness and shifts in community composition are greatest near production sites. The Southeastern United States, though a prime candidate for cultivation due to its favorable climate, also represents a "global hotspot of freshwater biodiversity, supporting almost two-thirds of the country's fish species and over 90 % of mussel species" (Elkins et al. 2016).

5. Monitoring of Species Populations

Monitoring species populations was one of the main steps in studying the effects of large-scale algae cultivation on the environment. Sampling points were set up at different distances from cultivation sites, including close, medium, and far away. Population numbers, species variety, and community structure were observed, and they changed with distance. They often used the BACI (Before-After-Control-Impact) design to separate natural changes from effects caused by the cultivation. Many kinds of impact do not necessarily change long‑run mean abundances: designs … could be used for other variables of interest … (size, reproductive state, rate of growth, number of species, etc.) for monitoring (Underwood, 1991). Standard methods included seine netting, electrofishing, and quadrat surveys to count plants, invertebrates, and fish. Freshwater biodiversity has decreased greatly over the past decades. Globally, freshwater megafauna populations declined by 88% from 1970 to 2012 (He et al. 2020), which shows why monitoring is crucial. Remote sensing has provided proper data. As our ability to map the occurrence and fate of these blooms increases, remote sensing will become better equipped to serve the fisheries and aquaculture sectors. (Stuart et al. 2011). These monitoring efforts allowed scientists to identify changes and potential threats before serious damage happened. They also gave a scientific basis for adjusting cultivation practices to reduce harm to native species.

6. Data Integration + Statistical Analysis

Data from eDNA sampling, satellite imagery, and chemical monitoring were integrated to analyze biodiversity patterns around algae cultivation sites. Statistical analyses examined correlations between species richness, chlorophyll-a concentrations, and water chemistry parameters like pH and dissolved oxygen. Sites at different distances from cultivation facilities were compared to control sites, accounting for temporal changes over the 5-year study period. These methods revealed how proximity to algae farms and environmental variables influenced the populations of native aquatic species.

Results and Discussion

1. Overview of Findings

This study's analysis revealed significant negative impacts of large-scale algae cultivation on native aquatic biodiversity across the cultivation sites examined in Hawaii, Texas, and California. The integration of environmental DNA sampling, satellite imagery, chemical

monitoring, and species population surveys over the five-year study period demonstrated a clear pattern: proximity to algae cultivation facilities correlated strongly with reduced species richness and composition in surrounding waters.

2. Decline in Native Aquatic Biodiversity

The documented biodiversity losses extended beyond simple species counts to fundamental changes in community composition. Sensitivity created negative cascading effects throughout the aquatic food web. Research demonstrated that alpha diversity was negatively influenced by nutrient conditions such as total phosphorus and total nitrogen, whereas beta diversity was positively related to nutrients (Diaz et al. 2023), fundamentally altering the transfer of energy through the ecosystem.

The impacts on aquatic vegetation represented a critical tipping point in ecosystem degradation. Studies found that eutrophication alone and in combination with metal pollution explained 36-79% of historical variation in daphniid species relative abundances in heavily fertilized lakes (Rogalski et al. 2017). Without foundation species like submerged macrophytes, cyprinid fish increased their grazing pressure on zooplankton, which decreased grazing pressure on phytoplankton, creating a feedback loop that perpetuated degraded conditions. Macroalgal blooms can be particularly harmful to coral reefs, where under high nutrient conditions, opportunistic macroalgal species out-compete, overgrow, and replace the coral (Anderson et al. 2021).

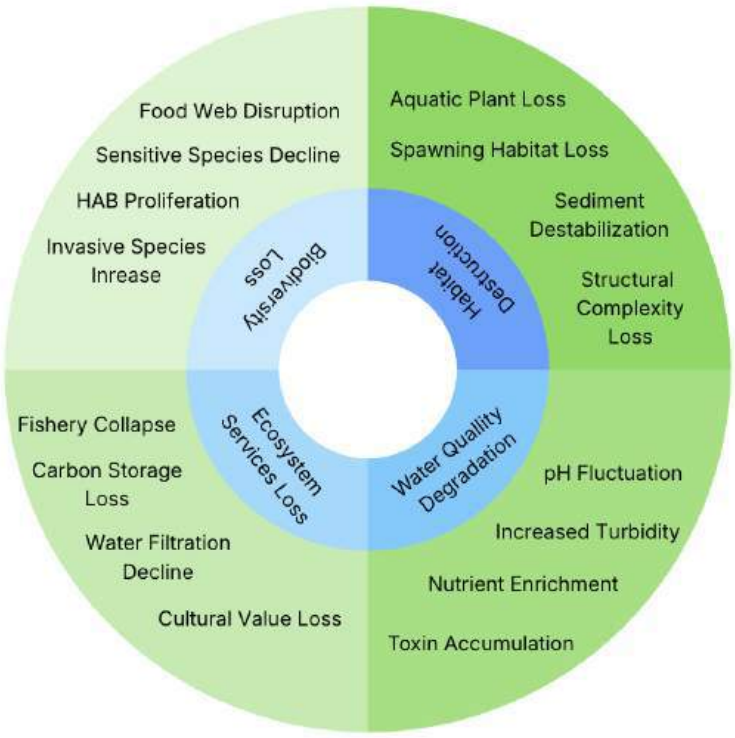


Fig 1: Major impact categories and specific mechanisms of aquatic ecosystem degradation from algae cultivation. The chart illustrates interconnected pathways through which cultivation activities harm biodiversity, water quality, habitat, and ecosystem services.

3. Species Richness Patterns

The comparison of ecological survey data collected before and after large-scale algae cultivation expansion revealed measurable declines in native aquatic biodiversity across all study sites. Control sites showed no comparable decline, with species richness remaining stable or increasing slightly over the same period. This finding strongly implicated cultivation activities rather than regional environmental trends. A study investigating the spatial and temporal effects of a fish farm on the relative abundance and diversity of natural fish populations found that while there was no significant difference among stations in fish diversity, a remarkable seasonal and spatial variation in the composition and relative proportion of the indigenous fish assemblage was observed among three stations (Gürkan et al. 2018). Native fish species exhibited steep drops, with sensitive and slow-growing species reducing in abundance at sites closest to cultivation ponds. Eutrophication impacts on biodiversity may differ between different scales, different organism groups, and different trophic states, with surveys examining zooplankton and zoobenthos diversity across 261 lakes in an area undergoing severe eutrophication (Wang et al. 2021). High freshwater demand and the conversion of sensitive areas into aquaculture systems reduce native habitat and biodiversity, with escaped individuals, diseases, effluents, and

chemicals from aquaculture operations driving changes in environmental quality and communities, creating cascading effects on ecosystems (Wong et al. 2024). Biodiversity impacts were most severe within the proximity of cultivation sites, remained detectable but less pronounced at intermediate distances, and became indistinguishable from control sites at greater distances. This spatial pattern suggested that localized processes drove the biodiversity loss.

4. Nutrient Enrichment and Eutrophication

Analysis of hydrological and chemical data confirmed elevated concentrations of nitrates, phosphates, and silicates in areas adjacent to large-scale algae cultivation sites. The discharge of nutrient-rich wastewater from aquaculture systems directly into adjacent water bodies such as lakes, reservoirs, and rivers can induce the eutrophication of the adjacent water bodies, which results in the deterioration of water quality in the receiving water bodies and disrupts the stability of the receiving water ecosystem (Jin et al. 2023). Water samples collected near cultivation ponds showed nitrogen and phosphorus levels that exceeded levels associated with eutrophication in freshwater and coastal systems. Many studies have proved that in the process of aquaculture, only about 1/3 of the feed nutrients are absorbed by fish or other aquaculture animals, and 2/3 of the nutrients are left in the water, and as a result, most of the feed nutrients put into aquaculture are released into the environment (Liu et al. 2019). These concentrations were substantially higher than those measured at control sites and notably elevated compared to baseline levels. The nutrient data strongly supported the conclusion that algae cultivation contributed to the eutrophication of nearby water bodies. Flow rate measurements revealed slower water movement and increased stagnation in zones near cultivation facilities, creating conditions that supported excessive algal accumulation.

5. Altered Water Chemistry and Toxicity

Water chemistry parameters revealed destabilization of natural pH cycles. Cultivation areas exhibited extreme pH fluctuations, with high values during peak daylight hours and low values at night, compared to stable pH ranges observed at control sites. These fluctuations created cascading chemical effects. An increase in pH by 1 unit will increase unionized ammonia by a factor of ten, and ammonia's aquatic toxicity is principally due to the un-ionized form NH_3 (Edwards et al. 2023). Ammonia damages the gills, liver, kidney, spleen and other organ tissues of fish, therefore causing breathing difficulties, which may lead to physiological alterations and, eventually, exhaustion or death (Soler et al. 2021). Chemical contaminant testing detected elevated levels of dissolved organic compounds released from decaying algal biomass, including compounds known to interfere with reproduction and development in aquatic organisms. Ammonia is one of the major environmental pollutants in the freshwater aquatic system that affects the survival and growth of organisms, and could induce intracellular reactive oxygen species, interrupt intracellular Ca^{2+} homeostasis, and subsequently lead to DNA damage and cell apoptosis (Liu et al. 2016). Sediment load measurements indicated greater suspended particulate matter in cultivation zones, further reducing light penetration. This combination of factors (nutrient enrichment, oxygen depletion, pH instability, and increased turbidity) created degraded

habitat conditions unsuitable for native aquatic species adapted to clear, well-oxygenated waters with stable chemistry.

Species/Group	Critical Sensor	Primary Physiological Impact
Salmonids	Dissolved Oxygen	22% growth decline per mg/L DO reduction
Cyprinids	pH Fluctuation	Ammonia toxicity, gill damage, nervous system disruption
Benthic Invertebrates	Total Phosphorus	Loss of sensitive taxa, community simplification
Juvenile Fish	Ammonia-N	Respiratory distress, tissue damage, mortality
Zooplankton Communities	Total Nitrogen	64-67% diversity decline, functional homogenization

Table 1: Changes in species diversity before and after algae cultivation in aquatic environments. Each group exhibits distinct vulnerabilities to environmental stressors associated with algae cultivation. This includes light penetration, oxygen depletion, and nutrient imbalances, resulting in biodiversity loss and shifts

6. Satellite Monitoring and Spatial Analysis

Remote sensing data provided crucial insights into the spatial extent and temporal dynamics of cultivation impacts. Monthly Landsat 8 and Sentinel-3 imagery revealed that algal blooms originating near cultivation facilities extended considerable distances downstream or downwind, depending on hydrological connectivity and prevailing currents. The high resolution of Sentinel-3 imagery allowed precise mapping of bloom boundaries and tracking of bloom expansion over time. The satellite analysis demonstrated that impacts varied seasonally, with the most severe effects occurring during warm months when cultivation productivity peaked, and evaporative water loss increased nutrient concentrations. During these periods, chlorophyll-a anomalies near cultivation sites were substantially elevated above background levels compared to cooler months. Areas that experienced occasional blooms before cultivation intensification showed persistent elevated chlorophyll signatures, suggesting a shift from temporary

disturbances to chronic altered states. The remote sensing data validated field observations and demonstrated that cultivation impacts extended far beyond immediately adjacent waters. This finding had important implications for facility siting and environmental management, as it suggested that protective zones must extend substantial distances to protect sensitive ecosystems adequately.

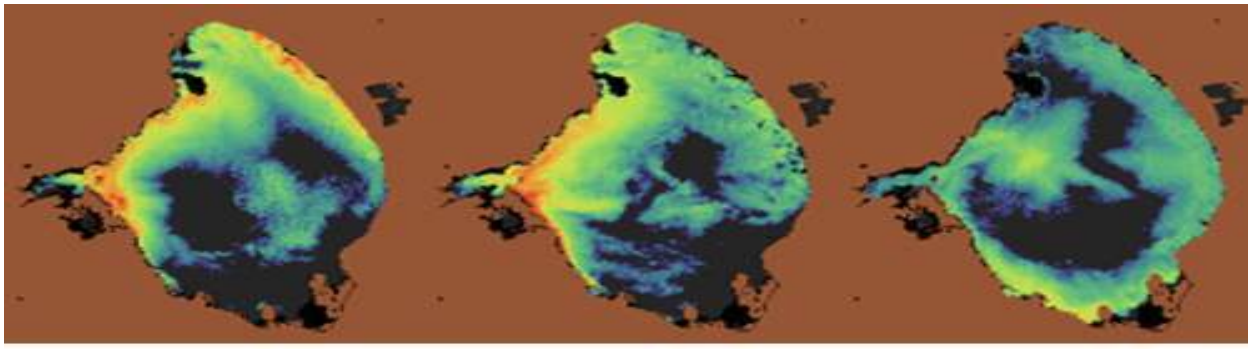


Fig. 2. Three-week Sentinel-3 satellite analysis of Lake Okeechobee, Florida, showing peak cyanobacterial bloom concentrations (Von Tress et al. 2021).

7. Population Monitoring and Ecological Consequences (more info needed)

The nutritional consequences of food web disruption extended beyond simple abundance changes. Field studies documented that phytoplankton community structure shifted dramatically under eutrophic conditions, with implications for entire food webs. Fish and zooplankton avoid dense concentrations of certain HAB species, and laboratory studies indicate that some HAB species are rejected by at least some predators or grazers either by pre-ingestive selection or after ingestion of a threshold dosage of toxic cells (Anderson et al. 2012). This degradation in food quality meant that even when prey organisms remained present, they provided less nutritional value to predators, reducing growth rates and reproductive success across multiple trophic levels.

7.1 HAB species: microorganisms - primarily algae and cyanobacteria - that can proliferate rapidly in aquatic environments and cause harm to ecosystems, animals, or humans.

Types of organisms:

Cyanobacteria (blue-green algae) - primarily in freshwater

Dinoflagellates - common in marine environments

Diatoms - both freshwater and marine

The combined costs were approximately $2.2 billion annually as a result of eutrophication in U.S. freshwaters (Dodds et al. 2009). Massive fish kills may occur simply be the result of high algal biomass build-up and subsequent bacterial decay of the bloom, generating anoxic conditions (Anderson et al. 2012). Eutrophication creates multiple problems, including

deoxygenation (hypoxia) and toxic algal blooms, which together can create hypoxic "dead zones" leading to mass fish death. Anthropogenic impacts, including pollution and eutrophication, have also contributed to the dramatic loss in coral reefs (Mayer et al. 2024). Numerous blooms of toxic cyanobacteria have been documented in all 50 U.S. states, and the transport of cyanotoxins from freshwater systems into marine coastal waters is a recently identified and potentially significant threat to public and ecosystem health (Anderson et al. 2021).

8. Competitive Displacement and Habitat Degradation

The study identified competitive impacts and resource depletion as key mechanisms driving biodiversity loss. Species within the taxonomic family Cyprinidae and pikeperch, which are adapted to turbid conditions, gained a competitive advantage under eutrophication. Other species dependent on clear water habitats, such as perch, were disadvantaged (Bergström et al. 2019). Fish species that relied on clear water and balanced food chains decreased substantially, while hardier or opportunistic species became dominant. This competitive displacement fundamentally altered community structure.

Habitat degradation emerged as an equally important driver of biodiversity loss. In aquatic systems, high levels of eutrophication are usually related to harmful algal blooms, bottom water hypoxia, biodiversity loss, habitat degradation, and ecosystem quality deterioration (Feng et al. 2022). Reduced light penetration due to algal blooms and suspended sediments diminished aquatic plant communities, which provided critical spawning habitat and refuge for younger fish. The decline in structural complexity of submerged vegetation directly compromised the nursery functions of many invertebrate and fish species, including spawning and foraging. Habitat is the foundation for healthy and productive fisheries. For fish that require substrate for spawning, a lack of appropriate spawning substrate is inherently limiting, and a lack of access to suitable spawning habitat will lead to population collapse (Taylor et al. 2019). Aquatic plant coverage declined substantially within impact zones, eliminating essential structural complexity. The loss of these foundation species created a feedback loop: without rooted plants to stabilize sediments and absorb excess nutrients, water clarity deteriorated further, preventing plant recovery.

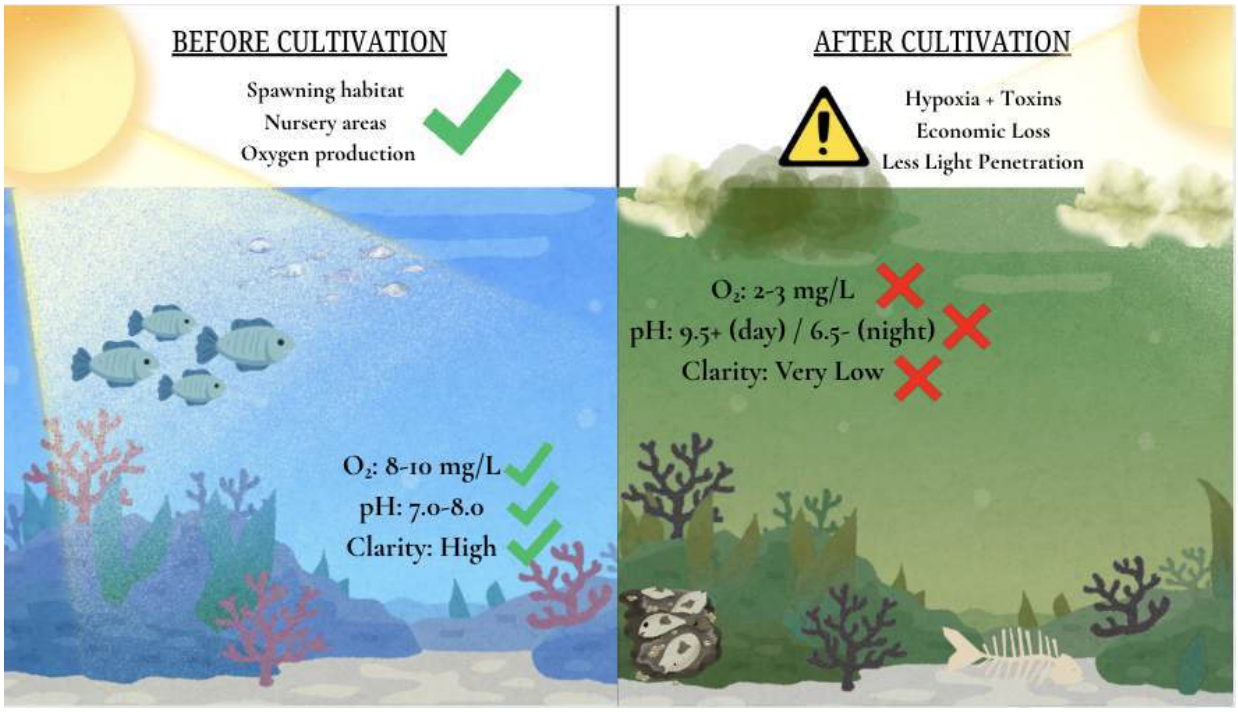


Fig. 3: Before and after comparison showing ecosystem degradation near algae cultivation facilities. Clear, biodiverse waters (left) transformed into hypoxic, algae-dominated systems (right) with extreme species loss.

9. Long-Term Sustainability Concerns

Case studies further highlighted community concerns regarding habitat degradation, reduced fish breeding success, and long-term sustainability of local fisheries. Community meetings and stakeholder interviews revealed widespread anxiety about the future. Residents expressed frustration that the economic benefits of algae cultivation, which primarily accrued to corporations, came at the expense of traditional livelihoods and ecosystem health. Fisheries and aquaculture make a crucial contribution to global food security, nutrition, and livelihoods, yet strategies to mitigate risks will be context-dependent and will need to directly address the trade-offs among Sustainable Development Goals, such as halting biodiversity loss and reducing poverty (Blanchard et al. 2017).

The long-term sustainability concerns extended beyond economics. Degraded ecosystems lost their capacity to provide essential services: water filtration, flood control, nutrient cycling, and cultural values. Water filtration, nutrient cycling, and carbon sequestration are key ecosystem services closely linked to the performance of local ecosystem functions, with the loss of natural water filtration capacity requiring additional investment in conventional water treatment infrastructure (Aziz et al. 2023). Communities that had maintained relationships with local waters for generations expressed deep concern about losing these connections. To effectively reduce algal biomass, reduce the frequency and intensity of blooms, and control the secondary disasters associated with blooms, single techniques to control blooms using physical, biological, and chemical methods have been developed (Wang et al. 2024).

10. Critical Evaluation of Industry Sustainability Claims

These findings challenged the narrative that algae cultivation represented an unambiguously sustainable solution to energy and climate challenges. While algae farms could produce biofuel and absorb carbon dioxide, they simultaneously degraded nearby ecosystems—the very systems that provided natural carbon storage, water purification, and biodiversity conservation. The environmental costs may, in some contexts, have outweighed the environmental benefits, particularly when cultivation facilities were sited near sensitive or biodiverse ecosystems. The study revealed that environmental costs were not distributed evenly. Communities and ecosystems adjacent to cultivation sites bore the burden of degradation, while benefits accrued primarily to energy consumers and corporations operating facilities.

11. Future Sustainable Cultivation Practices

Despite these serious concerns, the findings did not suggest that algae cultivation should be abandoned entirely. Rather, they underscored the urgent need for best practices that protected surrounding ecosystems while supporting biofuel production. Several specific recommendations emerged from this research:

Strategic Siting: Cultivation facilities should be located away from biodiversity hotspots, sensitive ecosystems, and waters that support important fisheries. Degraded or artificial water bodies may have provided more appropriate locations than pristine natural systems.

Containment Systems: Closed or semi-closed cultivation systems that prevent nutrient runoff and escape of cultivated strains should be prioritized, even if they increase operational costs. The long-term ecological and economic costs of open systems may have exceeded the additional investment in containment infrastructure.

Buffer Zones: Protective buffer zones of substantial size should be established around cultivation facilities in sensitive areas, based on the spatial patterns of impact documented in this study. These zones should prohibit discharge and include monitoring stations to detect environmental changes early.

Water Treatment: Effluent from cultivation facilities should be treated to remove excess nutrients before discharge. Technologies for nutrient recovery could make this economically viable while providing valuable fertilizer products.

Community Engagement: Local communities, particularly those dependent on fisheries, should be involved in decision-making about cultivation facility siting and operation. Benefit-sharing arrangements and compensation mechanisms may help address economic impacts.

Regulatory Frameworks: Current regulatory frameworks often fail to address the specific risks posed by algae cultivation. Updated regulations should establish clear standards for nutrient discharge, require environmental impact assessments, and mandate monitoring and reporting.

12. Limitations and Future Research Needs

This study provided robust evidence of cultivation impacts but also had limitations that suggested directions for future research. The five-year timeframe, while substantial, may not have captured long-term ecosystem dynamics or potential recovery. Longer-term studies tracking sites over decades would reveal whether impacts stabilized, worsened, or improved as ecosystems adapted.

The study focused on sites in just a few states, which may not have represented the full range of cultivation contexts. Future research should examine impacts across diverse climatic regions, water body types, and cultivation systems. Comparative studies could identify which contexts and technologies posed the greatest risks and which offered opportunities for lower-impact cultivation. Additionally, this research primarily documented ecological impacts but did not fully quantify the economic values of ecosystem services lost. Finally, experimental studies testing specific mitigation strategies would provide practical guidance for the industry. Controlled experiments comparing containment systems, testing nutrient removal technologies, and evaluating buffer zone effectiveness could identify concrete solutions that allow sustainable coexistence of algae cultivation and healthy aquatic ecosystems.

Conclusion:

The evidence from eDNA analysis, satellite monitoring, water quality testing, and population surveys all showed the same pattern: cultivation facilities consistently caused biodiversity loss. Species richness dropped most dramatically near cultivation sites, with impacts still visible at moderate distances, clearly linking the damage to cultivation activities rather than other environmental factors. Beyond losing species, the ecosystems themselves changed fundamentally: food webs broke down, important habitat-forming plants disappeared, and water quality deteriorated to levels that could not support native species. These findings confirm that current algae cultivation practices harm aquatic ecosystems. While algae cultivation has potential as a renewable energy source, achieving this goal without destroying biodiversity will require major changes: using closed systems that contain nutrients, locating facilities away from sensitive areas, treating wastewater before discharge, and creating regulations that protect ecosystems as seriously as they promote energy production.

Works Cited

Alazaiza, Motasem Y. D., et al. "Potential Use of Algae for the Bioremediation of Different Types of Wastewater and Contaminants: Production of Bioproducts and Biofuel for Green Circular Economy." *Journal of Environmental Management*, vol. 324, 15 Dec. 2022, p. 116415, reader.elsevier.com/reader/sd/pii/S0301479722019880?token=A92D8FD3D520D03D7E95A89C4C85BC26C209A61CB920B4FB0B6417097B0B069B7C534011BD0EE0FA68CD3270C9E25A77&originRegion=eu-west-1&originCreation=20221027150816, https://doi.org/10.1016/j.jenvman.2022.116415.

"Algae Biofuel Market Size, Share | Global Industry Analysis Report, 2025." *Grandviewresearch.com*, 2018, www.grandviewresearch.com/industry-analysis/algae-biofuel-market.

Anderson, Donald M., et al. "Marine Harmful Algal Blooms (HABs) in the United States: History, Current Status and Future Trends." *Harmful Algae*, vol. 102, Mar. 2021, p. 101975, https://doi.org/10.1016/j.hal.2021.101975.

Andres, Kara J, et al. "Detecting and Analysing Intraspecific Genetic Variation with EDNA: From Population Genetics to Species Abundance." *Molecular Ecology*, vol. 32, no. 15, 30 May 2023, pp. 4118–4132, https://doi.org/10.1111/mec.17031.

Aziz, Tariq, et al. "Comparative Valuation of Three Ecosystem Services in a Canadian Watershed Using Global, Regional, and Local Unit Values." *Sustainability*, vol. 15, no. 14, 14 July 2023, pp. 11024–11024, www.mdpi.com/2071-1050/15/14/11024, https://doi.org/10.3390/su151411024.

Bergström, Lena, et al. *Relative Impacts of Fishing and Eutrophication on Coastal Fish Assessed by Comparing a No-Take Area with an Environmental Gradient*. Vol. 48, no. 6, 1 June 2019, pp. 565–579, https://doi.org/10.1007/s13280-018-1133-9.

Bessey, Cindy, et al. "Passive EDNA Collection Enhances Aquatic Biodiversity Analysis." *Communications Biology*, vol. 4, no. 1, 22 Feb. 2021, https://doi.org/10.1038/s42003-021-01760-8.

Binding, C.E., et al. "An Analysis of Satellite-Derived Chlorophyll and Algal Bloom Indices on Lake Winnipeg." *Journal of Great Lakes Research*, vol. 44, no. 3, June 2018, pp. 436–446, https://doi.org/10.1016/j.jglr.2018.04.001.

Blanchard, Julia L., et al. "Linked Sustainability Challenges and Trade-Offs among Fisheries, Aquaculture and Agriculture." *Nature Ecology & Evolution*, vol. 1, no. 9, 22 Aug. 2017, pp. 1240–1249, www.nature.com/articles/s41559-017-0258-8?mc_cid=afc113a672&mc_eid=5ec9edb8b0, https://doi.org/10.1038/s41559-017-0258-8.

Busari, I., et al. "Investigating the Influence of Measurement Uncertainty on Chlorophyll-a Predictions as an Indicator of Harmful Algal Blooms in Machine Learning Models." *Ecological Informatics*, vol. 82, 25 July 2024, p. 102735,

www.sciencedirect.com/science/article/pii/S1574954124002772, https://doi.org/10.1016/j.ecoinf.2024.102735.

Camargo, Julio A., and Álvaro Alonso. “Ecological and Toxicological Effects of Inorganic Nitrogen Pollution in Aquatic Ecosystems: A Global Assessment.” *Environment International*, vol. 32, no. 6, Aug. 2006, pp. 831–849, www.sciencedirect.com/science/article/pii/S0160412006000602, https://doi.org/10.1016/j.envint.2006.05.002.

Chambers, Patricia A., et al. “Development of Environmental Thresholds for Nitrogen and Phosphorus in Streams.” *Journal of Environmental Quality*, vol. 41, no. 1, Jan. 2012, pp. 7–20, https://doi.org/10.2134/jeq2010.0273.

“Chapter 1. Considerations in the Selection of Sites for Aquaculture.” *Fao.org*, 2024, www.fao.org/4/x5744e/x5744e02.htm.

Diaz, Crisandra J., et al. “Developing Algae as a Sustainable Food Source.” *Frontiers in Nutrition*, vol. 9, 19 Jan. 2023, https://doi.org/10.3389/fnut.2022.1029841.

Dodds, Walter K., et al. “Eutrophication of U.S. Freshwaters: Analysis of Potential Economic Damages.” *Environmental Science & Technology*, vol. 43, no. 1, 12 Nov. 2008, pp. 12–19, www.mostreamteam.org/Documents/Research/WaterQuality/Eutrophication%20of%20US%20Freshwaters.pdf, https://doi.org/10.1021/es801217q.

Edwards, Thea M., et al. “Ammonia and Aquatic Ecosystems – a Review of Global Sources, Biogeochemical Cycling, and Effects on Fish.” *Science of the Total Environment*, vol. 907, 21 Oct. 2023, p. 167911, www.sciencedirect.com/science/article/pii/S0048969723065385, https://doi.org/10.1016/j.scitotenv.2023.167911.

Ertürk Gürkan, Selin, and Şükran Yalçin Özdilek. “The Effects of Aquaculture Activities on Species Composition and Diversity of Native Freshwater Fish Populations.” *Aquatic Living Resources*, vol. 31, 2018, p. 17, https://doi.org/10.1051/alr/2018005. Accessed 26 Sept. 2020.

Feng, Kai, et al. “Eutrophication Induces Functional Homogenization and Traits Filtering in Chinese Lacustrine Fish Communities.” *Science of the Total Environment*, vol. 857, 21 Oct. 2022, p. 159651, www.sciencedirect.com/science/article/pii/S0048969722067511#bbb0275, https://doi.org/10.1016/j.scitotenv.2022.159651.

Hannon, Michael, et al. “Biofuels from Algae: Challenges and Potential.” *Biofuels*, vol. 1, no. 5, Sept. 2010, pp. 763–784, https://doi.org/10.4155/bfs.10.44.

He, Fengzhi, et al. “The Global Decline of Freshwater Megafauna.” *Global Change Biology*, vol. 25, no. 11, 8 Aug. 2019, pp. 3883–3892, https://doi.org/10.1111/gcb.14753.

Jin, Siyu, et al. “Immobilization of Phosphorus in Water-Sediment System by Iron-Modified Attapulgite, Calcite, Bentonite and Dolomite under Feed Input Condition: Efficiency, Mechanism, Application Mode Effect and Response of Microbial Communities and Iron

Mobilization." *Water Research*, vol. 247, 21 Oct. 2023, p. 120777, www.sciencedirect.com/science/article/abs/pii/S0043135423012174, https://doi.org/10.1016/j.watres.2023.120777.

Leong, Yoong Kit, et al. "Reuniting the Biogeochemistry of Algae for a Low-Carbon Circular Bioeconomy." *Trends in Plant Science*, vol. 26, no. 7, July 2021, pp. 729–740, https://doi.org/10.1016/j.tplants.2020.12.010.

Liu, Chunhui, et al. "Aquaculture Feeds Can Be Outlaws for Eutrophication When Hidden in Rice Fields? A Case Study in Qianjiang, China." *International Journal of Environmental Research and Public Health*, vol. 16, no. 22, 13 Nov. 2019, p. 4471, pubmed.ncbi.nlm.nih.gov/31766308/, https://doi.org/10.3390/ijerph16224471. Accessed 9 July 2023.

Liu, Na, et al. "Importance of Controlling PH-Depended Dissolved Inorganic Carbon to Prevent Algal Bloom Outbreaks." *Bioresource Technology*, vol. 220, 1 Nov. 2016, pp. 246–252, www.sciencedirect.com/science/article/pii/S0960852416311877#s0060, https://doi.org/10.1016/j.biortech.2016.08.059.

Managing Wastewater in Coastal Urban Areas. Washington, D.C., National Academies Press, 1 Jan. 1993.

Marsullo, Matteo, et al. "Dynamic Modeling of the Microalgae Cultivation Phase for Energy Production in Open Raceway Ponds and Flat Panel Photobioreactors." *Frontiers in Energy Research*, vol. 3, 15 Sept. 2015, https://doi.org/10.3389/fenrg.2015.00041.

Mayer, Ian, and Martin Pšenička. "Conservation of Teleost Fishes: Application of Reproductive Technologies." *Theriogenology Wild*, vol. 4, 1 Jan. 2024, pp. 100078–100078, www.sciencedirect.com/science/article/pii/S2773093X24000096, https://doi.org/10.1016/j.therwi.2024.100078.

NOAA. "What Is Eutrophication?" *National Ocean Service*, NOAA, 16 June 2024, oceanservice.noaa.gov/facts/eutrophication.html.

Novoveská, Lucie, et al. "Overview and Challenges of Large-Scale Cultivation of Photosynthetic Microalgae and Cyanobacteria." *Marine Drugs*, vol. 21, no. 8, 1 Aug. 2023, p. 445, www.mdpi.com/1660-3397/21/8/445, https://doi.org/10.3390/md21080445.

"Progress in Understanding Harmful Algal Blooms: Paradigm Shifts and New Technologies for Research, Monitoring, and Management." *Annual Review of Marine Science*, vol. 4, no. 1, 15 Jan. 2012, pp. 143–176, https://doi.org/10.1146/annurev-marine-120308-081121.

"Read." *The National Academies Press*, 2026, www.nationalacademies.org/read/2049/chapter/9. Accessed 27 June 2026.

Rogalski, Mary Alta, et al. "Daphniid Zooplankton Assemblage Shifts in Response to Eutrophication and Metal Contamination during the Anthropocene." *Proceedings of the Royal Society B Biological Sciences*, vol. 284, no. 1859, 26 July 2017, pp. 20170865–20170865, https://doi.org/10.1098/rspb.2017.0865. Accessed 11 July 2025.

Rosenfeld, Jordan, and Rachel Lee. "Thresholds for Reduction in Fish Growth and Consumption due to Hypoxia: Implications for Water Quality Guidelines to Protect Aquatic Life."

Environmental Management, vol. 70, no. 3, 6 July 2022, pp. 431–447, https://doi.org/10.1007/s00267-022-01678-9.

Sahu, Ashish, et al. "Environmental DNA (EDNA): Powerful Technique for Biodiversity Conservation." *Journal for Nature Conservation*, vol. 71, Dec. 2022, p. 126325, www.sciencedirect.com/science/article/pii/S1617138122001984, https://doi.org/10.1016/j.jnc.2022.126325.

Satta, Giorgia, et al. "Red Light in Underwater Measurement Applications: A Short Review." *ACTA IMEKO*, vol. 14, no. 2, 16 June 2025, pp. 1–10, www.researchgate.net/publication/393219235_Red_light_in_underwater_measurement_applications_a_short_review, https://doi.org/10.21014/actaimeko.v14i2.2090.

Shi, Huajuan, et al. "Ammonia Nitrogen Stress Disrupts Respiratory and Immune System, Antioxidant Capacity, and Intestinal Microbiota in Juvenile Cichlid *Pseudotropheus Zebra*." *Aquaculture Research*, vol. 2025, no. 1, 1 Jan. 2025, https://doi.org/10.1155/are/3778308.

Soler, Patricia, et al. "Improving Water Quality Does Not Guarantee Fish Health: Effects of Ammonia Pollution on the Behaviour of Wild-Caught Pre-Exposed Fish." *PLOS ONE*, vol. 16, no. 8, 9 Aug. 2021, p. e0243404, https://doi.org/10.1371/journal.pone.0243404.

Stuart, Venetia, et al. "The Future of Fisheries Science in Management: A Remote-Sensing Perspective." *Ices Journal of Marine Science*, vol. 68, no. 4, 17 Jan. 2011, pp. 644–650, https://doi.org/10.1093/icesjms/fsq200.

Taylor, Jessica J., et al. "The Effectiveness of Spawning Habitat Creation or Enhancement for Substrate-Spawning Temperate Fish: A Systematic Review." *Environmental Evidence*, vol. 8, no. 1, 23 May 2019, https://doi.org/10.1186/s13750-019-0162-6.

The Southeastern Aquatic Biodiversity Conservation Strategy.

Tudge, Sophie Jane, et al. "The Impacts of Biofuel Crops on Local Biodiversity: A Global Synthesis." *Biodiversity and Conservation*, vol. 30, no. 11, 24 July 2021, pp. 2863–2883, https://doi.org/10.1007/s10531-021-02232-5.

Underwood, AJ. "Beyond BACI: Experimental Designs for Detecting Human Environmental Impacts on Temporal Variations in Natural Populations." *Marine and Freshwater Research*, vol. 42, no. 5, 1991, p. 569, https://doi.org/10.1071/mf9910569.

Vaquer-Sunyer, Raquel, and Carlos M. Duarte. "Thresholds of Hypoxia for Marine Biodiversity." *Proceedings of the National Academy of Sciences*, vol. 105, no. 40, 29 Sept. 2008, pp. 15452–15457, www.ncbi.nlm.nih.gov/pmc/articles/PMC2556360/, https://doi.org/10.1073/pnas.0803833105.

Von Tress, Natalie, et al. "Supporting Cyanobacterial Bloom Monitoring with Satellite Imagery | NC State Extension Publications." *Content.ces.ncsu.edu*, content.ces.ncsu.edu/supporting-cyanobacterial-bloom-monitoring-with-satellite-imagery.

Wang, Huan, et al. "Eutrophication Causes Invertebrate Biodiversity Loss and Decreases Cross-Taxon Congruence across Anthropogenically-Disturbed Lakes." *Environment International*, vol. 153, Aug. 2021, p. 106494,

www.sciencedirect.com/science/article/pii/S0160412021001197, https://doi.org/10.1016/j.envint.2021.106494.

Wang, Yuyao, et al. “Principles and Research Progress of Physical Prevention and Control Technologies for Algae in Eutrophic Water.” *IScience*, vol. 27, no. 6, 15 May 2024, pp. 109990–109990, www.ncbi.nlm.nih.gov/pmc/articles/PMC11152667/, https://doi.org/10.1016/j.isci.2024.109990.

Wong, Aleah, et al. “A Traits-Based Approach to Assess Aquaculture’s Contributions to Food, Climate Change, and Biodiversity Goals.” *Npj Ocean Sustainability*, vol. 3, no. 1, 31 May 2024, www.nature.com/articles/s44183-024-00065-7, https://doi.org/10.1038/s44183-024-00065-7.

Yaakob, Maizatul Azrina, et al. “Influence of Nitrogen and Phosphorus on Microalgal Growth, Biomass, Lipid, and Fatty Acid Production: An Overview.” *Cells*, vol. 10, no. 2, 14 Feb. 2021, p. 393, pmc.ncbi.nlm.nih.gov/articles/PMC7918059/, https://doi.org/10.3390/cells10020393.

Yevhen Maltsev, et al. “Nitrogen and Phosphorus Stress as a Tool to Induce Lipid Production in Microalgae.” *Microbial Cell Factories*, vol. 22, no. 1, 20 Nov. 2023, https://doi.org/10.1186/s12934-023-02244-6.

Zeng, Yinqing, et al. “Harnessing the Power of EDNA Technology for Macroalgal Ecological Studies: Recent Advances, Challenges, and Future Perspectives.” *Algal Research*, vol. 77, 1 Jan. 2024, pp. 103340–103340, https://doi.org/10.1016/j.algal.2023.103340. Accessed 4 Apr. 2024.

An Ethnographic Study on How Healthcare Classes Prepare Students to Work in the Healthcare Field By Ibrahim Ali

Introduction

Growing up with five siblings, injuries and bruises were inevitable. I broke my left hand at the age of eight and my right elbow at the age of ten. I broke my right elbow while my siblings and I were playing on a swing set. My sister wanted a turn, and because I wanted to be the cool older brother, I thought it would be impressive to jump off. One thing led to another, and my elbow landed on a rock. I remember waking up in the hospital after surgery, looking around as the doctors operated and communicated with one another. People often say one bad thing can lead to something good. For me, breaking my elbow led me to discover my passion for working in the healthcare field.

When I first started high school, it was the first time I got any exposure to health classes regarding medicine. It provided me with a deeper understanding of what a doctor is and how they communicate using medical terms. Thus, I chose to conduct an autoethnography, forming my question, "How do high school health classes prepare students interested in the healthcare field? How does it help to develop and shape an identity for the healthcare field?"

My research question aims to find out whether healthcare classes prepare people interested in the healthcare field, and if so, how. Thus, I chose to focus on the importance of being literate in a person's chosen workforce, leading me to draw from Mirabelli's "Learning to serve", "To be literate here requires something Other than a ninth-grade level of literacy." (Mirabelli 149) More than just a factual, or literal interpretation of the words on the page, it requires knowledge of specific practices". I also focused on people's sponsors, which draws from Brandt's "sponsors of literacy", "sponsors deliver the ideological freight that must be borne for access to what they have"(Brandt 168). Furthermore, I will be looking into different discourses in the healthcare community, which leads me to Gee's "What is literacy, "a discourse as an 'identity kit' which comes complete with the appropriate costume and instructions on how to act and talk so as to take on a particular role that others will recognize" (Gee 18). In the end, I aim to find how people's literacies and sponsors shape and form a person's chosen discourse community.

To elaborate on my healthcare discourse community. It is important to know some medical terminology, such as prefixes, suffixes, and root words. A prefix is the beginning of a word, such as tachy, which means rapid. A suffix is in the ending of a word, like it is, which means inflammation of, and a root is the central part of a word, like cardi, which means heart. To expand on health classes, there are three health classes provided at my school; each class is taken in a different year. In freshman year, you take health science, in sophomore year, you take medical terminology, and in your junior year, you take health science theory. In health science, they teach basic communication and mostly projects surrounding your chosen career. In medical terminology, you go over medical terms, CPR practice, and learn about the human body. Finally, in health science theory, it is a recap of the last two years combined, and we study in depth. You

start deciding whether you want to take the MA (medical assistant) program, which is a two-block period that you take in your senior year, which gets you volunteering hours, and you can get hands-on experience in the healthcare field.

Methods

To draw my findings and reasoning, I chose to conduct my interviews in a Google Form format, interviewing ten people and myself. This method seemed the most accurate and accessible for reaching out to a wide range of people rather than in person interviewing because of scheduling issues. The criteria I used to interview were one person in health science, two in medical terminology, and seven in health science theory. The reason I chose only students in health science classes is their greater understanding of healthcare and their experience volunteering in clinical settings. They also have different types of certifications due to the structure of the health classes, which adds credibility to their understanding of health care. Due to not being able to schedule issues with my participants, I chose to conduct my interview via form. Thus, leading me to ask through questions for my research, I kept all the questions identical for each participant, no matter their current health classes, to ensure I could get the right data I needed.

I chose to divide my interview form into three parts, the first part being questions like who they are, introducing their career, and why they chose it. I chose to ask to know what career they are interested in while finding out their drive behind it. The next part asking questions about our health classes led to me asking my main question "Has your chosen career been influenced due to our health classes and how?" my thought process while asking that question to find out if our health classes actually prepare and guide us to our chosen career or are they namesake classes that are all talk and that they have no meaning, finally the last part led me to ask "Have you had any experiences in a clinical setting? If so, how does it correlate to the things we do in class?" I wanted to know whether they had any real workforce experience in the medical field, such as volunteering or working there part-time, while finding out whether the things we do in our health classes correlate to an actual healthcare work field. By asking these questions, I aim to find whether our classes prepare us or even guide students interested in the healthcare field, and with that, this concludes my interview and research processes and leads me to my findings.

Findings

After getting my data from all ten of my participants, including myself, I began analyzing my responses and started synthesizing them. I was specifically focusing on the questions that help answer my research question, "How do high school health classes prepare students interested in the healthcare field? How does it help to develop and shape an identity for the healthcare field?" After looking for common similarities and differences, I started to notice patterns in both similarities and differences, which helped me get different perspectives on how to answer my research question.

A common factor I noticed was that more than half of the participants have been passionate about working in the healthcare field due to their experiences in their childhood. For example, participant 2 says, “ I've always liked science, and when I got injured, it made me curious as to what the issue could be.” This explains how people’s experiences in childhood highly impact their interests when they are teenagers. It also demonstrates why they chose health care classes and how taking health care classes solidified their passion for healthcare and perhaps piqued their interest more in the field. It gives students a chance to prepare for rigorous classes in the future and gives them a chance to decide if they want to pursue this career in the future, even when there are immense hardships and obstacles in the way.

However, some differences I noticed were that the other half of the participants were money driven. For instance, participant 3 stated, “ It makes a lot of money”. This factor is important, but when working in healthcare, passion is the key factor that drives one forward and motivates one to work. Without having a liking for one's work, it becomes difficult to continue to do any type of work. Simultaneously, one can start a career due to the money involved in it, and in the future, like it and develop a passion for it. It is important for one to know that money can be made in many careers, but loving and being passionate about something is unique to every individual.

Additionally, I noticed many participants showcased that they were interested and enjoyed group-driven activities. This is an important skill a person needs to work in the healthcare field. Healthcare requires teamwork, and high school healthcare classes incorporate this skill in assignments, hands-on activities, and presentations. The healthcare field is not an individual job. As I mentioned above, it requires teamwork and a variety of people to successfully treat a patient, and high school healthcare classes focus on developing this skill and also give an opportunity to help students identify if they work better individually or prefer working with people. All of these aspects are vital to succeed in the healthcare field. Our Healthcare classes focus on implementing these skills from the high school stage, so students can acclimate to their chosen healthcare field easily and not go through many difficulties, as healthcare is a demanding field. In retrospect my question was answered from different points of views but not to the extent that I wanted it to be, most of the participants were already interested in the health care field before attending any health care classes but, they also said being in these classes helped solidify their resolve to work in the healthcare field while learning communications skills and getting certifications in the processes.

Literature Review

To connect my research back to literacy elements, one aspect that has affected me immensely is the sponsorship from my teacher. My health science teacher influenced me the most because of how she explained difficult and challenging medical topics. She supported me and motivated me when I was struggling with certain areas and guided me through the class. Sponsorship is vital in everybody's life in order for people to grow academically and personally. This correlates back to Brandt’s "Sponsors of Literacy" when she states, “sponsors nevertheless

set the terms for access to literacy and wield powerful incentives for compliance and loyalty. Sponsors are a tangible reminder that literacy learning throughout history has always required permission, sanction, assistance, coercion, or, at a minimum, contact with existing trade routes" (Brandt 167).

Moreover, literacy is a core part of every aspect of life. It is used in personal, professional, and academic life, and people rely heavily on it to communicate about various topics. In healthcare, communication plays a crucial role when dealing with patients, doctors, nurses, and technicians. Knowing medical terminology and hospital-required terms is essential in excelling in one's career. It also ensures that the quality of patient care is not impacted and makes sure that everybody in the hospital or clinical setting performs their duties. Gee's "What is literacy" states "Discourses are inherently 'ideological." They crucially involve a set of values and viewpoints in terms of which one must speak and act, at least being in the discourse; otherwise one doesn't count as being in it."(Gee 19). This correlates to my research as it shows how one being literate is vital in one's discourse community to be a part of it.

Additionally, being literate in your discourse community is a key part. For example, as I mentioned above, when working in the healthcare field, it is important to know how to use medical terms and how to communicate with patients respectfully. In Mirabelli's "Learning to Serve," he also discusses how literacy is important in a person's discourse community. He states that literacy "requires knowledge of specific practices such as methods of food preparation" (Mirabelli 150). He explains that in the discourse community of waiters, it is important to know the menu, understand how to customize dishes, and be able to recommend dishes to customers. These are all parts of being literate in that discourse community. Likewise, this connects to the healthcare workforce because skills such as giving an injection or knowing how much fluid should be in a syringe based on a patient's weight are part of participating in the healthcare discourse community. Regardless of the discourse, whether someone is a waiter or a healthcare worker, a person must be literate within that discourse to be successful.

Conclusion

I chose to pursue my research question to highlight the importance of healthcare classes taken by students to give them a chance to decide if healthcare is a job they wish to do. By taking these classes, students learn about careers they intend to pursue and also learn valuable life skills, such as performing CPR, how to react in emergency situations, or how to respond to someone who is having medical difficulties even if they diverge career paths. These classes also focus on communication and presentation, which are a big part of most careers. I wanted to use this research paper to shed light on healthcare classes so students can benefit from them and do not need to struggle to decide on a career in their future. A future question that could stem from my research would be to analyze what skills doctors use in their day to day workplace. This expands on my research as it would assess the skills you learn in class to see if they would be helpful in the real clinical setting. Even more so as a student myself I realized how beneficial taking health classes has been and would like the same for all students.

Works Cited

Brandt, Deborah. “Sponsors of Literacy.” *National Council of Teachers of English*, vol. 49, no. 2, 1998, pp. 165-185. JSTOR, doi.ord/10.2307/358929

Gee, James Paul. “What is Literacy?” *Journal of Education,* vol. 171, no. 1, 1989, pp. 1-9.

Mirabelli, Tony. "Learning to Serve." *What They Don't Learn In School*, edited by Jabari Mahiri, Peter Lang, 2005, pp.143-162.

Evaluating Structural Risk and Capital Access Barriers: Why Turkish Startups Experience Higher Early-Stage Attrition Than European Counterparts By Ayşenur Kurplikaya

Introduction and Economic Context

In the contemporary globalized economy, early-stage technology and commercial startups serve as the primary engines of systemic innovation, localized employment generation, and macroeconomic resilience. Within highly developed financial ecosystems—most notably Western Europe—a deeply integrated network of venture capital (VC) firms, robust corporate venture arms, institutional banking mechanisms, and state-backed investor tax incentives creates a structured financial cushion. This framework allows nascent enterprises to scale safely through their predictably volatile early operational years. Conversely, emerging markets present a completely divergent landscape for early-stage corporate management, where structural support layers are often absent or highly volatile.

The Republic of Türkiye occupies a compelling, highly dualistic position within this global paradigm. On one hand, the nation boasts immense underlying human capital, characterized by the largest youth demographic pool in Europe and a continuous output of nearly one million university graduates annually. This demographic reality provides Turkish ventures with an exceptionally competitive, highly skilled developer, engineering, and entrepreneurial talent base. Since 2020, this ecosystem has demonstrated its high-tier capability by successfully producing multiple tech "unicorns" valued over a billion dollars, proving that Turkish conceptual business models possess definitive global viability.

However, beneath these headline success stories lies a stark operational reality. The broader Turkish startup ecosystem suffers from a disproportionately higher rate of early-stage corporate attrition—meaning companies fail, stall, or undergo forced operational liquidation within their first 24 to 36 months—when evaluated against Western European counterparts. This divergence cannot be simplistically attributed to a deficit in talent, entrepreneurial drive, or original product-market fit. Instead, it is the direct byproduct of localized, systemic economic friction points unique to the regional emerging market landscape. This paper argues that Turkish startups experience heightened early-stage attrition due to severe localized capital access barriers, macro-economic currency and inflationary volatility, and distinct regulatory bottlenecks that collectively hinder sustainable, long-term corporate scaling.

The Comparative Ecosystem: Capital Lifecycles

To isolate the exact mechanisms driving early-stage corporate attrition, academic literature evaluates economic health through two primary variables: capital access liquidity and regulatory compliance friction. In Western European markets, institutionalized frameworks provide multi-layered safety nets for capital formation. For example, the European Investment Fund (EIF) systematically channels liquidity directly into regional venture capital funds, actively stabilizing seed and Series A funding rounds available to localized innovators. Furthermore, legislative tax incentives, such as the United Kingdom’s Seed Enterprise Investment Scheme

(SEIS), dramatically lower the downside risk for individual angel investors, incentivizing private capital deployment into high-risk, early-stage corporate entities. Consequently, the operational survival curve for European ventures is heavily insulated by an abundance of decentralized private and institutional investor choices.

Conversely, emerging market literature highlights a severe structural disconnect during the capitalization lifecycle. While initial entrepreneurial intent in Türkiye remains exceptionally high, local ventures face an aggressive "funding canyon" immediately following their initial product development phases. Empirical market metrics reveal that the structural conversion rate from the Seed stage to the Early VC stage in Türkiye sits at a critical deficit. While conversion rates consistently hover around 50% in mature markets like the United Kingdom, Germany, and France, Türkiye's seed-to-early conversion rate stands at a fraction of that figure, frequently dropping below 15%.

Furthermore, data indicates that a majority of all localized startup funding is completely centralized within Istanbul-based ventures, leaving regional or smaller-scale founders heavily isolated from capital access networks. This extreme domestic capital scarcity means that when a Turkish startup exhausts its initial pre-seed or angel-tier capital, it faces an absolute vacuum of later-stage institutional funding alternatives. When this structural funding deficit intersects with severe regional macroeconomic factors—such as an annual inflation rate fluctuating at high double-digits and tightening central bank overnight lending rates—the cost of basic corporate operations accelerates beyond sustainable thresholds. This structural imbalance leaves Turkish startups uniquely exposed to systemic financial exhaustion before achieving operational self-sustainability.

Macroeconomic Pressures and Operational Friction

To understand the baseline attrition rates of early-stage corporate entities in the Turkish market, the research must evaluate the direct operational friction caused by localized macroeconomic factors. In mature Western European ecosystems, corporate balance sheets enjoy relative currency stability, allowing for predictable long-term financial modeling and cross-border software integration. Conversely, early-stage enterprises in Türkiye operate under highly dynamic monetary environments. Because critical operational infrastructure—such as cloud computing servers, premium raw materials, specialized direct-to-consumer (DTC) supply chains, and global digital advertising platforms—is systematically priced in major foreign currencies like the United States Dollar (USD) or the Euro (EUR), local startups experience immediate structural pressure.

When a startup generates its primary revenue streams in local currency while paying its baseline operational expenses in foreign currency, it triggers a severe structural mismatch. This currency disparity rapidly accelerates the company's "burn rate"—the rate at which an early-stage company exhausts its supply of cash before turning a profit. For a small commercial enterprise, sudden shifts in international supplier costs can completely wipe out planned profit margins within a single business cycle. While an established business might have the capital

reserves to absorb these fluctuations, a bootstrapping early-stage startup faces immediate vulnerability, often forcing premature liquidation.

Regulatory and Gatekeeping Bottlenecks

Beyond purely financial capital barriers, institutional and structural roadblocks significantly hinder the scaling phase for emerging commercial models. In European markets, cross-border payment integration and international commercial shipping infrastructure are deeply standardized, allowing small business owners to access a massive, multi-country customer base with minimal administrative friction. For a Turkish startup, however, scaling outward to capture foreign currency reserves presents complex logistical hurdles.

The domestic regulatory framework introduces specific compliance requirements for cross-border e-commerce and micro-export logistics. Furthermore, the limited access to global payment processing infrastructure heavily restricts local startups from handling international consumer transactions smoothly. Because a small business cannot easily automate recurring international sales or seamlessly process foreign currency checkout flows, expanding beyond the local market requires significant administrative and manual oversight. For a growing brand run by a small team, this extra friction point shifts focus away from strategic product development and directs it toward fighting bureaucratic operational chaos. Consequently, many high-potential ventures remain trapped in localized markets, unable to scale their revenues fast enough to survive domestic economic shifts.

Case Synthesis: Comparative Operational Tracking

To clearly see how these macro-pressures play out on a day-to-day basis, we can compare the growth path of a direct-to-consumer (DTC) or e-commerce startup operating in Western Europe versus one operating in the Turkish ecosystem. Consider an early-stage apparel or consumer goods brand launching simultaneously in Berlin, Germany, and Istanbul, Türkiye. The German enterprise operates within the unified Eurozone. When sourcing premium raw textiles from regional suppliers, utilizing global software-as-a-service (SaaS) platforms for inventory tracking, or deploying digital ad spend on international platforms, its operational costs remain completely flat and predictable. Furthermore, because the European market relies on centralized, frictionless cross-border payment gateways, the Berlin-based startup can seamlessly capture revenues from consumers across France, Italy, and Spain on day one without facing payment dropping risks or foreign currency conversion fees. This financial predictability allows management to accurately calculate their exact customer acquisition cost (CAC) and secure long-term inventory contracts.

Conversely, the Istanbul-based counterpart faces a fragmented operational landscape. While the brand may source domestic labor and initial manufacturing locally, its core scaling infrastructure—such as cloud hosting, advanced design software, and international digital marketing algorithms—is billed globally in USD or EUR. As a result, even if the Turkish startup

experiences surging domestic sales volume, a sudden shift in the local currency exchange rate can instantly double its software and international marketing liabilities overnight. Furthermore, when the Turkish brand attempts to hedge against local currency risks by exporting to European consumers, it hits immediate cross-border payment bottlenecks. The lack of localized integration with dominant global payment gateways means the brand must use complex, multi-tiered checkout workarounds. This introduces extra steps for international buyers, lowers checkout conversion rates, and subjects the founder to heavy cross-border transaction fees. Consequently, the Turkish startup is forced to burn through its limited seed capital significantly faster just to maintain basic operational stability, demonstrating why the domestic survival curve remains so steep.

Conclusion and Future Outlook

In summary, the higher early-stage attrition rates observed among Turkish startups relative to Western European ecosystems are not indicative of a failure in entrepreneurial vision or technical execution. Rather, they are the predictable results of operating within a structurally demanding emerging market. While European founders are insulated by robust, state-backed capital allocation programs and seamless cross-border financial rails, Turkish entrepreneurs must constantly navigate a severe secondary funding canyon, currency-driven cost overruns, and steep operational barriers to reach international consumers.

To bridge this gap and preserve the immense potential of Turkey's young tech and commercial sectors, future economic policy must focus on decentralizing capital away from singular metropolitan hubs, expanding local angel investor tax credits, and establishing smoother pathways for cross-border digital trade. Until these structural barriers are addressed, early-stage ventures will continue to face an uneven playing field, where survival is determined less by the quality of an innovation and more by a firm's capacity to withstand macroeconomic friction.

Works Cited

KPMG Turkey M&A Team. Turkish Startup Investments Review 2025. KPMG International, Mar. 2026. https://assets.kpmg.com/content/dam/kpmg/tr/pdf/2026/03/turkish-startup-investments-review-2025.pdf.

KPMG Turkey M&A Team. Turkish Startup Investments Review Q4 2024. KPMG International, Feb. 2025. https://assets.kpmg.com/content/dam/kpmgsites/tr/pdf/2025/02/turkish-startup-investments-review-q4-2024.pdf.

Startups.watch. Turkish Startup Ecosystem - 2025 Q1 Report - TURCORN 100. Republic of Türkiye Ministry of Industry and Technology, Apr. 2025. https://www.turcorn.gov.tr/upload/Node/100286/files/_2025-Q1_Report_v1-0.pdf.

Startups.watch. Turkish Startup Ecosystem - 2023 Year in Review Report. DigitalTalks Ecosystem Reports, Jan. 2024. https://www.digitaltalks.org/wp-content/uploads/2024/01/Startups.watch-2023-Raporu.pdf.

Relationship Between Stress and Buying Patterns By Kairrah Chaudhary

Abstract

It was intended to explore the possible impact of perceived stress on online impulse buying behavior in the present study, as well as to investigate a mediation role of self-control in such influence. Previous research has evidenced a link between the studied factors, but recent studies imply that this relation might depend on the strategies used by the subjects to cope with their emotions. For this reason, the authors conducted correlational and prediction analyses based on data obtained from the use of cross-sectional design and a sample of 22 respondents aged 15-42. Subjects filled in the questionnaires measuring their level of perceived stress, impulse buying behavior, and self-control. All three scales showed high reliability. The correlation and prediction analyses involved calculation of Pearson's product-moment correlation coefficients, simple linear regressions, and hierarchical multiple regression to test for moderation effects. Results did not reveal any link between the level of stress and impulse buying behavior or self-control and impulse buying behavior. An association between self-control and perceived stress was observed, although it was insignificant. Prediction analyses revealed that neither of the two constructs could predict impulse buying behavior, nor did self-control moderate this association.

Keywords : Buying behavior, Stress, negative coping, self-control

1. INTRODUCTION

1.1 Stress and Decision-Making in Daily Life

Stress is considered a psychological condition that affects several areas of daily activities, such as emotions, cognitions, and actions. Stress appears to be most widespread among young people because of various problems related to their age category, such as educational stress, financial instability, and identity crisis. High levels of stress may contribute to the desire of individuals to cope with it through various temporary strategies for dealing with emotional discomfort, like impulse consumption. The process of purchasing products appears to be emotional rather than completely rational, and therefore the purchase may serve as one of the mechanisms of coping with negative emotions. According to integrative studies on online consumers' behaviour, highly sensitive to stress persons may buy impulsively to relieve such negative feelings as anxiety and stress (Pacheco et al., 2022). Empirical research proves the assumption that stress affects the impulse of buying (Moran et al., 2015).

1.2 What Is Impulse Buying?

Impulse buying is defined as the act of unanticipated, unplanned consumption, which happens due to an instant desire, rather than planned necessity or logical thinking. This type of purchasing behaviour usually involves quick decision-making and emotions experienced by a customer at the moment of purchase. Due to the development of online stores and increased

popularity of Internet shopping, cases of impulse purchases have become quite frequent. It is expected that many customers make unplanned purchases, simply because they happen on-the-spot and in an accessible way. Impulse buying is therefore considered a vital issue for modern consumers (Moran et al., 2015; Pacheco et al., 2022).

1.3 Stress and Impulse or Compulsive Buying: What We Know

Research evidence emerging from the fields of consumer psychology and behavioural addiction suggests that there is a positive correlation between perceived stress and impulsive/compulsive purchasing habits. People who perceive themselves as under high levels of stress can use shopping to cope with their feelings of anxiety and sorrow. It is established through empirical research that perceived stress is positively correlated with the propensity of engaging in online compulsive buying, depending on the coping style and unique traits of individuals (Zheng et al., 2019). In a recent scoping review, it was found that correlational studies tend to reveal a connection between stress and compulsive buying-shopping disorder, where people with maladaptive buying behaviour demonstrate higher levels of perceived stress when compared to the control group (Thomas et al., 2024).

1.4 Aim and Research Question

The present study aims to examine whether perceived stress is associated with impulse buying tendencies. Based on prior literature, it is hypothesised that higher perceived stress will be associated with stronger impulse buying tendencies.

2. LITERATURE REVIEW

2.1 Evidence That Stress Is Positively Related to Impulse or Compulsive Buying

There is sufficient empirical evidence to support the notion that stress has a positive association with impulsive and/or compulsive buying behavior. First, there are questionnaire-based studies that have shown that stress predicts online compulsive buying behavior, and coping behaviors partially mediate such a relationship (Zheng et al., 2019). Second, a scoping study on compulsive buying-shopping disorder has revealed a significant correlation between stress and symptom intensity in numerous studies and that people with problematic purchasing behaviors tend to experience more stress compared to control groups (Thomas et al., 2024). Experimental studies also reveal that stress is positively related to purchasing behavior. In this regard, Moran et al. (2015) conducted an experiment and found that stressed participants were prone to making more impulsive purchases when browsing online stores.

2.2 Stress as an Emotional Trigger for Impulse Buying

Such psychological dynamics have been observed to be mostly driven by emotional regulation processes. According to integrative theory, people who are highly susceptible to stress resort to impulsive purchases as a way of dealing with their negative feelings, like stress and

boredom (Pacheco et al., 2022). Moran et al. (2015), on the other hand, identify impulsive purchases as a means of addressing the emotional distress caused by stressful experiences, such as dejection.

2.3 Mediating Role of Negative Coping

Stress has been recently found to be associated with impulsive purchases through coping. This finding was demonstrated by Zheng et al. (2019), who found that stress was mediated by the use of ineffective coping strategies. Similar findings have been discovered in studies on stress induced by social media in relation to compulsive purchases. The authors noted that such stress had a detrimental effect on compulsive buying via the use of maladaptive coping strategies, thus implying that there were alternative ways through which stress could affect consumer behavior, one of which was emotional regulation (Purwaningdyah & Pratminingsih, 2024).

2.4 Moderators: Self-Esteem, Self-Control, and Emotional Intelligence

Further, individual characteristics determine the effects of stress on consumer behavior. According to Zheng et al. (2019), self-esteem moderates the effect of stress on compulsive buying. Specifically, high self-esteem reduces the effect of stress on compulsive buying behavior. Likewise, when studying Gen Z consumers, Purwaningdyah & Pratminingsih (2024) noted that low levels of self-control increase the effect of stress on impulsive buying behavior. Also, some scholars argue that the effect of stress on impulsivity might be mediated by emotional intelligence, which might reduce the effect by enhancing emotional regulation (Shams et al., 2021).

2.5 Mixed and Inconsistent Findings

Even though there is an abundance of scientific information regarding the links between stress and impulse buying, the existing body of research draws attention to several issues related to methodology. In particular, Thomas et al. (2024) emphasize significant heterogeneity in design, measurements, and methods used by different studies. In addition, relatively few longitudinal or experimental studies have been conducted so far. Correlational research tends to find a positive association between stress and buying, while regression or structural equation modelling research gives inconsistent results.

2.6 Identifying the Gap

While previous studies have identified stress as a factor that influences impulsive and compulsive buying behaviors, there is little literature about the effects of stress on people aged between 18-25 years. A number of studies tend to investigate stress as a factor in a broad range of samples or specific segments, including but not limited to females, employees, and Generation Z people (Zheng et al., 2019; Shams et al., 2021). Also, it needs to be explored if stress acts on people's impulsivity by means of mediating variables.

2.7 Why This Research Is Important

Stress is prevalent in young adults because of the numerous transitions they experience in their lives. Buying on impulse can create financial pressure and dissatisfaction. This study explores the link between stress and impulse buying to help develop strategies that promote sound financial habits and stress management among young adults.

3. METHODOLOGY

3.1 Participants

The participants for this research were drawn from the general population ranging in ages between 15 to 42 years old. These included people from India, the UK, Dubai and from Instagram members within a group that I am a member of. All these participants received the link to the Google form and were instructed to share it as much as possible among the same age category. The screening question, which was the first question of the survey, asked the participant to identify his/her age range between 15-42.

3.2 Research Design

This study was aimed at investigating the relationship between perceived stress, negative coping, self-control, and online impulsive buying behaviour. It was suitable for answering the research question, as it enables the analysis of multiple factors measured at one moment of time. Thus, there are multiple variables under study that could be correlated with each other. In particular, perceived stress is an independent variable, and online impulsive buying behaviour is a dependent one; moreover, negative coping was considered as a potential mediator, whereas self-control was suggested as a moderator.

3.3 Measures

a. Perceived Stress

The measure of perceived stress entailed a series of eight questions that were answered on a five-point Likert scale ranging from 1 ("never") to 5 ("very often"). These questions addressed the extent to which subjects experienced such feelings of being stressed as having a hard time relaxing, feeling anxious, and being unable to cope with multiple pressures.

b. Online Impulse Buying

The impulsiveness tendency in online purchases was measured through seven statements based on a 5-point Likert scale ranging from 1 for strongly disagreeing to 5 for strongly agreeing. These statements focused on aspects like impulse purchase behavior, urges to make purchases during browsing, and impulse purchases without consideration.

c. Negative Coping

Negative coping mechanisms were assessed using a scale with five questions which used a 5-point Likert Scale with answers ranging from 1 (strongly disagree) to 5 (strongly agree). These questions are based on the use of online shopping as a coping mechanism in dealing with stress

and its consequences. This involves purchasing products to change one's state of mind and to forget one's worries.

d. Self-Control

The self-control trait was assessed based on five questions that ranged from 1 to 5 using a 5-point Likert scale. The questions reflected one's ability to resist temptations, manage his impulses, and remain focused on his long-term objectives. There were two questions that were reversed before analyzing them.

3.4 Procedure

Participants were recruited using a survey conducted online using Google Forms. The survey was shared through an Instagram group and via family members, friends, and a school located in India through convenience sampling. After viewing the information page describing the aim of the study, participants were presented with two screening questions assessing age and informed consent. Respondents who did not satisfy the requirements or those who declined to give consent were led to the end of the survey. Those who satisfied the requirements and gave consent continued with the survey consisting of six questionnaires, namely the perceived stress scale, impulsivity in online shopping behavior, negative coping strategy, self-control, and attention item. An indication of progression bar was included to increase the completion rate of the survey. No contact details were recorded, and all the results were anonymously stored in a Google Sheet.

3.5 Data Analysis

All data analyses were done using R Studio. Prior to conducting hypothesis tests, data screening was done to identify missing cases, outliers, attention checks, and violations of statistical assumptions. Those participants whose answers on the attention check question turned out to be wrong were excluded from data analysis.

In order to address the research questions, correlation tests were performed to determine the association between perceived stress and online impulse buying behavior. In case when data violated normality assumption, Spearman's rank correlation test was used. To test the hypotheses concerning mediation model, a simple mediation test was conducted in which perceived stress acted as an independent variable, negative coping as a mediator, and online impulse buying as an outcome. The bootstrapping approach was applied to find the significant indirect effect.

The study tested the mediation model where the relationship between perceived stress and online impulse buying moderated by self-control was determined. Interaction term between perceived stress and self-control made results more precise. Results of hypothesis testing were accepted as statistically significant at a p-level lower than 0.05.

4. Results

4.1 Data Screening and Participants

The total number of survey responses collected was 23, and one of the respondents opted out of giving consent before participating in the study, reducing the number of responses to be analyzed from N = 23 to N = 22. One respondent did not answer the question on the attention check variable, although the data was used for the analysis of all variables.
Participants consisted of 17 females (77.3%), four males (18.2%), and one individual who did not identify themselves as male or female (4.5%), with an average age of M = 23.14 (SD = 8.73, Mdn = 19, range = 15-42). Students comprised the largest proportion of the sample (n = 14; 63.6%), followed by working adults (n = 6; 27.3%) and individuals under other conditions (n = 2; 9.1%). Among the options available, the largest group reported spending between AED 0-50 per month on the Internet (n = 7; 31.8%), followed by AED 100-200 (n = 6; 27.3%), AED 50-100 (n = 5; 22.7%), and AED 200 or more (n = 4; 18.2%). Regarding daily social media usage time, the largest portion of participants used between one to

4.2 Scale Reliability

Internal consistency was examined using the Cronbach's alpha test before deriving scores. The reliability of the scale for assessing stress perception (with 8 items) was high (α=.857). The scale measuring the propensity to make impulsive online purchases (with 5 items) showed very high reliability (α=.915). The scale measuring self-control (with 5 items and two reverse coded items) had acceptable reliability (α=.725).

4.3 Descriptive Statistics and Normality

For each scale, the descriptive statistics are presented in Table 1. Perceived stress scale (range: 8-40) had a mean value of M = 25.18 (SD = 5.93, Mdn = 24.00, range: 17-38), suggesting that the participants felt a moderate level of stress. Online impulse buying scale (range: 5-25) had a mean value of M = 13.95 (SD = 5.21, Mdn = 13.50, range: 5-25). Self-control scale (range: 5-25) had a mean of M = 16.68 (SD = 3.27, Mdn = 16.00, range: 12-23), where reverse scores of two items were taken into account.
A Shapiro-Wilk test was performed to verify the assumption of normal distribution, as it applies to smaller samples. All assumptions of normality were met for all scales: perceived stress (W = .934, p = .152), online impulse buying (W = .967, p = .647), self-control (W = .930, p = .122). Thus, Pearson product-moment correlations were conducted.

4.4 Bivariate Correlations

There was no significant relationship between perceived stress and online impulse buying (r = .171, p = .446, 95% CI [–.270, .553]). There was, however, a non-significant negative association between perceived stress and self-control (r = –.387, p = .075, 95% CI [–.695, .041]). Furthermore, there was also no significant relationship between online impulse buying and self-control (r = –.275, p = .216, 95% CI [–.624, .166]). Using Spearman's rank-order correlation analysis showed similar results (stress-impulse buying: ρ = .274, p = .217; stress-self-control: ρ = –.363, p = .097; impulse buying-self-control: ρ = –.280, p = .207).

4.5 Primary Analysis: Perceived Stress as a Predictor of Online Impulse Buying

A linear regression analysis was performed to determine if higher levels of perceived stress are associated with higher levels of impulse purchases made on the Internet. Perceived stress was mean-centered prior to analysis. However, the entire model was not significant (R^2 = .029, F(1, 20) = 0.606, p = .446), nor was perceived stress a predictor of online impulse buying (β = 0.150, p = .446, 95% CI [–0.253, 0.554]).

4.6 Moderation Analysis: Self-Control as a Moderator

Hierarchical multiple regression was used to test if self-control moderates the link between perceived stress and impulsive online shopping behaviors. The two independent variables were mean-centered prior to data analyses in order to control for the possibility of multicollinearity. An interaction variable was created by multiplying the two variables. Variance inflation factors for all predictors fell within acceptable levels (perceived stress: VIF = 1.29; self-control: VIF = 1.21; interaction: VIF = 1.10).

As per Step 1, the impact of perceived stress (β = 0.067, p = .752, 95% CI [–0.371, 0.505]) and self-control (β = –0.390, p = .317, 95% CI [–1.185, 0.404]) on online impulse buying was non-significant and combinedly explained only a small portion of variance in online impulse buying behavior (R^2 = .081, adj R^2 = –.016, F(2, 19) = 0.832, p = .450).
As per Step 2, the inclusion of the interaction term did not have a significantly positive contribution to the model (ΔR^2 = .039, F-change(1, 18) = 0.799, p = .383). The interaction term was non-significant (β = 0.053, p = .383, 95% CI [–0.072, 0.179]), as were both the main effects (stress: β = 0.125, p = .578; self-control: β = –0.332, p = .403

5. Discussion

5.1 Overview of Findings

In the current study, the effect of perceived stress on online impulsive buying among young adults was investigated along with self-control as a possible moderator of this effect. Contrary to the hypothesized effect, perceived stress did not significantly predict online impulsive buying (r = .171, p = .446). Neither self-control nor an interaction between stress and self-control were found to have a significant role in predicting online impulsive buying. A negative but non-significant correlation was also noted between perceived stress and self-control (r = –.387, p = .075).

5.2 Interpreting the Null Association Between Stress and Impulse Buying

The absence of a correlation between perceived stress and online impulse buying goes against earlier findings which revealed positive relationships between stress and impulsive or compulsive buying behavior (Zheng et al., 2020; Purwaningdyah & Pratminingsih, 2024). Nonetheless, such a finding does not make this study incongruous but rather gives insight. The literature shows that stress does not necessarily lead to impulsive buying, but rather, the impacts of stress are mediated by maladaptive coping mechanisms (Maharani et al., 2023; Zheng et al.,

2020) and personal factors like self-esteem and self-control (Nyrhinen et al., 2023; Gulfraz et al., 2022). In light of such a phenomenon, focusing solely on the relationship between stress and impulse buying would yield insufficient knowledge of the whole process involved. Importantly, according to Maharani et al. (2023), despite the fact that stress-related elements, such as failure of self-control and self-reward, significantly predicted online impulse purchases, the perception of stress was not found to have any significant effect. Therefore, this result corroborates the current findings, which imply that there is no direct connection between stress and impulsive buying behaviors, and stress can be considered a remote variable operating psychologically.

5.3 The Mediating Role of Coping: A Plausible Pathway

Mediation was not investigated in this study, but studies have already found evidence that negative coping mediates the relationship between stress and impulsive buying behavior. According to Zheng et al. (2020), negative coping mechanisms partially mediate the influence of stress perception on online compulsive buying behaviors. Purwaningdyah and Pratminingsih (2024) found a similar result when they found out that stress induced by social media use increases impulsive buying behavior by means of maladaptive coping mechanism. According to Jabutay et al. (2024), perceived stress mediates the relationship between social isolation and hedonic browsing that predicts e-impulse buying. Thus, based on the results mentioned above, it may be argued that stress-induced impulse buying is a process involving two stages, so conducting a test to examine a potential stress-bad buying relationship might be meaningless in this particular case. To examine the potential mediation effects, future studies are recommended to include measures of coping mechanisms.

5.4 The Role of Self-Control: Non-Significant Moderation and a Marginal Trend

Despite the findings indicating that the presence of self-control does not moderate the effect of perceived stress on online impulse shopping, contrary to past studies that suggest self-control to be a negative moderating factor for impulsive buying (Gulfraz et al., 2022; Nyrhinen et al., 2023; Amelia et al., 2025), the existence of a marginal negative correlation between perceived stress and self-control ($r = –.387$, $p = .075$) supports the concept of ego depletion theory since stress weakens self-regulation, hence making an individual more vulnerable to impulsive tendencies (Verplanken et al., 2011).

The inability to find significant moderation may also be due to a lack of variability in self-control among the sample members since there was only moderate self-control scores among the respondents ($M = 16.68$). Research indicates that significant moderation was established from samples that demonstrated greater self-control and larger number of participants (Nyrhinen et al., 2023; Gulfraz et al., 2022). It is the same conclusion drawn by Amelia et al. (2025), whereby self-control can serve as the main protective factor to emotional buying only when there is self-control variation.

5.5 Emotional Mechanisms and Broader Contextual Factors

Beyond the stress-buys relationship, other emotional and situational factors have been associated in the literature, either independently of stress or alongside it. According to Kimiagari et al. (2021), both cognitive and affective responses were key mediators of online impulse buying, indicating emotional arousal but not stress as a motivator of unplanned purchases. Negative emotions and personality traits, in turn, according to Wang et al. (2022), served as channels for stress-induced impulsive actions, whereas Amelia et al. (2025) noted fear of missing out (FOMO) as a prominent trigger in a digital environment. Time pressure and countdowns, among other situational elements, increase arousal and motivate individuals to buy impulsively (Sun et al., 2023b). Self-reported stress measurements cannot account for situational factors that impact impulse purchases and, therefore, can only partially explain the stress–buying association, which presents an interesting area for future exploration using an experimental design.

5.6 Protective Factors and Intervention Implications

Other studies in the larger body of literature have identified personal resources that could potentially mitigate the stress-related impulse buying phenomenon. In their study, Vihari et al. (2022) showed that mindfulness mitigated the negative impacts of problem internet usage and acted as an additional safeguard to mitigate impulse buying among online shoppers, where emotional intelligence was the other moderating variable. Similarly, Wang et al. (2023), in a randomised control trial study design, found that self-compassion interventions decreased the practice of impulse buying through improved self-control and decreased materialistic tendencies, thereby establishing causation for this relationship. Likewise, Kurniawati et al. (2025) showed that emotion regulation and self-control independently predicted impulse buying behavior. Therefore, a potential way of mitigating stress-induced impulse purchase is through mindfulness-based or self-compassion programmes aimed at developing self-control capacities among such individuals.

5.7 Ethical and Practical Implications

The results from the current study are associated with another larger ethical issue about how digital sites promote impulsive buying. Both Wang et al. (2022) and Gulfraz et al. (2022) demonstrate that digital spaces consistently use emotional and environmental factors like personalized recommendations and urgent messages for increasing impulsive shopping behaviors, which ultimately have negative effects on people's financial health. Impulse purchases can be defined as a self-regulatory failure that occurs due to a persuasive environment in which consumers suffer psychologically and financially while being forced not to think consciously about their decisions.

5.8 Limitations

There are several weaknesses associated with the current research that deserve highlighting. First of all, the small sample size (N = 22) made it impossible to conduct properly

powered analysis. As suggested by Button et al. (2013), the use of a small sample makes the study more vulnerable to committing a Type II error which implies the failure to detect a real effect. Furthermore, according to Szűcs et al. (2016), most of the published psychological research is underpowered when it comes to detecting small or medium-sized effects. Thus, the study conducted on the current sample had an estimated power of about 11% to detect an effect of r = .171, implying that the null result could not be used as proof of non-existence of a relationship between variables. At the same time, based on the ideas advanced by Smith et al. (2018), the current results could be considered informative.

Secondly, the age range was broader than expected (15-42 years), thus leading to variations in both stress and buying behaviour at different life stages and development periods. Thirdly, due to its cross-sectional nature, the study cannot establish causality, and, moreover, being based solely on self-reports poses the threat of social desirability bias as well as potential inaccuracies when reporting past events. Fourthly, the planned mediator analysis, representing the ultimate examination of the proposed mechanism between stress and impulse purchases, is impossible to carry out.

5.9 Future Directions

Further research needs to ensure that the participant sample remains between the ages of 16 and 25 years and have a minimum sample size of N = 80–100 for adequate statistical power when conducting correlational and moderation tests, as well as bigger samples when doing mediation analysis. Future research designs will be able to investigate how chronic stress varies in terms of impulse buying in a longitudinal framework and even conduct experiments similar to what Moran et al. (2015) did. Adding the variable "negative coping" will enable researchers to perform mediation tests.

6. Conclusion

In the current study, there was no statistical support for the existence of any direct association between stress perception and online impulsive buying behavior, while self-control did not have a moderating role in this relationship either. It seems that such results should be attributed to insufficient statistical power and heterogeneous sampling rather than the true lack of effect; in addition, the sign of the correlation coefficient (r = .171) is quite congruent with the existing empirical research. On the whole, the empirical literature indicates the stress- and emotion-related self-regulatory failure underpersuasive conditions, whereby negative coping strategies, hedonistic search behaviors, and exhaustion of emotional energy turn stress into impulsive buying.

7. References

Amelia, R., Yafiz, M., & Kamilah. (2025). Emotional triggers and self-control in digital consumption within Islamic communities. *Journal of Social Commerce*, *5*(3). https://doi.org/10.56209/jommerce.v5i3.175

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Gulfraz, M. B., Sufyan, M., Mustak, M., Salminen, J., & Srivastava, D. K. (2022). Understanding the impact of online customers' shopping experience on online impulsive buying: A study on two leading e-commerce platforms. *Journal of Retailing and Consumer Services*, Article 103000. https://doi.org/10.1016/j.jretconser.2022.103000

Hussey, I. (2023). A systematic review of null hypothesis significance testing, sample sizes, and statistical power in research using the Implicit Relational Assessment Procedure. *Journal of Contextual Behavioral Science*. https://doi.org/10.1016/j.jcbs.2023.06.008

Jabutay, F. A., & Limpachote, T. (2024). Young adults' perceived social isolation and e-impulse buying on social commerce platforms: Linking roles of perceived stress and hedonic browsing. *Young Consumers*, *25*(6), 807–823. https://doi.org/10.1108/YC-11-2023-1910

Kimiagari, S., & Sharifi Asadi Malafe, N. (2021). The role of cognitive and affective responses in the relationship between internal and external stimuli on online impulse buying behaviour. *Journal of Retailing and Consumer Services*, Article 102567. https://doi.org/10.1016/j.jretconser.2021.102567

Kurniawati, A., et al. (2025). Regulasi emosi dan self control terhadap perilaku impulse buying. *FENOMENA*.

Maharani, S. A. D., & Utami, N. P. (2023). Coping mechanisms of stress: The impact on online purchase impulsivity. *Journal of Business, Management, and Social Studies*, *3*(3). https://doi.org/10.53748/jbms.v3i3.70

Moran, B., & Kwak, L. E. (2015). *Effect of stress, materialism, and external stimuli on online impulse buying* [Unpublished manuscript]. Department of Marketing, School of Business, Southern Connecticut State University. https://www.southernct.edu/sites/default/files/a/directory/kwake1/MoranKwak2015.pdf

Nyrhinen, J., Sirola, A., Koskelainen, T., Munnukka, J., & Wilska, T.-A. (2023). Online antecedents for young consumers' impulse buying behaviour. *Computers in Human Behavior*, Article 108129. https://doi.org/10.1016/j.chb.2023.108129

Pacheco, D. C., Moniz, A. I. D. S. A., Caldeira, S. N., & Silva, O. D. L. (2022). Online impulse buying — Integrative review of psychological factors. In A. Mesquita, A. Abreu, & J. V. Carvalho (Eds.), *Perspectives and trends in education and technology* (Smart Innovation, Systems and Technologies, Vol. 256, pp. 451–460). Springer. https://doi.org/10.1007/978-981-16-5063-5_44

Purwaningdyah, S. W. S., & Pratminingsih, S. A. (2025). Stressed and shopping: Examining social media-induced stress and online impulsive buying in Gen Z. *JBTI: Jurnal Bisnis: Teori dan Implementasi*, *15*(3). https://doi.org/10.18196/jbti.v15i3.24822

Shams, H. A., Hussain, S., Feroze, T., Saki, S. A., Shams, M. A., Abbasi, K. A., & Munir, S. (2021). Stressors, stress and impulsive buying behavior: Moderating role of emotional intelligence. *International Journal of Business and Economic Affairs*, *6*(3), 164–174. https://ijbea.com/ojs/index.php/ijbea/article/view/219

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-*N* design. *Psychonomic Bulletin & Review*, *25*, 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Sun, B., Zhang, Y., & Zheng, L. (2023). Relationship between time pressure and consumers' impulsive buying — Role of perceived value and emotions. *Heliyon*. https://doi.org/10.1016/j.heliyon.2023.e23185

Sun, J., Li, T., & Sun, S. (2023). Factors affecting users' impulse purchases in online group buying: Online consumer reviews, countdowns and self-control. *Asia Pacific Journal of Marketing and Logistics*, *36*(1), 224–240. https://doi.org/10.1108/APJML-07-2022-0560

Szűcs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*. https://doi.org/10.1371/journal.pbio.2000797

Thomas, T. A., Schmid, A. M., Kessling, A., Wolf, O. T., Brand, M., Steins-Loeber, S., & Müller, A. (2024). Stress and compulsive buying-shopping disorder: A scoping review. *Comprehensive Psychiatry*, Article 152482. https://doi.org/10.1016/j.comppsych.2024.152482

Verplanken, B., & Sato, A. (2011). The psychology of impulse buying: An integrative self-regulation approach. *Journal of Consumer Policy*, *34*, 197–210. https://doi.org/10.1007/s10603-011-9158-5

Vihari, N. S., Sinha, N. K., Tyagi, A., & Mittal, S. (2022). Effect of mindfulness on online impulse buying: Moderated mediation model of problematic internet use and emotional intelligence. *Frontiers in Psychology*, *13*, Article 1012331. https://doi.org/10.3389/fpsyg.2022.1012331

Wang, Y., Pan, J., Xu, Y., Luo, J., & Wu, Y. (2022). The determinants of impulsive buying behaviour in electronic commerce. *Sustainability*, *14*(12), Article 7500. https://doi.org/10.3390/su14127500

Wang, Y., Zhou, J., Gu, X., et al. (2023). The effect of self-compassion on impulse buying: A randomized controlled trial of an online self-help intervention. *Mindfulness*, *14*, 1542–1551. https://doi.org/10.1007/s12671-023-02139-y

Zheng, Y., Yang, X., Liu, Q., Chu, X., Huang, Q., & Zhou, Z. (2020). Perceived stress and online compulsive buying among women: A moderated mediation model. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2019.09.012

Investigating the Role of TikTok in the Growth and Success of Gen Z Fashion Brands: A Case Study of Pink Palm Puff (PPP) By Raeka Ren

Abstract

This Study investigates the role of TikTok in driving the growth of engagement for Generation Z fashion brands, focusing on the case study of Pink Palm Puff (PPP). A mixed-methods approach was employed, integrating quantitative analysis of PPP's recent TikTok performance, examining metrics such as views, likes, comments, and shares, alongside follower growth, with a qualitative evaluation of content types, brand aesthetic, and audience engagement. The results show that PPP's success is closely tied to how well it utilizes viral trends, works with influencers, and maintains a consistent brand style. This has led to more visibility and audience loyalty. These insights suggest that TikTok could be a cost-effective and high-impact marketing tool for new fashion brands, offering practical implications for entrepreneurs and digital marketers seeking to connect with Gen Z audiences.

Keywords: TikTok, Generation Z, fashion brands, Influencer marketing, consumer culture, digital marketing.

Introduction

Digital marketing is evolving rapidly and transforming how brands connect with customers and compete in the market. Generation Z has been the most affected by this change, among all age groups. Gen Z is the first generation to grow up every day in the digital age. They are always connected through mobile devices and social media, which can alter how they discover, assess, and engage with brands. Unlike previous generations, they are less likely to be influenced by traditional advertising formats such as television, print, or radio, and more likely to interact with brands through digital platforms that offer immediacy, interactivity, and personalization. TikTok, in particular, has emerged as a powerful platform for reaching Gen Z audiences. Its unique algorithm allows even small brands to achieve viral reach by promoting content based on user interests rather than follower count. TikTok is a great platform for fashion brands to create short, visually appealing videos, collaborate with influencers, and participate in trends, ultimately increasing awareness of their products and driving sales. Through creative storytelling and the ability to join viral conversations, brands can establish a cultural presence that resonates with younger consumers. The objective of this study was to examine the impact of TikTok on the development and success of Gen Z fashion brands, using Pink Palm Puff (PPP) as a case study. This study asks: How does TikTok content influence the growth and engagement of Gen Z fashion brands, specifically through the case study of Pink Palm Puff (PPP)? It argues that TikTok plays a significant role in the success of Gen Z fashion brands by amplifying visibility through influencers, fostering emotional engagement through aesthetic appeal, and integrating brands into Gen Z consumer culture.

Literature Review

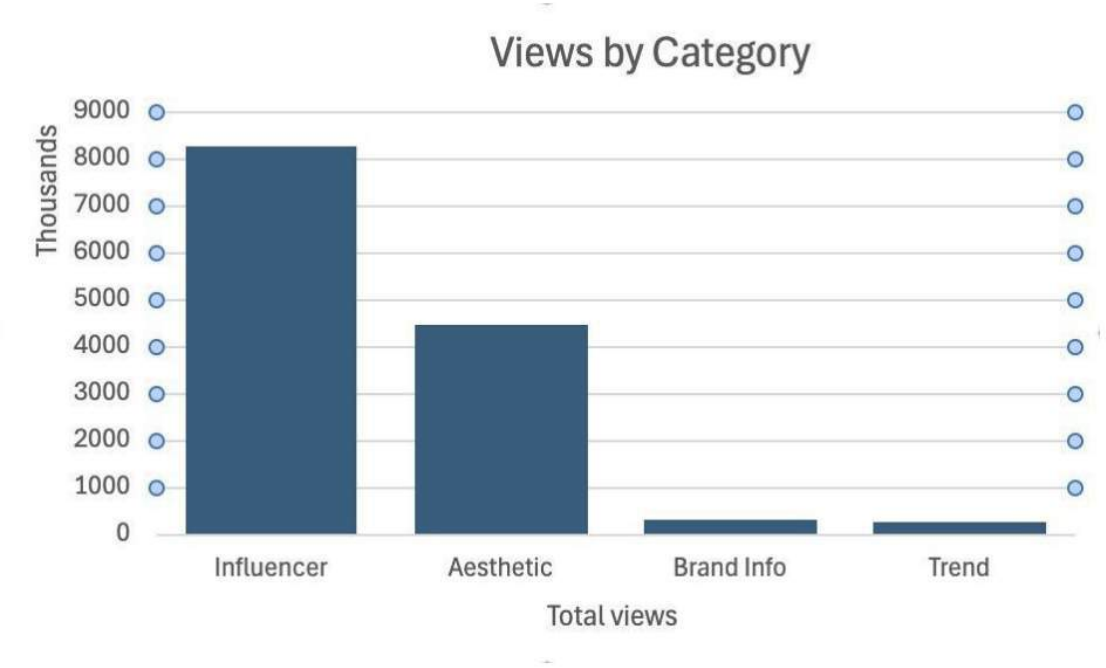
The rise of TikTok has fundamentally reshaped the digital marketing landscape, particularly for brands targeting Generation Z. Unlike earlier platforms, TikTok's recommendation algorithm prioritizes users' interactions, such as watch time, re-watches, likes, and comments, over follower count. This can allow both established and emerging brands to achieve significant visibility. Koc (2023) emphasizes that this personalization can enable smaller fashion labels, such as Pink Palm Puff, to go viral by investing heavily in traditional advertising. Similarly, Ying (2025) demonstrates how TikTok's algorithm directly shapes Gen Z's purchase intentions, thereby strongly reinforcing the platform's role as a powerful tool for consumer influence.

Beyond TikTok's algorithmic mechanics, TikTok stands out as a unique marketing ecosystem. A 2025 review by Lutfiina et al. (2025) synthesizes findings from over 15 peer-reviewed studies, concluding that TikTok enhances brand visibility through the use of viral content, storytelling, and influencer participation. The study also identifies challenges, such as algorithm dependency and the need for content consistency, which highlights the importance of strategy in sustaining growth.

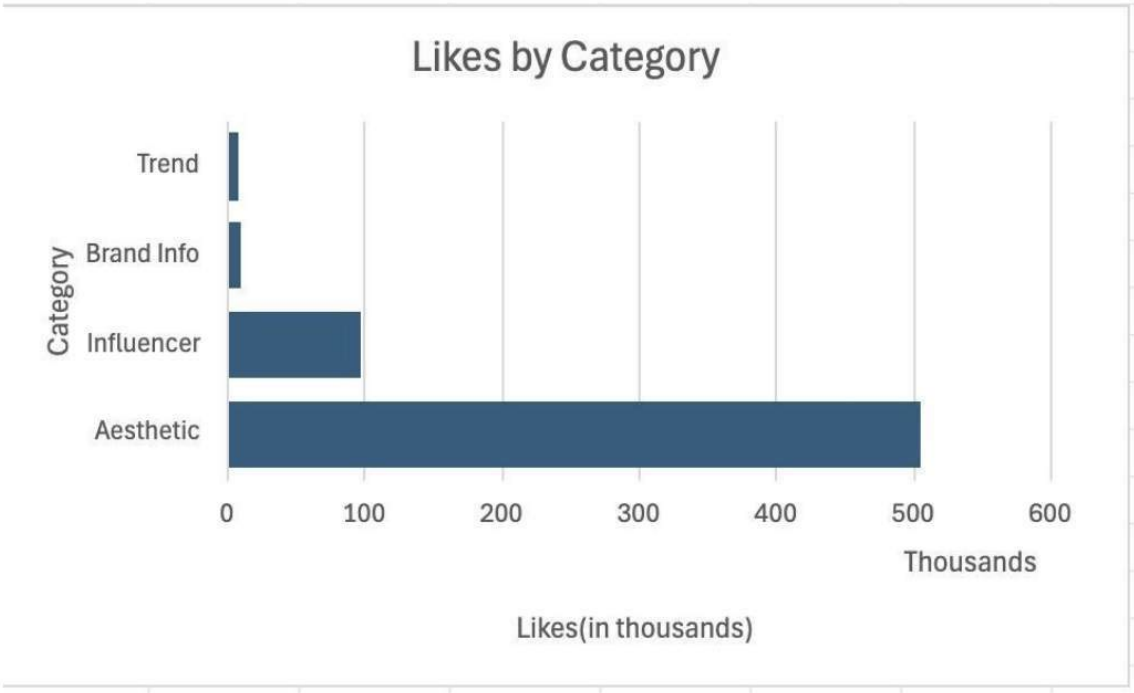
Understanding this algorithmic potential requires consideration of Gen Z's consumer behaviour. Priyanka(2023) highlights the fact that consumers of the Gen Z generation are less receptive to traditional forms of advertising and are more likely to engage with content that is visually appealing, trend-driven, and aligned with their values. This generation is also susceptible to issues of social responsibility and inclusivity. As a result, they are exerting pressure on brands to adopt communication strategies that are also creative and transparent. These kinds of discoveries help to explain why some audiences respond so strongly to PPP's consistent aesthetic, its use of short-form videos, and its participation in viral trends.
Influencer marketing represents another key factor in Gen Z brand engagement. A case study on Scarlett Whitening demonstrates that collaborations with both macro- and micro-influencers significantly enhance trust, credibility, and purchasing intent among Gen Z consumers (Runnan, 2023). Influencers are perceived as more relatable and genuine than traditional celebrity endorsements, making them more persuasive to this group. For PPP, working with influencers not only increases the brand's exposure but also places it in the cultural spaces that Gen Z uses on TikTok. The literature consistently emphasizes that influencers act as crucial intermediaries between brands and young consumers, enhancing both visibility and emotional engagement. These insights can help explain why audiences respond so strongly to PPP's consistent aesthetic, use of short-form videos, and participation in trends.

Results:

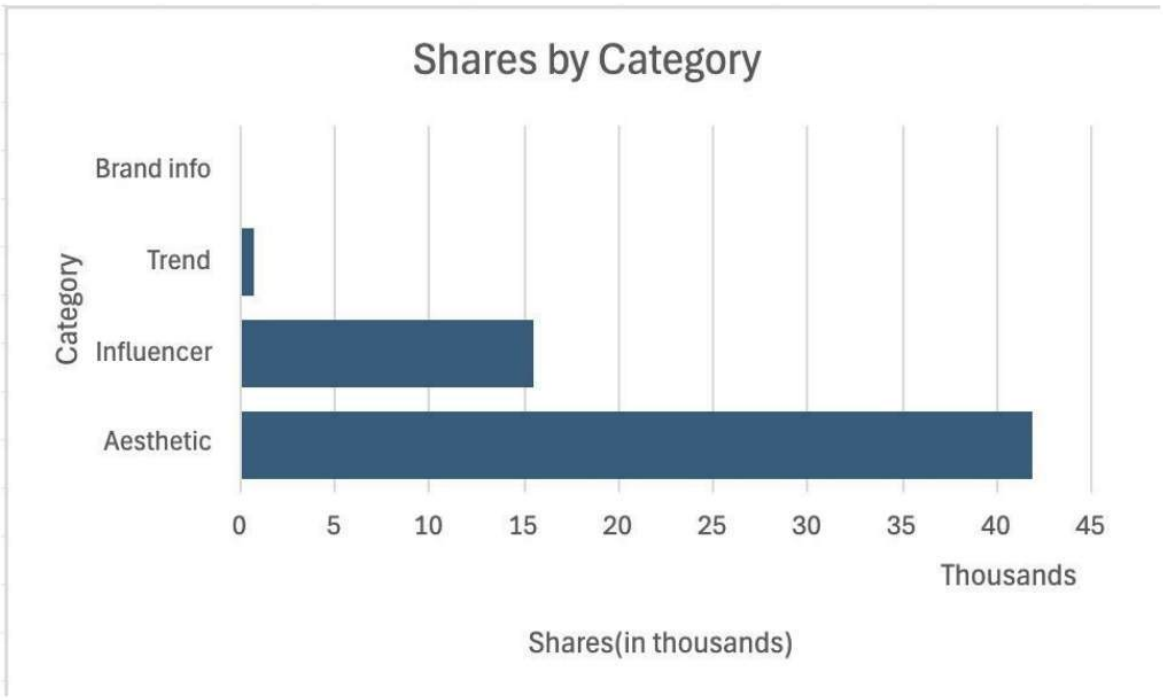
The analysis of PPP's TikTok content shows a clear difference in performance across categories



Videos featuring influencers amassed over 8 million views, illustrating the significant appeal of influencer-driven content relative to alternative strategies. Aesthetic-focused videos garnered approximately 4.4 million views, whereas videos centred on brand information and trends exhibited the poorest performance, each remaining below 500,000 total views. This suggests that people are more likely to interact with PPP when the content is associated with well-known individuals rather than just brand-driven messages. When examining engagement metrics, however, a different pattern emerges.



Videos that focused on aesthetics got the most likes, with over 500,000. Videos that focused on influencers received significantly fewer likes, around 100,000. Brand information and trend content were once again behind. This means that influencers can help people see things, but aesthetic content can help people feel connected to them and appreciate them. Shares strike a balance between reach and relatability.



As can be seen in this graph, engaging videos received the highest number of shares (approximately 42,000), followed by content from influencers (approximately 16,000). Videos based on trends and those providing information about brands received very few shares. This reaffirms the idea that viewers find aesthetic posts to be the most "shareable," which contributes to the organic growth of PPP through the process of peer-to-peer diffusion.

Individual influencers are vital for PPP's brand visibility, in addition to content categories. It's essential to note that TikTok stars like Embreigh (3.8 million followers) and Preslee Faith (2.7 million followers) play a significant role in PPP's success. Embreigh worked on a PPP collection that sold out within minutes of being released. This illustrates the significant direct buying power that influencers possess. Because Embreigh and Preslee both have large, very active teen audiences, their connection to PPP puts the brand at the centre of Gen Z consumer culture. This trend is significant because it indicates that Gen Z is generally more consumer-oriented. When they make purchases, they are heavily influenced by how relatable, attractive, and genuine they perceive the product to be. When teens see trusted influencers showcasing products or entire collections, it makes them want to buy them too: "If they buy it, I want it too." The success of PPP demonstrates that TikTok is more than just a means of advertising; it is also a platform where people can discover their identity, and where purchases are made based on a combination of entertainment, peer culture, and influencer-driven demand. Ripple effect

Methodology:

A mixed-methods approach was employed, integrating quantitative and qualitative analysis. On the quantitative side, we analyzed PPP's TikTok posts. We categorized the engagement data (views, likes, shares, and comments) into three groups: videos that focus on

aesthetics, videos that focus on brands and trends, and videos that feature collaborations with influencers. This made it possible to compare the number of people who saw and interacted with each pattern. On the qualitative side, we examined PPP's brand style, their use of trends, and partnerships with influencers, particularly McVeigh (3.8 million followers) and Preslee Faith (2.7 million followers), to assess how these factors impacted brand visibility and consumer behaviour. These methods, combined, provided both numerical evidence of performance and a deeper understanding of how TikTok influences cultural meaning and consumer identity among Gen Z.

Discussion

The results of this case study underscore the pivotal role TikTok plays in driving brand growth among Gen Z audiences. The analysis clearly demonstrates that the success of the Pink Palm Puff (PPP) is not solely due to aesthetic appeal, but rather the strategic leveraging of TikTok's cultural dynamics, including algorithmic amplification, influencer partnerships, and consumer-driven trends.

The differences in performance between content categories demonstrate that TikTok is more than just a platform for promoting things; it is also a place where cultural meaning is created and shared. Content created by influencers garnered the most views (millions), while videos that focused on aesthetics received more likes and shares, indicating that they elicited more emotional engagement. This difference implies that influencers increase reach, while aesthetic storytelling improves connection and loyalty. These strategies help PPP garner both significant attention and a stronger connection with customers.

Second, the involvement of well-known TikTok stars, such as Embreigh (3.8 million followers) and Preslee Faith (2.7 million followers), highlights the close connection between PPP's growth and digital personalities already popular among teenagers. Embreigh's PPP collaboration, which is selling out in minutes, demonstrates the significant economic influence that influencers wield. This pattern aligns with the findings of other researchers on influencer marketing, which suggest that being relatable and appearing authentic are key factors in convincing Gen Z consumers. For PPP, influencers are more than just brand partners. They are also role models who make products more appealing to the public.

Ultimately, these findings are linked to the broader consumer culture of Generation Z. People don't simply purchase items because they are necessary; instead, they buy to establish their identity, feel a sense of belonging, and engage with the culture. When teens see their friends or trusted influencers promoting a product, buying it becomes a way for them to express themselves and feel good about themselves. TikTok amplifies this effect by combining entertainment, peer culture, and commerce into a single platform. This reinforces the idea that shopping is more of a social activity.

These findings provide valuable information; however, the study has some limitations. The analysis focused solely on TikTok and didn't examine other marketing channels, such as Instagram or brand-owned websites. TikTok trends also change rapidly, so strategies that work today may not remain effective for long. Future research could enhance this study by comparing

PPP with other Gen Z brands across various platforms. or by investigating the evolution of consumer loyalty following the initial surge of influencer-driven visibility.

The study reveals that TikTok is a cost-effective and high-impact platform for emerging fashion brands to promote themselves. PPP has become a successful Gen Z fashion brand and a model for how new businesses can grow in the digital age by leveraging influencer partnerships, visually consistent branding, and a deep understanding of consumer culture.

Conclusion

This research demonstrates how TikTok significantly contributes to the expansion of Gen Z fashion brands, exemplified by the case of Pink Palm Puff (PPP). The main findings reveal that collaborating with influencers generates significant attention for your brand, while content that focuses on appearance tends to evoke stronger emotional involvement. Combining these two approaches enhances brand loyalty and sales. The fact that Embreigh's PPP collection sold out and Preslee Faith's promotion demonstrates the power of influencer-led campaigns in persuading teens to purchase products.

Overall, TikTok influences Gen Z fashion brands by serving as a cultural hub and a platform for marketing products. Its algorithm helps brands reach a large audience with their content, and it also influences how young people express their identity, follow trends, and participate in consumer culture.

These insights show that for new business owners and marketers, success with Gen Z requires more than just traditional advertising. Instead, even small businesses can compete in the crowded fashion market by leveraging genuine influencer partnerships, consistent brand aesthetics, and short-form videos that stay current with the latest trends. When used wisely, TikTok is a cost-effective and efficient way for new brands to gain visibility, grow rapidly, and remain relevant to Gen Z.

Work Cited

"(PDF) The Role of User Interactions in Social Media on Recommendation Algorithms: Evaluation of TikTok's Personalization Practices From Users' Perspective." *ResearchGate*, https://www.researchgate.net/publication/375775130_The_Role_of_User_Interactions_in_Social_Media_ on_Recommendation_Algorithms_Evaluation_of_TikTok s_Personalization_Practices_From_User's_Perspective. Accessed 24 August 2025.

Author links open overlay panelLim Jia Ying a *et al.* (2025) *Social Media Marketing in industry 4.0: The role of Tiktok in Shaping generation z's purchase intentions*, *Procedia Computer Science*. Available at: https://www.sciencedirect.com/science/article/pii/S1877050925002868 (Accessed: 28 August 2025).

Priyanka, A.L. (no date) *(PDF) Marketing to gen Z: Understanding the preferences and behaviors of Next Generation*, *Research Gate*. Available at: https://www.researchgate.net/publication/383434299_Marketing_to_Gen_Z_Understanding_the_Pr eferences_and_Behaviors_of_Next_Generation (Accessed: 28 August 2025).

Lutfiana, P. (no date) *(PDF) the influence of Tiktok content on digital marketing of electronic products on e-commerce: A systematic literature review*, *Research Gate*. Available at: https://www.researchgate.net/publication/386291698_THE_INFLUENCE_OF_TIKTOK_CONTENT _ON_DIGITAL_MARKETING_OF_ELECTRONIC_PRODUCTS_ON_E-COMMERCE_A_SYSTEM ATIC_LITERATURE_REVIEW (Accessed: 28 August 2025).

Li, R. (no date) *(PDF) the impact of social media influencers on gen Z's online purchase decisions*. Available at: https://www.researchgate.net/publication/387778483_The_Impact_of_Social_Media_Influencers_o n_Gen_Z's_Online_Purchase_Decisions (Accessed: 28 August 2025).

The Effects of Achondroplasia on Quality of Life By Georgia Wei

Abstract

Achondroplasia is a heritable condition that affects human limb development. The condition is caused by a genetic mutation in the fibroblast growth factor receptor 3 (*FGFR3*) gene. This genetic mutation causes hyperactivity of the FGFR3 protein, which results in slowed bone growth. Many individuals with achondroplasia have disproportionate limb growth, resulting in shorter arms and legs but average-sized torsos. Numerous studies throughout have found achondroplasia to be associated with a lower quality of life. In this study, the quality of life of individuals with achondroplasia and those without it are compared using a modified Quality of Life in Short Stature Youth (QoLISSY) survey. The survey focused on physical, social, emotional, and future categories. From the results, the study found that individuals with achondroplasia report experiencing a lower overall quality of life when evaluated with a convenience comparison group, particularly in the social, physical, and future categories. These results suggest future treatment options to look more into targeting improvements in quality of life for individuals with achondroplasia in these categories specifically, including providing both physical and mental support.

Keywords

achondroplasia, quality of life, survey, physical quality of life

Introduction

Achondroplasia is a form of skeletal dysplasia that affects bone growth and development. It is characterized by shorter limbs (arms and legs), smaller hands and feet, short stature, a larger head, and a flat nose (Cleveland Clinic). Along with these characteristics, individuals with achondroplasia may be more likely to develop breathing problems (such as sleep apnea), obesity, bowed legs, extra fluid in the brain, and hearing problems (McDonald and De Jesus). Globally, around 250,000 people are affected by achondroplasia (Waller et al.).

Achondroplasia is caused by mutations in the fibroblast growth factor receptor type 3 (*FGFR3*) gene, which plays a crucial role in bone development. The FGFR3 protein regulates bone growth by inhibiting the proliferation of cartilage cells, thereby restricting the length of bones (MedlinePlus). In people with achondroplasia, mutations in the *FGFR3* gene cause the receptor to become overactive, leading to excessive inhibition of bone growth, resulting in the characteristic short stature.

Over 99% of achondroplasia cases are caused by one of two mutations in *FGFR3*, both of which result in the same protein change (MedlinePlus). About 80% of cases result from spontaneous mutations, while the remaining 20% are inherited from a parent with the condition (McDonald and De Jesus).

Achondroplasia can be diagnosed through various methods, including a physical examination, prenatal examination, genetic testing, and other x-rays or imaging scans (Cleveland

Clinic). Currently, there are no cures for achondroplasia. However, there are still many different treatment options available targeted towards specific symptoms or other complications caused by achondroplasia. Current treatment options include surgical procedures to help with excess fluid in the brain, a narrower foramen magnum, a compressed spine, and bowing in the legs; and treatments to help with hearing problems (McDonald and De Jesus). For targeting the limbs specifically, limb-lengthening surgeries could be considered (Paley). There are also a few ongoing drug trials targeted towards increasing growth in children with achondroplasia. Currently, vosoritide is the first approved medication option for treating achondroplasia, which limits signaling from the mutated FGFR3 gene (McDonald and De Jesus).

Although achondroplasia has been characterized for over 50 years, there is still a lot to learn about this condition, including exploring treatment options that can further improve the wellbeing and livelihood of individuals with achondroplasia (Pauli). In this paper, how achondroplasia can affect individuals and their quality of life across social, physical, and emotional aspects is discussed in hopes to inform future treatment options.

Quality of life is used as a way to measure an individual's well-being through evaluations of one's physical, emotional, and social health. While there are numerous ways to measure quality of life, it is typically assessed through standardized questionnaires. Standardized questionnaires enable researchers to compare results to a reference population and draw conclusions on how social, economic, and cultural factors can significantly affect one's quality of life.

The burden of living with medical conditions can result in changes to individuals' quality of life. Achondroplasia is a condition that affects physical appearance through its impacts on bone growth, which could also impact the wellbeing of individuals. As a result, several studies have examined the quality of life of those affected by achondroplasia.

Literature Review

Quality of Life in Children with Achondroplasia and their Parents

One study conducted in Germany examined the quality of life in children with achondroplasia, as well as their parents. The children that participated in this study ranged from ages 5 through 14 and had a medical diagnosis of achondroplasia but did not have any other significant medical conditions that could affect the quality of life assessment (Witt et al.). They were recruited through the German Association for People of Short Stature and their Families. In the study, quality of life was measured through two standardized questionnaires. The first type of questionnaire was called the Pediatric Quality of Life Inventory version 4.0 (PedsQL 4.0) questionnaire, which was completed by children ages 8-14 or by the parents of children ages 5-7. The second type was called the Short-Form-8 (SF-8) questionnaire, which was completed by all parents of patients participating in the study.

The PedsQL 4.0 questionnaire is designed to evaluate the physical, emotional, social, and school functioning in children and adolescents, through either self-reported or parent-reported

responses. The SF-8 questionnaire is used to measure parental quality of life by focusing on overall health, physical activity limitations, daily work difficulty, bodily pain, energy levels, social activity limitations, and emotional difficulties.

In total, 73 families participated in this study, with 73 responses received from the parents and 47 responses received from the children. Based on the PedsQL 4.0 responses, it was found that children with achondroplasia experience a significantly lower total quality of life relative to the healthy reference population based on the reports of the parents (M = 63.70, SD = 15.83 compared to a reference value of 82.70, $t(72) = -10.25$, $p \leq .01$) and the children themselves (M = 73.76, SD = 18.04 compared to a reference value of 83.84, $t(46) = -3.83$, $p \leq .02$) (Witt et al.). From the SF-8 responses, it was found that parents in the study similarly experienced a lower quality of life mentally in comparison to the reference population (M = 46.51, SD = 10.22 compared to a reference value of 53.25, $t(72) = -5.64$, $p \leq .01$), though their quality of life physically was the same as the reference population's (M = 50.50, SD = 8.49 compared to a reference value of 50.30, $t(72) = .20$, $p = .85$) (Witt et al.).

Physical and Mental Health in Adults with Achondroplasia

A different study conducted in Portugal examined the quality of life of adults with achondroplasia, specifically focusing on their physical activity, physical health, and mental well-being (Alves et al.). There were a total of 16 adults with achondroplasia, ten women and six men, who participated in this study. The participants ranged from ages 20-57, and all of them were medically diagnosed with achondroplasia.

It was noted that across the participants, 88% had no children, 44% lived with parents, 69% had a university degree, 63% had a job, 31% had received limb-lengthening treatments, and 25% experienced sleep apnea. The body compositions of the participants were measured by a Tanita MC780-PMA bioelectrical impedance analyser (a tool that analyzes body composition through 31 collected measurements). The participants also completed various questionnaires, including a socio-demographic questionnaire, the Short Form-36 (SF-36) version 2, the Brief Symptoms Inventory (BSI), and the International Physical Activity Questionnaire (IPAQ). This study found there was a strong correlation between the participants' general health and their M_MET (moderate-intensity activities), with M_MET being a huge indicator of a participant's general health. However, with regards to mental well-being, they did not find a significant correlation between physical activity and mental health. Lastly, in comparison to the general population of Portugal, it was found that adults with achondroplasia typically experience lower health-related quality of life and greater psychological distress. Specifically, the SF-36 scores for pain were much lower for adults with achondroplasia (48.69 ± 13.8) versus the general population (71.44 ± 24.27), and it was found that adults with achondroplasia experienced phobic anxiety 40% higher, psychoticism 32% higher, and anxiety 20% higher compared to the general population.

Impacts of Achondroplasia on Health-Related Quality of Life

A study across multiple countries in Latin America (Argentina, Brazil, and Colombia) examined the quality of life in both children and adults. The 172 participants of this study were all diagnosed with achondroplasia (Llerena Jr. et al.). Of the participants in this study, 7% had undergone a limb-lengthening procedure.

Different questionnaires were sent to the children and adults, with the Pediatric Quality of Life Inventory (PedsQL), Quality of Life in Short Stature Youth (QoLISSY), Pediatric Functional Independence Measure (WeeFIM), and Adolescent Pediatric Pain Tool (APPT) used for children and the 5-level EQ-5D version (EQ-5D-5L), Nottingham Health Profile (NHP), Brief Pain Inventory-Short Form (BPI-SF), and Work Productivity and Activity Impairment (WPAI) used for adults. The questionnaires were completed by the participants themselves or by their parents if the patient was a minor.

From the results of this study, the QoLISSY scores for individuals with achondroplasia were on average between 60-70, which was much lower than the average score of 82 for individuals with average stature (Llerena Jr. et al.). Among these scores, participants scored the lowest in the physical and social categories. The WeeFIM scores also increased with age, suggesting that individuals with achondroplasia gain more independence as they age, which increases their quality of life. In regards to pain, over 50% of adolescent participants reported they experience pain in at least one site, and 10% of adult participants experience at least three pain sites. The EQ-5D-5L Visual Analogue Scale (VAS) scores of the participants were also lower than those of the reference population, but their overall health status was similar. The participants also had high NHP scores, suggesting they have poorer subjective health (measured based on the sleep, energy, mobility, pain, isolation and emotional responses of the participants) which was more severe for participants who had limb-lengthening procedures. Overall, individuals with achondroplasia have a lower quality of life and are linked to various complex medical conditions, pain, discomfort, and less independence.

There are many existing studies regarding quality of life in people with achondroplasia. However, current questionnaires for those with achondroplasia focus on aspects regarding one's size and height. These questions make it difficult to compare to a reference population due to their bias towards these aspects that predominantly affect those with achondroplasia. To circumvent this, this study was conducted to examine how having achondroplasia can impact an individual's quality of life across multiple aspects (social, emotional, physical, etc.), refraining from including questions that specifically measure an individual's feelings toward their height.

Methods and Materials

To better quantify the impact of achondroplasia on one's quality of life across multiple aspects, the Quality of Life in Short Stature Youth (QoLISSY) questionnaire (Bullinger et. al) was modified and used to develop a survey focused on social, physical, and emotional health, and prospects for the future. Survey topics from the modified survey came from the QoLISSY

questionnaire, but any questions regarding height or size were removed to allow for better comparison to those without achondroplasia (Appendix A). The modified survey was reviewed by a mentor.

Survey participants were recruited through the Achondroplasia (https://www.facebook.com/groups/61999603825/) and the Dwarfism Facebook Group (https://www.facebook.com/groups/dwarfismworld) as well as broadcasting to peers and colleagues. All survey responses were completed between July and September of 2025. Both those with and without achondroplasia were encouraged to complete the survey. The participants were aware that participation was voluntary and that all survey responses would be anonymous, collected through an online Google Form. As the surveys were anonymous, names, location, or other personal identifying information were not collected to ensure confidentiality. The survey consisted of 12 quality of life questions and 4 demographic questions (Appendix A). The quality of life questions asked participants to respond how frequently they experienced the statements in the past 30 days using a Likert scale, a tool used to quantify the beliefs and experiences of others using a scale. The Likert scale used was from 1 to 5, with 1 representing an individual who never experiences the statement, and 5 representing an individual who experiences the statement very often. All quality of life questions except for one were positively worded, with a higher score suggesting a higher quality of life. As question 1 "I get teased and bullied by others" is phrased negatively, negative reverse scoring was used to ensure that higher mean scores across all quality of life survey questions suggest a higher quality of life. For responses that were left blank, overall mean scores were calculated for those questions using the remaining responses. The 4 demographic questions collected data regarding the participant's age, race/ethnicity, current gender identity, and whether or not they are diagnosed with achondroplasia. For the questions regarding race/ethnicity and current gender identity, the participants had the option to choose "Prefer not to answer".

In order to see whether the collected data is significant, two-tailed *t*-tests were run. A significance threshold of $p \leq 0.05$ was used.

Results

46 people completed the survey, though one person was excluded due to uncertainty in achondroplasia diagnosis. Of the remaining 45 participants, 24 (53.3%) have been diagnosed with achondroplasia, 34 (75.6%) identify as women, and 28 (62.2%) are white (Fig 1). The age of participants ranged from 5 to 61, with a mean age of 35.19 ± 15.96 for participants without achondroplasia ($n = 21$, 95% Cl [28.36, 42.02]) and 31.17 ± 13.83 for participants with achondroplasia ($n = 24$, 95% Cl [25.63, 36.70]) (Fig 2). The majority of participants completed the survey themselves, but anyone under the age of 10 had a caregiver complete the survey on their behalf.

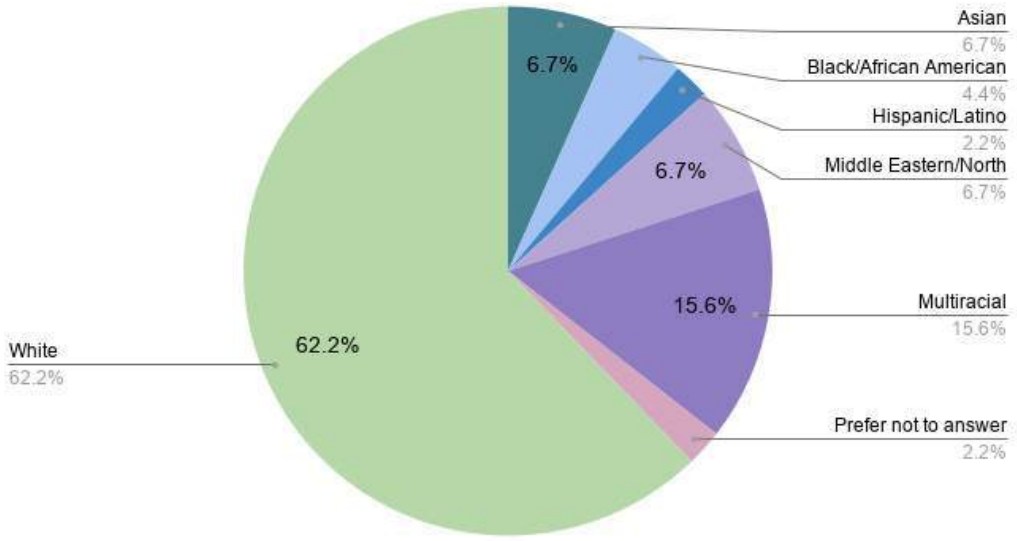


Fig 1. The self-reported racial/ethnic background of study participants.

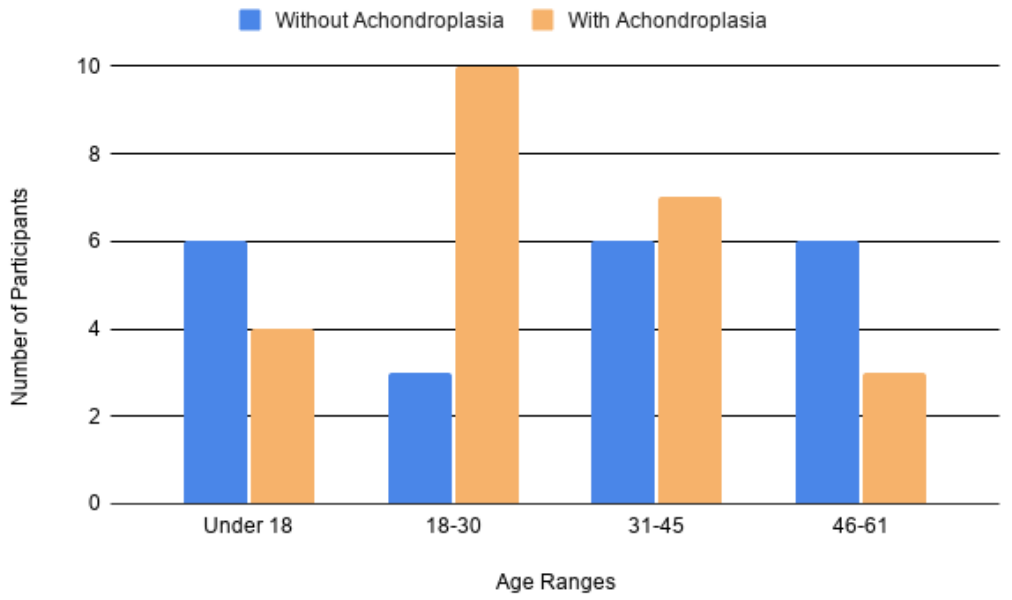


Fig 2. The age of participants from this study represented by different age ranges, separated by participants with and without achondroplasia.

Across all quality of life questions, the average scores of participants without achondroplasia from each category ranged from 3.87 to 4.46 with a mean score of 4.21 ± 0.85 (95% Cl [4.01, 4.42]), while average scores of participants with achondroplasia from each category ranged from 3.33 to 3.49 with a mean score of 3.42 ± 1.02 (95% Cl [3.32, 3.67]). The

mean scores of participants without achondroplasia were also greater than the mean scores for those with achondroplasia across each question individually (Appendix B).

To get a better understanding of which aspects of quality of life most affected those with achondroplasia, the survey questions were grouped into four categories: social, physical, emotion, and future (Appendix A). Those with achondroplasia reported a lower score for each category, with the greatest difference in physical quality of life aspects (score difference of 0.99, $p \leq 0.0001$) (Fig 3).

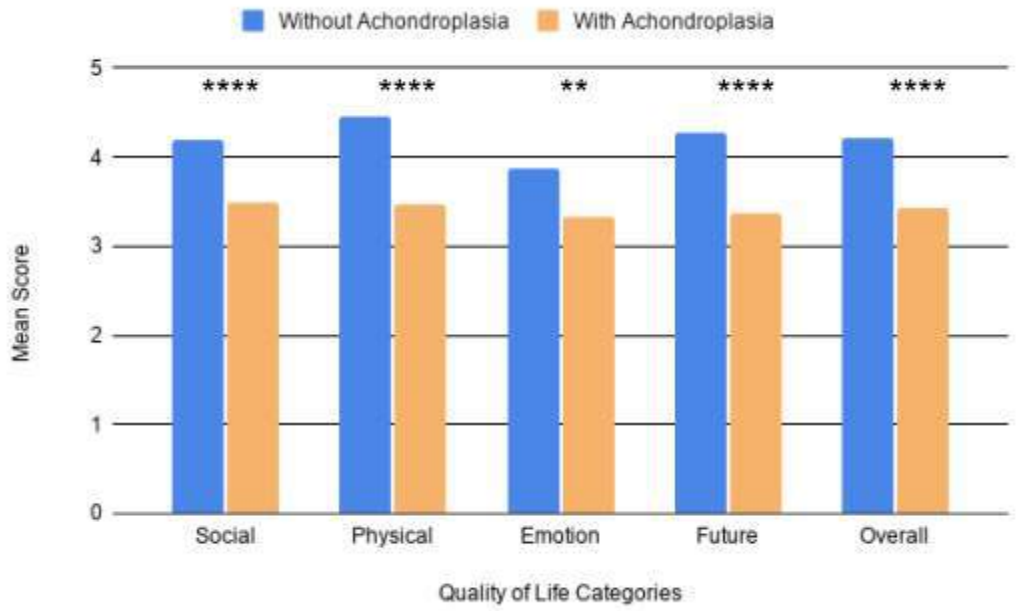


Fig 3. The mean scores of participants across the four assessed quality of life categories and across all quality of life questions overall, separated by participants with and without achondroplasia. Two-tailed unpaired *t*-test, **: $p \leq 0.01$, ****: $p \leq 0.0001$.

When broken down by individual questions, those with achondroplasia reported the greatest difference in quality of life scores in questions regarding energy (difference of 1.13, $p \leq 0.001$) and clothing (difference of 1.29 , $p \leq 0.0001$) (Fig 4). In contrast, the difference in scores between participants with and without achondroplasia were the smallest for the questions regarding confidence and appearance satisfaction (difference < 0.5, $p > 0.05$) (Fig 4).

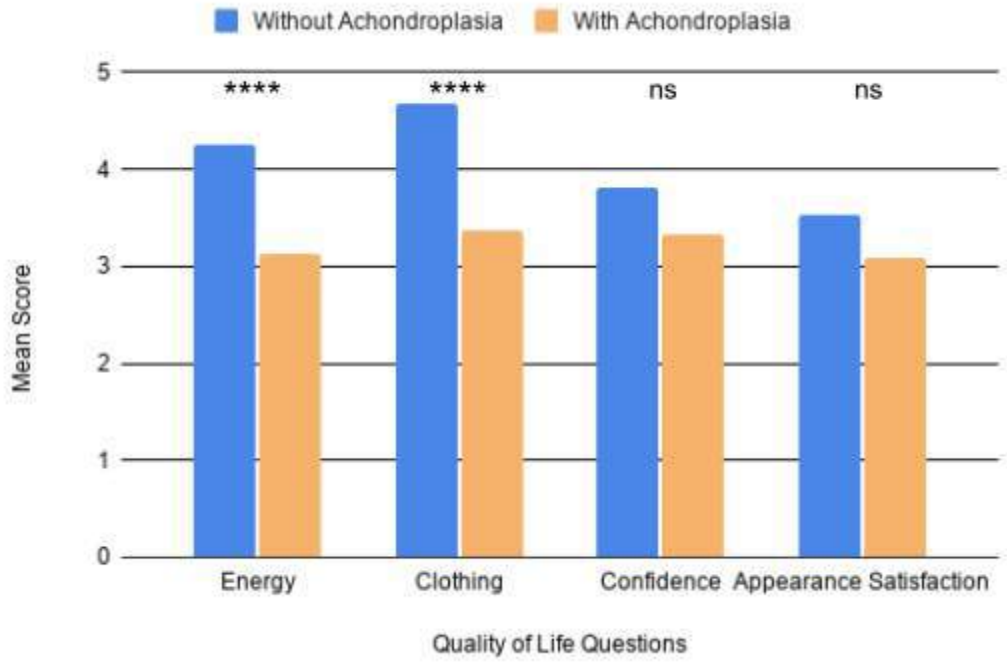


Fig 4. The mean scores of participants for the specific quality of life questions regarding energy, clothing, confidence, and appearance satisfaction, separated by participants with and without achondroplasia. Two-tailed unpaired *t*-test, ***: $p \leq 0.001$, ****: $p \leq 0.0001$, ns (not statistically significant): $p \geq 0.05$.

Discussion

The responses from this survey reveal a lower quality of life (QoL) for participants with achondroplasia in comparison to participants without. The results of the study suggest that out of the areas included in the survey, achondroplasia most heavily impacts an individual's quality of life physically, their ability to find appropriate clothing, and their ability to have enough energy to fulfill daily tasks.

However, although the smallest difference in scores were found in the questions about confidence and appearance satisfaction (difference of 0.48, $p > 0.05$ for confidence; difference of 0.44, $p > 0.05$ for appearance satisfaction), this does not suggest that people with achondroplasia have high levels of confidence and appearance satisfaction. Instead, with mean scores of 3.33 ± 1.09 for the question about confidence and 3.08 ± 1.14 for the question about appearance satisfaction from participants with achondroplasia, the small difference in scores may be attributed to random variation, alongside the lower scores for these two questions from participants without achondroplasia as well (mean scores of 3.81 ± 0.81 for confidence and 3.52 ± 0.75 for appearance satisfaction). This suggests the possibility that both people with and without achondroplasia can have similar ratings of confidence and satisfaction with their appearance.

Some limitations of this study include that it had a small sample size and targeted recruitment to an achondroplasia support group and local community. By targeting a support group, the results could be skewed towards those more severely affected by the condition. Additionally, the reference population was more geographically limited. Differences in age distribution (Fig 2.) could have skewed results as well, as different age groups will answer and interpret questions differently due to differences in their social roles, independence, and daily activities.

In addition to improving studies, current treatment options are still limited in their accessibility and success at increasing quality of life, which is why actionable changes are needed to improve the quality of life of people with achondroplasia while waiting for better treatment options. Some strategies include education surrounding achondroplasia, allowing for increased inclusivity in both schools and workplaces, confidence building, and body positivity. Additionally, providing accommodations for physical activities, particularly in school-aged children, and making the surrounding school and/or work environment more accessible could drastically improve individuals' quality of life. On a grander scale, there is a need for increased diversity in clothing sizes, allowing for more options when picking out clothes. Together, these efforts can bring about positive changes for those with achondroplasia and improve their quality of life.

Conclusion

The results of this study are consistent with other published quality of life studies (Witt et al.; Alves et al.; Llerena Jr. et al.). Together, these studies suggest that having achondroplasia can lead to a lower quality of life across different aspects (social, physical, emotional, future).

Moving forward, studies need to expand the participant pool and diversity of participants in order to better understand how an achondroplasia diagnosis impacts quality of life across larger regions and age groups, and in combination with other sociodemographic factors. Additionally, future studies could explore how novel or existing treatment options may improve a patient's quality of life, providing a better understanding of what the needs are for those with achondroplasia.

Conflict of Interest

The author declares that there are no conflicts of interest related to this work.

Appendix

Appendix A
Online Quality of Life Survey Questionnaire

Demographic Questions

The following questions are asked for demographic purposes only. Again, all of your responses for this entire survey are anonymous.

1. What is your age?
2. Have you been diagnosed with achondroplasia?
    1. Yes
    2. No, and I am not diagnosed with any types of dwarfism
    3. No, but I am diagnosed with a different type of dwarfism
    4. Unsure
3. Race/Ethnicity (select one)

Please choose the option that BEST describes your identity.

1. American Indian or Alaska Native
2. Asian
3. Black or African American
4. Hispanic or Latino
5. Middle Eastern or North African
6. Native Hawaiian or Other Pacific Islander
7. White
8. Multiracial
9. Prefer not to answer

4. Current Gender Identity (select one)

Please choose the option that BEST represents your current self.

1. Woman
2. Man
3. Nonbinary
4. Prefer not to answer

Quality of Life Questions

The following questions will ask you to respond based on a 1-5 scale. Select the number that honestly reflects how frequently you experienced these things in the last 30 days.

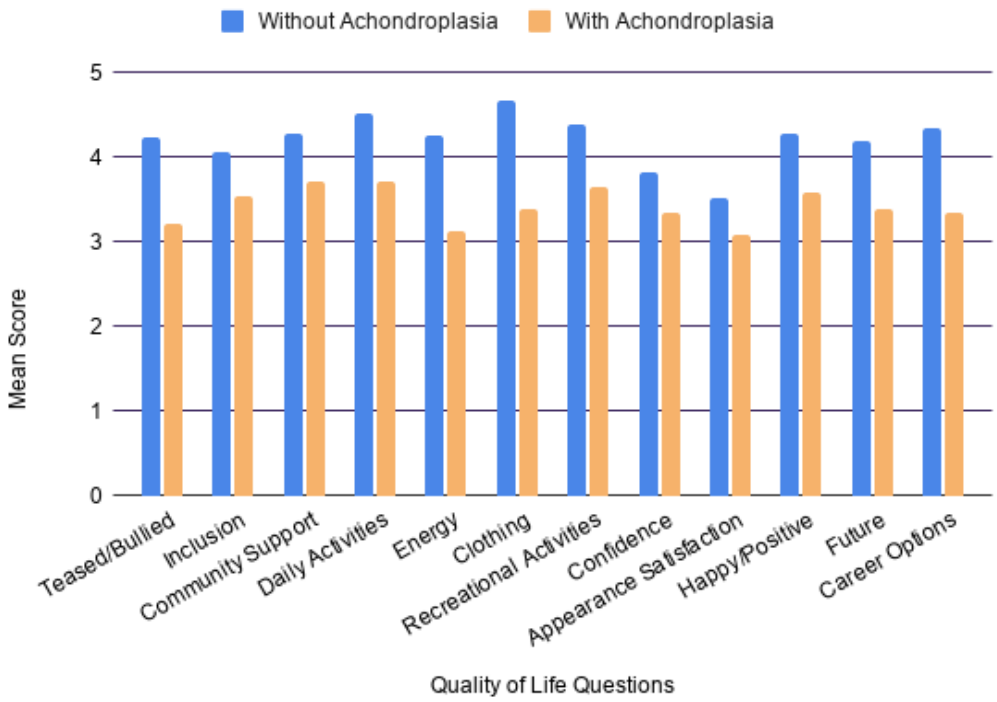
1. I get teased and bullied by others. (Social)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often

2. I feel included by my classmates and others my age. (Social)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
3. I feel supported by my community in who I am. (Social)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
4. I am able to carry out my daily activities without difficulty. (Physical)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
5. I have energy throughout the day to complete my daily activities without fatigue. (Physical)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
6. I can find and wear clothing that fits me comfortably and is appropriate for my age. (Physical)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
7. I am able to participate in recreational activities that I enjoy (e.g. playing sports, hanging out with friends, extracurriculars). (Physical)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
8. I feel confident in myself. (Emotional)

    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
9. I feel satisfied and pleased with the way that I look. (Emotional)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
10. I feel happy and positive. (Emotional)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
11. I feel optimistic about my future. (Future)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often
12. I feel like I have numerous career options (or for when I am older). (Future)
    1. 1 - Never
    2. 2 - Rarely
    3. 3 - Sometimes
    4. 4 - Often
    5. 5 - Very often

Appendix B

Mean Quality of Life Scores Across Each Survey Question



Works Cited

Alves, Inês, et al. "Physical Activity and Psychosocial Outcomes in Adults with Achondroplasia: An Exploratory Study." *International Journal of Environmental Research and Public Health*, vol. 21, no. 9, 2024, p. 1160. *PubMed Central*, https://doi.org/10.3390/ijerph21091160.

Bullinger, Monika, et al. "Assessing the Quality of Life of Health-Referred Children and Adolescents with Short Stature: Development and Psychometric Testing of the QoLISSY Instrument." *Health and Quality of Life Outcomes*, vol. 11, art. 76, 2013. *BioMed Central*, https://doi.org/10.1186/1477-7525-11-76.

Cleveland Clinic. "Achondroplasia: Symptoms, Treatment, Causes & Diagnosis." *Cleveland Clinic*, 10 Dec. 2021, https://my.clevelandclinic.org/health/diseases/22183-achondroplasia. Accessed 17 July 2026.

Llerena, Jaime, Jr., et al. "Lifetime Impact Study for Achondroplasia (LISA): Findings from an Observational and Multinational Study Focused on Health-Related Quality of Life in Individuals with Achondroplasia in Latin America." *Genetics in Medicine Open*, vol. 2, 2023, p. 100843. *ScienceDirect*, https://doi.org/10.1016/j.gimo.2023.100843.

McDonald, E. J., and O. De Jesus. "Achondroplasia." *StatPearls*, StatPearls Publishing, 23 Aug. 2023, https://www.ncbi.nlm.nih.gov/books/NBK559263/. Accessed 17 July 2026.

MedlinePlus. "FGFR3 Gene: Fibroblast Growth Factor Receptor 3." *National Library of Medicine*, 24 June 2020, https://medlineplus.gov/genetics/gene/fgfr3/. Accessed 17 July 2026.

Paley, Dror. "Extensive Limb Lengthening for Achondroplasia and Hypochondroplasia." *Children*, vol. 8, no. 7, 2021, p. 540. *MDPI*, https://doi.org/10.3390/children8070540.

Pauli, Richard M. "Achondroplasia: A Comprehensive Clinical Review." *Orphanet Journal of Rare Diseases*, vol. 14, art. 1, 2019. *BioMed Central*, https://doi.org/10.1186/s13023-018-0972-6.

Waller, D. Kim, et al. "The Population-Based Prevalence of Achondroplasia and Thanatophoric Dysplasia in Selected Regions of the US." *American Journal of Medical Genetics Part A*, vol. 146A, no. 18, 2008, pp. 2385–89. *Wiley Online Library*, https://doi.org/10.1002/ajmg.a.32485.

Witt, S., et al. "Quality of Life of Children with Achondroplasia and Their Parents - A German Cross-Sectional Study." *Orphanet Journal of Rare Diseases*, vol. 14, art. 194, 2019. *BioMed Central*, https://doi.org/10.1186/s13023-019-1171-9.

Does Interest-Rate Sensitivity Organise Volatility Transmission in Indian Equity Markets? By Aarav Shah

Abstract

Volatility connectedness studies between equity sectors in India have been done many times, but often by simply adding all available sectors and describing the resulting structure. None has asked whether the network is organised around a stated economic characteristic. This paper subjects it to the test. Prabu et al. (2020) categorize Indian sectors based on their bank credit dependency as Type 1: Auto, Banks, Financial Services and Realty; and Type 2: Information Technology, Media, Fast Moving Consumer Goods, Metal and Pharmaceuticals. The generalised variance decomposition of Diebold and Yilmaz (2012) is used to estimate connectedness over rolling 200-day windows, and the volatility is captured by a GARCH(1,1) conditional standard deviation using six NSE sectoral indices over 4,554 trading days from April 2008 to September 2026. There are six sectors that allow exactly ten partitions into groups of three, and the modularity and within-group co-movement minus cross-group contagion are scored for each. The credit-dependence partition is the top performer on both, ahead by 6.572 percentage points, and it is the largest difference out of the nine partitions. It is the first moving block bootstrap in 96.4 per cent of replications, and the first in 56.1 per cent of rolling windows (as compared to 10 per cent by chance). The result does not depend on the standardisation, the forecast horizons and the volatility measure based on a range. The modularity of each partition is negative, so it is a comparative statement, not a statement of two clean communities – credit dependence is the best of the available partitions. A second test examined Reserve Bank of India tightening and easing phases and found no difference, but because there were only seven usable phases, the design can detect differences only above 5.54 percentage points, and that question remains open.

Keywords: volatility connectedness; sectoral indices; interest-rate sensitivity; monetary policy; India; variance decomposition; JEL classification: G10, E52, C32

1. Introduction

Stock market sectors don't operate independently of one another. If one becomes volatile, then other ones tend to become volatile shortly thereafter. Diebold and Yilmaz (2009, 2012, 2014) devised a method to do so, decomposing the error variance of each sector's prediction into components attributable to all other sectors and viewing the result as a network. It has been applied to the Indian sectoral indices by several studies (Chatziantoniou, Gabauer and Marfatia, 2022; Chowdhury and Irfan, 2022; Maurya, Bansal and Mishra, 2025; Mishra, Anand K and Kappagantula, 2024).

A design is shared by these studies. Once results are available, every available sector enters the system, the degree of connectedness is estimated, and the structure is described. The approach is exploratory. It displays the network that shows up, but it does not assess if the network is organized around some specific economic characteristic. This paper poses the second question. We select a grouping of Indian sectors prior to estimating them, based on the way they are financed, and then check its validity for the transmission of volatility.

1.2 The grouping and the problem

The grouping is provided by Prabu, Bhattacharyya and Ray (2020). They split the market into two types, based on the expectation that the policy would affect the auto, banks, financial services and realty group (Type 1), but not the information technology, media, fast moving consumer goods, metal and pharmaceuticals group (Type 2), and then tested how the Indian sectoral indices reacted to monetary policy announcements. The split is based on the bank credit dependence. The ratio of bank loans to total liabilities was on average 22.7 per cent for realty, 14.8 per cent for auto and 3.2 per cent for information technology between 2005 and 2016. Their findings were similar to that of the split and Khuntia and Hiremath (2019) concluded the same with another design. Both test sectors one at a time, no checking to see if responsive sectors move together.

A second opening is from method. Buchwalter, Diebold and Yilmaz (2026) expanded the connectedness framework to clusters, where shocks are correlated within the cluster but not across clusters. Toward the end of their paper they say that "applying the incorrect clustering may produce misleading results" and that "empirical cluster classification is an open question. Rather than assuming one clustering, this paper enumerates all the possible clusterings, and asks which is best from the data.

1.2 Data, method and findings

The balanced sample of trading days is from 15 April 2008 to 15 September 2026 for six NSE sectoral indices. Nifty Bank, Nifty Realty and Nifty Auto form Type 1. Type 2 consists of Nifty FMCG, Nifty Pharma and Nifty IT. Nifty Financial Services is not listed in the data extract, as Prabu, Bhattacharyya and Ray consider it a Type 1. Volatility is the conditional standard deviation from a GARCH(1,1) model (Bollerslev, 1986) in logs. Connectedness is defined following Diebold and Yilmaz (2012) based on a rolling 200-day VAR(2) and the generalised forecast error variance decomposition (Koop, Pesaran and Potter, 1996; Pesaran and Shin, 1998) that does not require any ordering of the sectors.

There are exactly ten ways for the six sectors to be divided into two groups of three. Each is evaluated in terms of modularity (Newman and Girvan, 2004), which is the difference between observed within-group connectedness and a degree-preserving null model and the simpler difference between within-group co-movement and cross-group contagion. Next, we identify 21 RBI policy phases based on the RBI's published statements and examine whether co-movement increases when the RBI tightens within the Type 1.

On both indicators, the Type 1 and Type 2 split beat the entire spectrum of the other nine by more than 6.572 percentage points. It is the first among 500 replications of the moving-block bootstrap (p = 0.036), and among 4354 rolling windows it ranks first 56.1 per cent of the time (10 per cent expected by chance). The result is at three forecast horizons, with standardisation of each series, and with a second volatility measure. Co-movement is 14.42 per cent of system variance within Type 1 compared to 5.41 per cent within Type 2. There is one qualification to that – all the partitions are of negative modularity, and the market is not split into two neat communities. The credit-sensitivity split is the most favorable of 10 possible divisions, and not a full description of the network.

The second question is one that cannot be answered here. The number of usable policy phases within the estimation window is seven (three phases of policy tightening and four of policy easing), and the smallest difference that design can detect is 5.54 percentage points against an observed 2.52. The difference, which is virtually all due to the 2010-2011 tightening cycle, changes sign under the other measure of volatility. The design is not powerful enough, not that the effect doesn't exist. The overall results indicate that the organising principle of the Indian equity markets for the transmission of volatility is credit dependence, and this dependence is not only when the policy is in force, but also throughout the policy cycle.

1.3 Contribution

The paper is an experimental test of a grouping that has been selected a priori, in an area where Indian sectoral work has been exploratory. It is complementary, not a replacement, to the data-driven cluster detection of Gudmundsson and Brownlees (2021) and tests a theory-specified grouping against all alternatives that the data allows. It also incorporates a conditioning on the RBI policy cycle, which has not been done in any prior study.

Three nearby studies are to be separated. They call it after the event - Mishra, Anand K and Kappagantula (2024) find banking and financial services transmitting and pharmaceuticals, FMCG and IT receiving, so that's our grouping in their output, and they're working on returns. Panda, Anjaly B., Panda and Mohanty (2026) apply an event study approach in eight countries and Tomar and Kesharwani (2022) model each sector independently. None assesses whether there is any change in the inter-sectoral relationships over the policy cycle.

Asymmetric in Indian literature typically refers to the leverage effect, which is the concept that bad news will have a larger impact on volatility than good news of an equal magnitude (Sahoo and Kumar, 2023). In this context, it means only to tighten and to ease. Lastly, a word on the development of the study. It was created to test the monetary policy hypothesis, and the partition test started out as a way to see if the grouping made sense. That finding should therefore not be interpreted as a pre-registered test, but rather as an exploratory finding that has passed the robustness checks in Section 4. Both outcomes are reported and hypothesis that failed is stated as initially formulated.

The second section examines the literature and formulates the hypotheses. The description of data and method is given in Section 3. Section 4 presents results. Section 5 discusses them, and Section 6 concludes.

2. Review of Literature

2.1 Measuring connectedness

Diebold and Yilmaz (2009) proposed the spillover index, which decomposes the forecast error variance of each market into the share of each of the other markets. Variable ordering was an issue with its Cholesky factorisation. Diebold and Yilmaz (2012) did this with a generalised decomposition, along the lines of Koop, Pesaran and Potter (1996) and Pesaran and Shin (1998), and henceforth, the measures used are: total, directional, net, and net pairwise. Diebold and Yilmaz (2014) showed that the decomposition matrix is the

adjacency matrix of a weighted directed network. There are two refinements that are not used: the time-varying parameter VAR of Antonakakis, Chatziantoniou and Gabauer (2020), which was not used because the second hypothesis requires window specifications to be related to a dated phase, and the frequency decomposition of Baruník and Křehlík (2018).

This paper builds on the framework of Buchwalter, Diebold and Yilmaz (2026). They enable network nodes to cluster and shocks to be correlated within clusters, but orthogonal across them, and connectedness can thus be decomposed into three elements that add up to one: own variance, co-movement within a cluster, and contagion across clusters. The cross-cluster component became the largest contributor to the total connectedness in sixteen country equity markets, whereas the within-cluster component remained unchanged. They also state that applying the incorrect clustering will affect the results, and that the empirical cluster classification is an open question. This paper also borrows from the community detection literature since it asks whether the proposed grouping describes network structure.

Modularity was introduced by Newman and Girvan (2004), who compare the connections made within proposed groups with the connections made by a degree-preserving null. It is positive if a grouping explains more than chance, and negative if it explains less, and corrects for some nodes having more links to the system than others, which is important here because Indian sectors vary widely in their extent of involvement. Assortativity is reported along with it by Newman (2002). Buchwalter, Diebold and Yilmaz (2026) see the detection of groups from the data, as done by Gudmundsson and Brownlees (2021) and Brownlees, Gudmundsson and Lugosi (2022), as the way forward. The estimation also relies on common econometric tools as described in Section 4 (Engle, 1982; Bollerslev, 1986; Garman and Klass, 1980; Dickey and Fuller, 1979; Phillips and Perron, 1988; Kwiatkowski, Phillips, Schmidt and Shin, 1992; Newey and West, 1987; Künsch, 1989).

2.2 The inter-sectoral connectedness

The situation is similar with the applications to the Indian sectors. The TVP-VAR is employed here because Chowdhury and Irfan (2022) employed TVP-VAR on sectoral returns from the NSE for the period of 2011 to 2020 with a generalised decomposition, as no theory orders the sectoral shocks. Chatziantoniou et al. (2022) identified that spillover increases when crises happen, during the 2014 election and demonetisation, while Maurya et al. (2025) discussed fourteen sectors from 2007 to 2023.

The matter of Mishra, Anand K and Kappagantula (2024) is the most important. Based on the twelve NIFTY sectoral indices, they employed an asymmetric TVP-VAR model during the pandemic period from 2017 to 2023 and discovered that during the pandemic, the overall connectedness has increased from 67.07 to 76.10 per cent, and the net positions of the models obtained during the pandemic replicate the sectoral grouping tested in this paper: the banking sector, financial services and private banks were the transmitters, and the pharmaceuticals, fast-moving consumer goods and information technology were the receivers. They explain that pattern only after the estimates have been made and work on returns, not volatility. Asymmetry in Sahoo and Kumar (2023) and in the literature generally refers to the leverage effect, while in this paper it is used only for the difference between tightening and easing.

2.3 Interest-rate sensitivity as a feature of sectors

The grouping and the reasons are provided by Prabu, Bhattacharyya and Ray (2020). They pointed out that the aggregate index obscures differences between the sectors, and they split the market before making any estimation: Type 1 included auto, banks, financial services and realty; Type 2 included information technology, media, fast-moving consumer goods, metal and pharmaceuticals.

The basis is dependence on bank credit. The average ratio of bank loans to total liabilities was 22.7 per cent for realty, 14.8 per cent for auto, 7.3 per cent for fast-moving consumer goods and 3.2 per cent for information technology between 2005 and 2016. Housing made up 51.7 per cent of personal bank credit and vehicle loans another 11.0 per cent, against 1.2 per cent for consumer durables. Based on identification using heteroscedasticity, they identified significant effects for banking, financial services and auto and no effects for the Type 2 sectors, on the announcements between 2004 and 2016.

There are two studies in favor of the split. Khuntia and Hiremath (2019) investigated RBI announcements in relation to five NSE sectoral indices and identified that banking, financial services and auto had the highest prominence in the communication of policy, which is important as auto would otherwise be dependent on a single source. Chauhan, Suri, Twala, Priyadarshi and Ali (2025) employed an autoregressive distributed lag model in monthly NSE data from 2012-24 and concluded that foreign institutional investment had an impact on all the sectors except the IT sector. It is an open review and is therefore only considered as supporting evidence.

In addition to return, rate sensitivity is also apparent in volatility; hence, this paper quantifies volatility connectedness. Akimov, Lee and Stevenson (2019) directly included interest rate volatility in the variance equation of a GARCH-in-mean model, Ghosh, Latha and Gupta (2018) and Uddin, Qamar and Ahmad (2022) applied GARCH(1,1) for Indian non-banking financial companies and the Nifty Bank Index, respectively.

2.4 Monetary transmission and network structure

Monetary transmission in India is not uniform, and the non-uniformity varies by direction. Das (2015) reported asymmetric adjustment in bank lending rates, while Singh (2011) examined the differences between bank lending rates in tight and easy policy cycles. Mishra, Montiel and Sengupta (2016) used a structural VAR and Pandit, Mittal, Roy and Ghosh (2006) examined the bank lending channel. Tomar and Kesharwani (2022) divided the policy variables into positive and negative parts and measured them individually for each sector.

A smaller body of literature asks whether policy alters the structure of the network, and the results are mixed. Anghel and Caraiani (2024) identified weak sign asymmetry around FOMC meetings, while Kamboj, Rao and Agrawal (2026) identified asymmetry at the asset-class level. Others measured policy spillover based on network topology (Hou, Li, Wu, Zang and Quach, 2025), shocks in the form of shifts in the yield curve (Inoue and Rossi, 2019), and conditional connectedness across countries (Nyakurukwa and Seetharam, 2023). Closest in terms of subject is Panda, Anjaly B., Panda and Mohanty (2026) – which includes 8 countries, including India – but employs an event study and does not generate any variance decomposition.

2.5 Research gaps

Four gaps follow. The first is in design: all the Indian sectoral connectedness studies reviewed here read the network after all sectors have been entered, and no one specifies a grouping beforehand and checks if the network respects it. The second is in approach: Buchwalter, Diebold and Yilmaz (2026) identify cluster misclassification as the primary challenge in clustered connectedness research. Gudmundsson and Brownlees (2021) and Brownlees, Gudmundsson and Lugosi (2022) do so by searching for groups in the data, but no study has evaluated an a priori group defined by theory against the alternatives a dataset offers. The third is conditioning: the Indian studies have been on samples with complete RBI cycles, but condition on crises, elections and the pandemic; the transmission literature ends at bank interest rates. The fourth is in the sector-level work: Prabu, Bhattacharyya and Ray (2020), Khuntia and Hiremath (2019) and Tomar and Kesharwani (2022) all measure a sector at a time – and none of them can say whether the responsive sectors move together.

2.6 Hypotheses

Two hypotheses follow. The grouping constructed around interest-rate sensitivity should be different from the alternatives, if volatility transmission is organised around interest-rate sensitivity. The number of ways in which 6 sectors can be divided into two sets of 3 sectors is exactly 10.

H1: The excess of within-group co-movement over cross-group contagion is greatest for the Type 1 and Type 2 partitions of six sectors into two groups of three. The null hypothesis is that it doesn't do any better than any other partition.

The second is a question of cause. The studies by Singh (2011) and Das (2015) reveal that easing and tightening have different impacts on transmission, and Prabu, Bhattacharyya, and Ray (2020) reveal that policy is effective for the Type 1 sectors and not for the Type 2 sectors. Combined, they suggest that Type 1 sectors should become more clustered when they all experience a credit squeeze simultaneously.

H2: The co-movement for the Type 1 group is higher in RBI tightening periods than in easing periods. The null is that co-movement is not different across phases and the grouping is structural and not cyclical.

3. Data and Methodology

3.1 Data

The daily closing values of six sectoral indices of NSE are used in the study. Group 1 consists of Nifty Bank, Nifty Realty and Nifty Auto, while Group 2 consists of Nifty FMCG, Nifty Pharma and Nifty IT. The classification is based on Prabu, Bhattacharyya and Ray (2020) and is predetermined prior to the estimation. They also have Nifty Financial Services in Type 1, which is not included in the data extract, and Media and Metal in Type 2, which are also not included in the data extract. Three sectors from each group are therefore used in the test, instead of the four and five of the original classification.

The sample runs from 15 April 2008 to 15 September 2026. Nifty Realty was launched in April 2008, and the start date is determined by it since a balanced panel needs all six series. Merging on common trading days leaves 4,554 observations and 4,553 returns, with ten dates having been dropped for missing values.

Returns are differences in closing values on a daily basis multiplied by 100. Summary statistics are given in Table 1. Each series is negatively skewed, leptokurtic, and Jarque-Bera rejects normality for each of the six series at 1 per cent. Realty has the highest (37.9 per cent annualised), and FMCG has the lowest (17.6 per cent).

Table 1. Descriptive statistics of daily log returns (per cent), 4,553 observations

Index	Type	SD	Annual SD	Min	Max	Skewness	Ex. kurtosis
Nifty Bank	1	1.663	26.4	-18.31	17.24	-0.20	10.63
Nifty Realty	1	2.388	37.9	-27.06	21.25	-0.52	8.88
Nifty Auto	1	1.422	22.6	-14.91	14.00	-0.20	8.43
Nifty FMCG	2	1.107	17.6	-11.20	8.30	-0.18	7.00
Nifty Pharma	2	1.193	18.9	-9.35	11.16	-0.15	6.61
Nifty IT	2	1.510	24.0	-12.49	11.72	-0.26	6.38

Notes: Annualised SD uses a 252-day year. Mean daily returns lie between -0.004 and 0.059 per cent. Jarque-Bera rejects normality for all six at $p < 0.001$.

According to the results of the Augmented Dickey-Fuller and Phillips-Perron tests, a unit root in the log volatility series can be rejected at 1 per cent for all six sectors, and thus, these series are entered in levels. All six are also found to be non-stationary in levels according to the KPSS test (Kwiatkowski, Phillips, Schmidt and Shin, 1992), with statistics ranging from 1.73 for Pharma to 12.40 for Realty compared to the 5 per cent critical value of 0.463. It is the typical signature of long memory in volatility, as the two families of tests disagree. The VAR is estimated in levels, based on the results of the ADF and Phillips-Perron tests, and this lack of agreement is reported in Section 6 as a limitation.

3.2 Measuring volatility

Each of the sectors' returns is modelled as a GARCH(1,1) process with Student-t errors (Engle, 1982; Bollerslev, 1986) and the conditional standard deviation is used as an estimate of volatility. The persistence ranges from 0.965 (FMCG) to 0.996 (Bank), which are all normal for daily equity data and explain the stationarity results above.

An alternative approach was considered that involves a range. The Garman and Klass (1980) estimator is based on the open, high, low and close of the daily time series. This was not a suitable index to calculate on, as these values are incomplete prior to October 2011 for Auto, Pharma and FMCG, and it would have cost the financial crisis and the 2010-2011 tightening cycle. This is instead used as a robustness check on the shorter period.

3.3 Estimating connectedness

The definition of connectedness is based on Diebold and Yilmaz (2012). The six log volatility series are placed into a VAR(p) with p chosen by Akaike information (AIC2) and Schwarz information (SIC) criteria (two lags is the baseline, one lag a robustness check).

The forecast error variance for each sector is decomposed at a horizon of H steps following the generalised version of Koop, Pesaran and Potter (1996) and Pesaran and Shin (1998) where all the variables are assumed to be ordered first. The order of the sectors is important since there is no theory that stipulates which sector's shocks occur first, as pointed out by Chowdhury and Irfan (2022). The share of sector i's forecast error variance due to shocks from sector j is

$$\theta ij(H) = [\ \sigma jj^{-1}\ \Sigma h\ (e'i\ Ah\ \Sigma\ ej)^2\ ]\ /\ [\ \Sigma h\ (e'i\ Ah\ \Sigma\ A'h\ ei)\ ] \ldots. \quad (1)$$

where Ah is the moving average coefficient matrix of the VAR, Σ is the residual covariance matrix and ei is a selection vector. The generalised decomposition does not require that rows sum to one, so that for each row, each element is divided by the sum of the row, e.g., $\tilde{\theta}ij(H) = \theta ij(H) / \Sigma j\ \theta ij(H)$.

The average off-diagonal share (in percentage terms) is the total connectedness index. We follow Buchwalter et al. (2026) in disaggregating the off-diagonal parts of the shares into co-movement within a group and contagion across groups, with the sum being one with the diagonal shares.

This is done on 200-day rolling windows, resulting in 4,354 observations of each measure from February 2009 onwards. The distance for the baseline horizon is 10 steps.

3.4 Testing the partition

There are exactly 10 ways of dividing 6 sectors into 2 sets of 3. They are all assessed based on the full-sample decomposition by two measures. The main measure is modularity (Newman and Girvan, 2004), which compares the connectedness of the groups to that which a degree-preserving null would generate. The standard statistic used to determine if a proposed grouping explains network structure is modularity, which has a built-in null, because sectors are not all equally connected overall. The secondary measure is the simple difference between within-group co-movement and cross-group contagion, a measure which is easier to read but lacks a reference distribution. The H1 says that the partition rank is 1,1 on both.

Two types of inference are required, as a score of 1 out of 10 is not sufficient to obtain a p-value < 0.10. The first is a moving-block bootstrap (Künsch, 1989) of 500 replications with a block length of 250 days, which preserves the serial dependence that makes ordinary resampling invalid. The statistic is the percentage of replications that place the Type 1 and the Type 2 in the top rank. The second is based on rolling windows. The null hypothesis is that no partition is special, and the theoretical value of the rank of the partition is that the partition is the lowest of all partitions in some 10 per cent of partitions, where the partitions do not overlap each other.

An obvious objection to a partition test of this sort is. Bank, Realty and Auto will be the most volatile sectors and any such test will tend to select the group whose members move together most. There are three diagnostics that address this: the same estimation on the raw correlation matrix, which would see if correlation alone picks the grouping; estimation on series standardised to unit variance, which would remove any scale differences; and

re-estimation on the residuals after extracting the first principal component, which would test whether the grouping survives without the common market factor.

3.5 Policy phases

A total of 21 policy phases were identified from the Reserve Bank of India (RBI) policy statements from April 2008 to September 2026 and categorized into tightening or easing/pause. Prior to estimation, three rules were set. A phase has to be a minimum of sixty days of trading. Phases are determined based on the RBI's own reasons, and not by looking at the volatility series – which would be circular. The crisis phases are marked with the dates as easing (2008-2009), rupee defence tightening (2013-2014) and easing (2020-Covid).
The first phase, which is the end of the 2008 tightening cycle, is not included in the results as it is within the first rolling window. That means that there are three non-crisis tightening phases and four easing phases in the core comparison, as will be reiterated in the limitations. One of them, the 2018 tightening, is only 41 trading days and thus does not meet the minimum of 60 days. It is kept because if it were removed, then there would only be two tightening phases, and the drop-one exercise in Section 4.3 demonstrates that the difference would change from -3.30 to -3.25.

The comparison for H2 only considers non-crisis phases, where co-movement is compared between the tightening and easing periods of Type 1. There are 2 units of analysis, and they are not synonymous. Daily rolling windows provide thousands of observations, but consecutive windows share 199 of the 200 days and are therefore not independent, and no standard error correction can fix that. Phase gives seven really independent observations and is the honest unit. Both are reported, with Newey and West (1987) standard errors at 200 lags on the daily version, but only the test at the phase level should be interpreted as inference. It is accompanied by a power calculation: with three and four phases, the design can only detect large effects. A drop-one exercise is a comparison performed using each phase separately.

3.6 Robustness

Four design decisions are tested: two window lengths (150 and 200 and 250 days), two VAR orders (1 and 2), three forecast horizons (5, 10 and 20 steps), and the use of the Garman and Klass estimator instead of the volatility measure on the subsample for 2011-2026. So that the two GARCH measures can be compared like for like, the subsample is re-estimated on the same period. These three diagnostics, described in Section 3.4, are reported. The results are given in Section 4.

4. Results

4.1 Full-sample connectedness

The connectedness matrix for the entire sample is shown in Table 2. The total connectedness index is 39.03 per cent, meaning that approximately 40 per cent of the forecast error variance of any sector is due to the other five sectors. This is in the vicinity of the 38.7 per cent reported by Guru and Das (2021) for BSE sectors and less than the 67-76 per cent reported by Mishra, Anand K and Kappagantula (2024) using a time-varying parameter VAR on returns.

Table 2. Full-sample connectedness matrix, per cent

Receiver	Bank	Realty	Auto	FMCG	Pharma	IT	FROM others
Nifty Bank	53.52	15.47	15.96	6.85	5.51	2.69	46.48
Nifty Realty	15.35	56.66	11.99	6.05	6.11	3.84	43.34
Nifty Auto	15.54	12.21	50.67	8.56	9.69	3.34	49.33
Nifty FMCG	10.51	6.65	10.51	60.33	7.81	4.19	39.67
Nifty Pharma	5.97	5.59	8.21	7.37	69.38	3.47	30.62
Nifty IT	4.83	4.95	5.34	5.11	4.51	75.26	24.74
TO others	52.20	44.88	52.02	33.94	33.62	17.53	TCI 39.03
NET	5.71	1.54	2.69	-5.73	3.01	-7.21	

Notes: Rows are receivers, columns transmitters. Rows sum to 100. VAR(2), generalised FEVD, horizon 10.

The structure is seen prior to all testing. The three most solid connections are between Bank and Auto, Bank and Realty, and Realty and Auto, each of which has a relationship above 12 per cent in both ways, and all of which are within Type 1. The highest association with a Type 2 sector is with Auto (9.69 per cent) from Pharma.

The net positions are also divided up by group. The top net transmitter is Bank with 5.71, followed by Pharma (3.01), Auto (2.69) and Realty (1.54), while the top net receiver is IT with -7.21 and FMCG with -5.73. IT has the highest proportion of variance retained (75.26 per cent). The ordering is the same as Mishra, Anand K and Kappagantula (2024). There is one exception that doesn't fit: pharma, as Section 5 discusses.

4.2 The partition test

Table 3 shows the complete set of 10 partitions of 6 sectors into two sets of 3 sectors, arranged by modularity.

Table 3. All ten partitions, ranked by modularity

Rank	Group A	Group B	Modularity	Gap	Theory
1	Bank + Realty + Auto	FMCG + Pharma + IT	-0.017	0.635	Yes
2	Bank + Realty + IT	Auto + FMCG + Pharma	-0.077	-5.937	
3	Bank + Realty + FMCG	Auto + Pharma + IT	-0.099	-7.217	
4	Bank + Auto + FMCG	Realty + Pharma + IT	-0.103	-6.898	
5	Bank + Realty + Pharma	Auto + FMCG + IT	-0.113	-8.683	

6	Bank + Auto + Pharma	Realty + FMCG + IT	-0.116	-8.472	
7	Bank + FMCG + IT	Realty + Auto + Pharma	-0.127	-9.706	
8	Bank + Auto + IT	Realty + FMCG + Pharma	-0.128	-9.936	
9	Bank + FMCG + Pharma	Realty + Auto + IT	-0.135	-10.467	
10	Bank + Pharma + IT	Realty + Auto + FMCG	-0.153	-11.380	

Notes: Modularity follows Newman and Girvan (2004). Gap is within-group co-movement minus cross-group contagion, in percentage points. Group A always contains Bank, removing mirror-image duplicates.

The Type 1 and Type 2 partition have the highest scores on both indicators. Its modularity of -0.017 is over four times closer to zero than the runner-up at -0.077, and its gap of 0.635 beats the next best by 6.572 percentage points, compared to the rest of the nine. The two measures agree on the top ranking and the bottom four, with the only difference being that the third and fourth ranking scores are reversed and the fifth and sixth ranking scores are reversed.

It is important how big the difference is! Co-movement within Type 1 is 36.9 per cent of system connectedness, and Type 1 has 20 per cent of the directed pairs (concentration ratio of 1.85), while a Type 1 sector absorbs 46.4 per cent of forecast error variance from the rest of the system as opposed to 31.7 per cent for a Type 2 sector.

One aspect of Table 3 should be explicitly noted, as it restricts the claims that can be made for the result. All the modularity values are negative. This network, under a degree-preserving null, has sectors that link groups more than within groups, regardless of the partition selected. The network is disassortative; however, it is cut. What the finding is not, however, is that the market is divided into two communities, but that, of the ten possible divisions, the one based on credit dependence is always the least disassortative.

A moving-block bootstrap was conducted because the lowest ten partitions can give an exact p-value of 0.10. This is the inferential test which yielded p = 0.036 in 96.4 per cent of 500 replications, placing the theoretical partition first. A ranked partition for the 10 per cent and 50 per cent of the 4,354 windows is also reported but is considered descriptive: the partition was ranked first in 56.1 per cent of the windows, and the median rank was one, against 10 per cent expected by chance. A binomial test of non-overlapping windows yields 1.23×10^{-7}, but that assumes that windows 200 days apart are independent, which is at odds with this paper's own argument about the grouping spanning years. It is reported for information only and should not be relied upon.

The outcome is independent of the settings. The partition is ranked first at the five-step, 10-step and 20-step horizons with gaps of 1.92, 0.63 and -1.43, respectively, and ranked first at all three horizons using the Garman and Klass measure over the 2011-2016 subsample. The sign of the gap is moved along with the specification, not the ordering.

There are three diagnostics that determine if it's mechanical or economic. The same comparison on the raw correlation matrix of log volatility yields the theoretical partition second, not first, and so the decomposition is adding information to the raw correlation matrix. If all the series are standardised to have a variance of 1, then the partition first on both measures remains unchanged, and is not an artefact of Realty and Bank being the most volatile series.

The third is not omitted and is reported. If the first principal component (accounting for 71.4 per cent of the variance) is removed, the partition is moved from first to last in both measures. Two observations suggest this is mechanical: loadings on that component are not far from each other (range 0.356 for Pharma to 0.433 for Bank), so it is not a credit factor but rather a common market factor; and residuals after stripping off a common factor are forced towards negative correlation, which turns any block structure upside down. The modularity correlation before and after ranks is – 0.515, which is consistent with inversion.

4.3 The policy-phase comparison

Table 4 compares the tightening/easing periods (excluding crisis periods).

Table 4. Tightening against easing, non-crisis phases

Measure	Tightening	Easing	Difference	t	p
Co-movement within Type 1	12.289	15.585	-3.296	-2.14	[0.032]
Type 1 share of total	28.995	37.467	-8.473	-2.68	[0.007]
Total connectedness index	43.677	42.313	1.364	0.32	[0.752]
Phase level, co-movement	12.704	15.231	-2.527	-1.45	0.244
Phase level, share	36.839	37.862	-1.023	-0.10	0.925

Notes: The first three rows use daily rolling windows, where consecutive observations share 199 of 200 days. Their p-values are shown in brackets and are descriptive only, since overlapping windows are not independent. The last two rows treat each phase mean as one observation and are the inferential test: three tightening phases against four easing.

The daily comparison is pointing in the wrong direction for H2, as co-movement is lower during the tightening, not higher. Those are not tests and refer to the sample. The data in consecutive windows are very similar, and the number of independent observations is much less than 4,354, which is too close to dependence to correct using a Newey-West correction.

Both measures are not significant on the inferential test, not because of lack of evidence against an effect, but because of power. The pooled standard deviation is 2.08, and with three tightening and four easing phases, the standard error is 1.59 on five degrees of freedom, meaning that the smallest difference that could have been detected with 80 per cent power is 5.54 percentage points. The observed value of 2.52 is 46 per cent of that level, and this effect size would require about twelve phases per group to detect.

Two additional tests reveal that the descriptive difference is frail. The drop-one exercise shows that it is reduced to nearly -3.30 to -0.53, and -2.93 to -4.83 if the exercise is excluded for any other phase. The volatility measure is important, too. Using the 2011-2026 subsample, the sign changes: the Garman and Klass measure is 15.19 compared with 14.11, and the GARCH measure is -0.33. The two agree when the period remains constant; the difference is in the period, i.e., 2010-2011. Table 5 shows the same trend for different window lengths and lag orders.

Table 5. Window and lag robustness, co-movement within Type 1

Window	VAR lag	Tightening	Easing	Difference	t	[p]	Phase p
150	1	13.26	15.00	-1.74	-1.10	[0.27]	0.89
200	1	12.80	15.41	-2.61	-1.87	[0.06]	0.26
250	1	12.99	15.59	-2.59	-2.13	[0.03]	0.15
150	2	12.80	15.05	-2.25	-1.28	[0.20]	0.94
200	2	12.29	15.59	-3.30	-2.14	[0.03]	0.24
250	2	12.59	15.78	-3.19	-2.38	[0.02]	0.15

Notes: Bracketed p-values are descriptive, as in Table 4. The phase-level test is insignificant in every specification.

All six specifications are negative, with a minimum p-value at the phase level of 0.15 anywhere in the grid. The difference also increases as the length of the window increases, as would be the case if the measure were capturing slow co-movement instead of a policy response.

Thus, the answer to H2 is not provided. This is not to say that there is no difference, but rather that this design is not capable of detecting the difference of the size observed. There are seven policy phases that are usable, and the question is still open.

4.4 Summary

The two hypotheses end differently. The credit-sensitivity partition is always the best fitting (on two measures) among the 10 partitions of the market at all horizons under standardisation and under a second volatility proxy. None of the partitions has positive modularity, which is a relative statement. Due to lack of statistical power, the policy question is left open. The pattern over time is illustrated in Figures 1-3, and it is evident that Type 1 co-movement is greater than Type 2 in almost all windows.

Figure 1. Total connectedness index, 200-day rolling window

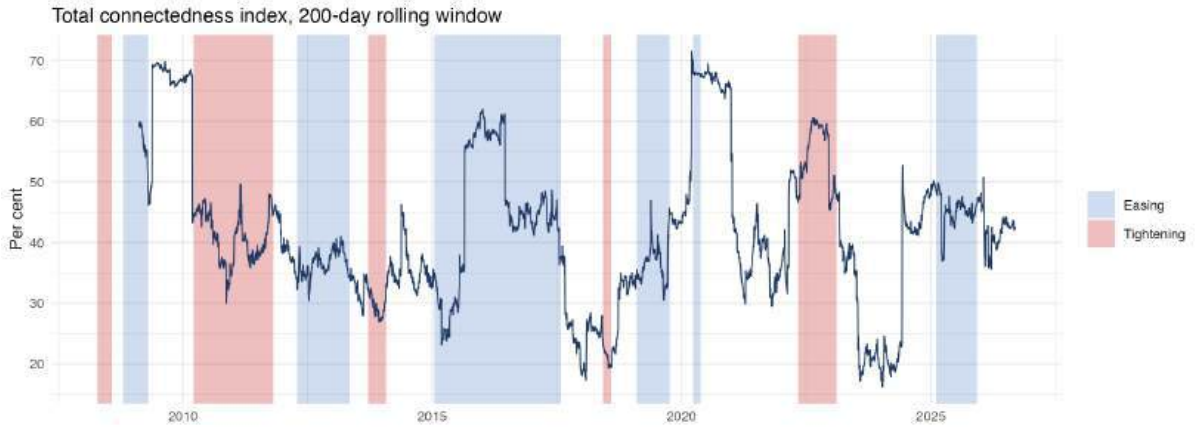


Figure 2. Co-movement within each group

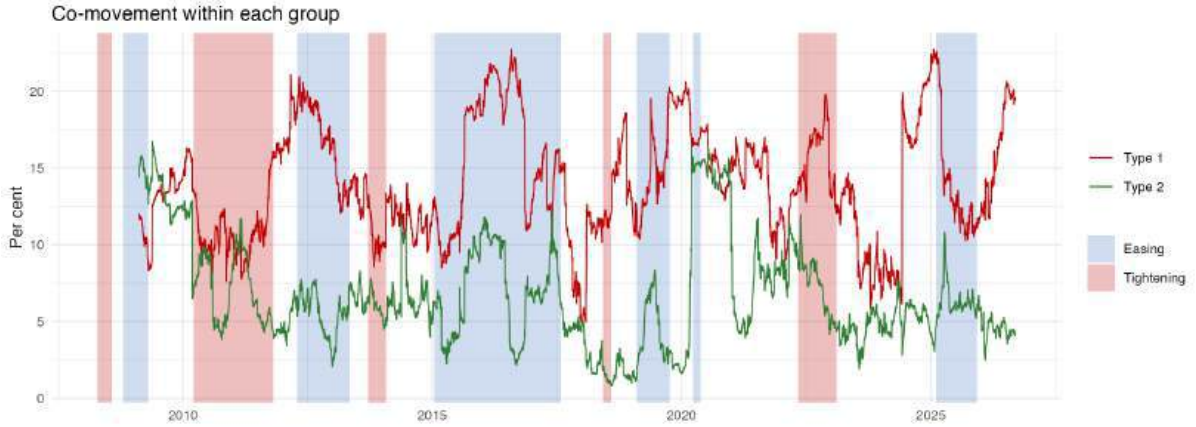
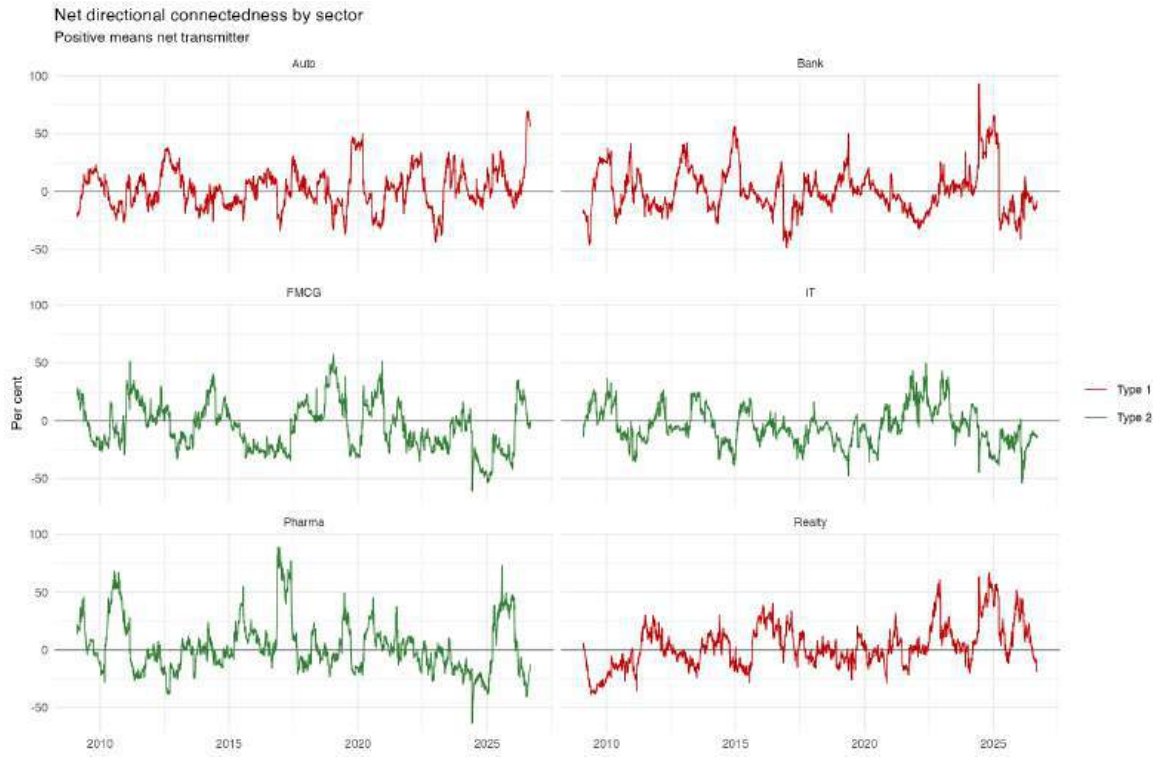


Figure 3. Net directional connectedness by sector



Notes: Shaded bands mark RBI tightening (red) and easing (blue) phases. In Figure 3, positive values indicate a net transmitter.

5. Discussion

The two outcomes are different and yet, together, they mean something more specific than might be initially thought, yet new. The volatility network fits best the one way the market can be split into two halves: the half that is credit dependent, and this one is always so, whatever the specification tested. There is no positive modularity in any division, and thus the market is not divided into two neat communities. What the evidence does support is a comparative statement: credit dependence is the most effective organising principle of those tested, but not the only one that exists.

5.2 Why the grouping holds

The weaker the partition result, the more important it is to understand what it does and doesn't indicate. It does not indicate that Bank, Realty and Auto are interest rate sensitive, as the same was already depicted by Prabu et al. (2020) and Khuntia and Hiremath (2019). Nor does it suggest that there are two communities in the market because each partition is disassortative. It demonstrates that out of all the possible ways to split the market in two, the one that is constructed on that response is the best.

The mechanism is dependence on credit. The bank loans to Prabu, Bhattacharyya and Ray are 22.7 per cent of total liabilities, while in the case of information technology it is 3.2 per cent, and housing and vehicle loans account for nearly two-thirds of personal bank loans. Banks provide that credit, developers and carmakers rely on it, and their clients borrow it to purchase from them. The volatility of these three sectors is correlated because they are hit by the same shocks and are on the same credit chain. Information technology provides foreign earnings, pharmaceuticals have inelastic products, and fast-moving consumer goods are funded by cash flow. The chain is empty.

The grouping is so stable because of this reading. Credit dependence is a characteristic of the financing of a firm, and financing structures vary over years, not quarters. If a grouping is constructed based on something that should show up in nearly all of the windows of the sample, as Figure 2 does, and the rolling-window rank test does as well. It also provides an explanation for the comparative, not absolute, nature of the effect. Credit dependence is not the largest element in the variance of these six series: Indian sectors share exposure to a large common market factor which accounts for 71.4 per cent of the variance.

5.2 Why the policy question stays open

The answer is that if the grouping is of credit, then co-movement should visibly strengthen in the case of a credit tightening. The honest first answer is that this study can't say, since it doesn't have the ability to detect a difference as small as the one observed. The smallest difference between the two is 5.54 percentage points; the observed difference is 2.52, three tightening phases compared to four easing phases. This would be a null and would be an exaggeration with 7 observations.

However, two substantive explanations can be offered, as either would predict a small effect. The first is that a channel is always 'open'. A permanently credit-dependent sector is vulnerable to credit conditions at all stages of the credit cycle. Easing is as much a credit

event as tightening, and both should bring the three sectors along in tandem, so a small differential is the outcome and not a puzzle.

The second is related to transmission. Das (2015) and Singh (2011) find that Indian policy rates do not get passed on uniformly and in a timely fashion to bank lending rates. In the event the policy rate changes, and lending rates lag behind and less completely, the glue that would hold these three sectors together is dampened before it reaches them. The lending rate asymmetry described may not be reflected in the structure of the equity network, as the equity market is two steps removed from the lending market.

The descriptive effect of the tightening cycle from 2010 to 2011, which covers the bulk of the difference, also coincided with the euro area debt crisis and the lead-up to the taper tantrum, so this period's volatility cannot solely be blamed on RBI policy.

Anghel and Caraiani (2024) checked for sign asymmetry in network connectedness surrounding FOMC meetings and found it to be weak. This study poses a related question in a different market, and also fails to come up with anything detectable, but in part because of a lack of power.

5.3 The sector which does not fit

Pharma sits in Type 2, but is a net transmitter at 3.01 per cent, and has a higher value than Realty. This can't be ignored or left unexplained.

Two points decrease the level of difficulty. The grouping is based not on the shock transmission but on the response to monetary policy: Prabu, Bhattacharyya and Ray show that pharmaceuticals are the least responsive group with a p-value of 0.63 and bank-loan dependence of 9.4 per cent, compared to auto and realty. Indian pharmaceuticals suffer from export-market regulation and currency fluctuations, both of which contribute to volatility, irrespective of credit. However, the net position of the pharmaceuticals is not steady over periods as Mishra, Anand K and Kappagantula (2024) report that the pharmaceuticals are the largest net receiver in their sample.

Because the partition test is based on within-group connectedness vs. cross-group connectedness, it is not influenced by the individual net positions. The two-group setup is a bit of a simplistic model which would be improved by more sectors.

5.4 What this adds

Three contributions follow. The first is to the Indian sectoral literature, which has plotted the network repeatedly, without trying to find out whether it is organised around anything. If a grouping is set up and compared to all other groupings, it becomes a test that can be used by other studies as a design with other characteristics.

Modestly, the second is the connectedness framework. The primary risk is cluster misclassification, as identified by Buchwalter, Diebold and Yilmaz (2026), and to detect groups from data can be referred to Gudmundsson and Brownlees (2021) and Brownlees, Gudmundsson and Lugosi (2022). Exhaustive enumeration is not a detection method and does not compete with these detection methods. It is an application when a grouping is already determined by theory and the system is small: 6 sectors provide 10 partitions, 8 sectors provide 35 partitions, 10 sectors provide 126 partitions, and it becomes impractical pretty quickly.

The third is negative. Indian sectoral connectedness studies take place under contingent shocks of crises, elections and the pandemic, and the transmission literature ends with the bank lending rate. This study relates the two and determines that seven policy phases don't answer the question. Proving it is underpowered is a smaller contribution than answering the question, but it does tell the next study what it needs: about 12 phases in each group or a continuous measure of policy surprise.

The practical implication can be specified. The return correlation within Type 1 is on average 0.678, while the return correlation between groups is on average 0.460, and within Type 2 is on average 0.414. The sample volatility of an equal-weighted portfolio of Bank, Realty and Auto is 25.7 per cent annually, that of a mixed portfolio of Bank, Pharma and IT is 18.2 per cent annually, and that of the least volatile combination of any three sectors – FMCG, Pharma and IT – is 15.8 per cent annually. Thus it is approximately 42 per cent more volatile to hold the three credit-sensitive sectors, compared to having one of them plus two from the other, for no apparent benefit. The connectedness estimates say the same thing in another way: a Type 1 sector explains 46.4 per cent of the forecast error variance as compared to 31.7 per cent for a Type 2 sector, and the co-movement within Type 1 accounts for 36.9 per cent of all system connectedness (a concentration ratio of 1.85), while only 20 per cent of the directed pairs are within Type 1. An investor with these three names is not as diversified as the labels imply, and whether this helps or makes it more comfortable when the policy rate is reduced is something that this study cannot answer.

6. Conclusion

This paper questioned if there is an organisational structure of volatility transmission between equity sectors in India and if it is sensitive to the monetary policy cycle. There is an answer to the first question. The second doesn't, due to statistical power.

On the first question, the evidence is consistent. The split that Prabu, Bhattacharyya and Ray (2020) have identified from the ten possible ways of dividing the six NSE sectoral indices into two groups of three is the top-ranked on modularity and the top-ranked on co-movement gap, with a difference of 6.572 percentage points from the next best split and the entire spread of the remaining nine splits. For example, 96.4 per cent of replications had it in first position and, in 4,354 rolling windows, it appeared in first position 56.1 per cent of the time compared to 10 per cent expected by chance, with a median rank of one. At three forecast horizons, standardisation and under a second volatility measure. The claim it supports is comparative, in the sense that every partition has a negative modularity value – that is, the market is not split into two clean communities – and that the credit dependence is the best of the partitions, not necessarily a complete description of the network.

On the second question, the study cannot give an answer. There are three tightening and four easing phases; the smallest difference is 5.54 percentage points, while the difference observed is 2.52. That gap is also precarious, having been added almost exclusively in the 2010-11 period and turning negative as per Garman and Klass. The design is not powerful enough; the effect is non-existent.

The grouping that fits is structural. The banks, the property developers, and the car companies are one credit chain, and financing structures are not quarterly but yearly.

Whatever the policy rate is doing, their volatility is more aligned with each other than any other group of these six sectors.

6.1 Limitations

There are six caveats which need to be clearly stated.

One of these is the number of samples at the phase level. The three non-crisis tightening periods are included within the estimation window, while the four easing periods are not, and there is not a lot of weight in the seven observations in the test.

The second one is the number of sectors. The margin is not defined, but the exact rank test is capped at p = 0.10 for ten partitions, so bootstrap was used. Financial Services and Media were not available in the data extract, and three sectors have been used to represent each of these groups, instead of the original four and five sectors.

The third one is stationarity. Dickey-Fuller and Phillips-Perron reject a unit root in all six log volatility series, while KPSS rejects level stationarity in all six. The disagreement is the typical hallmark of long memory, and levels of the VAR estimate this.

The fourth is identification. In Buchwalter, Diebold and Yilmaz (2026), the within-group measure reflects co-movement and does not distinguish contagion in the single-cluster case. They would be identified by their clustered identification.

The fifth refers to the common factor. The first principal component accounts for 71.4 per cent of the variance, and the partition is moved from first to last. Nearly uniform loadings and negative correlation constraint in the residuals from the common factor removal suggest that the reversal is most likely mechanical, but the partition result is contingent on the common factor being retained.

The sixth is attribution. The core comparison was limited to the time period between 2010 and 2011, a period during which the RBI was tightening, while also the euro area crisis and the lead-up to the taper tantrum were flagged and hence not part of the core comparison.

6.2 Future research

This work is followed by six future directions.

The most immediate is to widen the system. With eight sectors, the 35 partitions, and ten sectors, the 126 partitions, are sufficient for the rank test to be conventionally significant, independently of the others. Incorporating Metal, Energy, Media and Financial Services would also allow for more granular categorizations and see if the credit chain reading holds up with a larger swath of the market in play.

The second one is to use the clustered identification method of Buchwalter, Diebold and Yilmaz (2026), which would allow for separating co-movement from contagion within each cluster and for the study to address causality instead of correlation.

The third is to try out other combinations of the same design. There are other untested factors that are observable, such as export exposures, and state ownership and foreign institutional holdings are clean candidates because of the paired public and private banking indices in India.

The fourth is frequency decomposition. Baruník and Křehlík (2018) distinguish short-run from long-run connectedness, while a structural grouping should show up more strongly at low frequencies, as per the reading provided here.

The fifth is to re-examine H2 with more tightening episodes or continuous policy surprises, similar to Prabu, Bhattacharyya and Ray (2020), using the proxy of interest rate swaps, which cannot be done in the phase classification.

The sixth one is economic value. Mishra, Anand K and Kappagantula (2024) construct minimum connectedness portfolios based on their estimates, and the same can be used to examine whether excluding the Type 1 cluster brings better performance or not.

Works Cited

Akimov, A., Lee, C. L., & Stevenson, S. (2019). Interest rate sensitivity in European public real estate markets. Journal of Real Estate Portfolio Management, 25(2), 138–150. https://doi.org/10.1080/10835547.2020.1803694

Anghel, D. G., & Caraiani, P. (2024). Monetary policy shocks and the high-frequency network connectedness of stock markets. International Review of Economics and Finance, 96, 103558.

Antonakakis, N., Chatziantoniou, I., & Gabauer, D. (2020). Refined measures of dynamic connectedness based on time-varying parameter vector autoregressions. Journal of Risk and Financial Management, 13(4), 84. https://doi.org/10.3390/jrfm13040084

Baruník, J., & Křehlík, T. (2018). Measuring the frequency dynamics of financial connectedness and systemic risk. Journal of Financial Econometrics, 16(2), 271–296. https://doi.org/10.1093/jjfinec/nby001

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327.

Brownlees, C., Gudmundsson, G. S., & Lugosi, G. (2022). Community detection in partial correlation network models. Journal of Business and Economic Statistics, 40(1), 216–226.

Buchwalter, B., Diebold, F. X., & Yilmaz, K. (2026). Clustered network connectedness: A new measurement framework with application to global equity markets. Journal of Econometrics. https://doi.org/10.1016/j.jeconom.2026.106243

Chatziantoniou, I., Gabauer, D., & Marfatia, H. A. (2022). Dynamic connectedness and spillovers across sectors: Evidence from the Indian stock market. Scottish Journal of Political Economy, 69(3), 283–300. https://doi.org/10.1111/sjpe.12291

Chauhan, S., Suri, P., Twala, B., Priyadarshi, N., & Ali, M. (2025). Exploring the relationship between macroeconomic indicators and sectoral indices of Indian stock market. F1000Research, 14, 180. https://doi.org/10.12688/f1000research.160668.2

Chowdhury, S. S. H., & Irfan, M. (2022). A study on the time-varying volatility connectedness between the sectors in the Indian stock market. Montenegrin Journal of Economics, 18(3), 107–118.

Das, S. (2015). Monetary policy in India: Transmission to bank interest rates (IMF Working Paper No. 15/129). International Monetary Fund. https://doi.org/10.5089/9781513598796.001

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366), 427–431.

Diebold, F. X., & Yilmaz, K. (2009). Measuring financial asset return and volatility spillovers, with application to global equity markets. The Economic Journal, 119(534), 158–171. https://doi.org/10.1111/j.1468-0297.2008.02208.x

Diebold, F. X., & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. International Journal of Forecasting, 28(1), 57–66. https://doi.org/10.1016/j.ijforecast.2011.02.006

Diebold, F. X., & Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. Journal of Econometrics, 182(1), 119–134. https://doi.org/10.1016/j.jeconom.2014.04.012

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987–1007.

Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. Journal of Business, 53(1), 67–78.

Ghosh, R., Latha, K., & Gupta, S. (2018). Interest rate sensitivity of non-banking financial sector in India. Vikalpa: The Journal for Decision Makers, 43(3), 152–170. https://doi.org/10.1177/0256090918792803

Gudmundsson, G. S., & Brownlees, C. (2021). Detecting groups in large vector autoregressions. Journal of Econometrics, 225(1), 2–26.

Guru, B. K., & Das, A. (2021). COVID-19 and uncertainty spillovers in Indian stock market. MethodsX, 8, 101199. https://doi.org/10.1016/j.mex.2020.101199

Hou, Y., Li, W., Wu, D., Zang, Y., & Quach, L. (2025). The spillover effect of US monetary policy on the international financial market: Evidence from network analysis. Journal of Management Science and Engineering, 10, 111–125.

Inoue, A., & Rossi, B. (2019). The effects of conventional and unconventional monetary policy on exchange rates. Journal of International Economics, 118, 419–447. https://doi.org/10.1016/j.jinteco.2019.01.015

Kamboj, H., Rao, N. V. M., & Agrawal, M. (2026). Central bank communication, policy action, and spillover: Evidence from Indian financial market. International Review of Economics and Finance, 111, 105649.

Khuntia, S., & Hiremath, G. S. (2019). Monetary policy announcements and stock returns: Some further evidence from India. Journal of Quantitative Economics, 17, 801–827. https://doi.org/10.1007/s40953-019-00158-y

Koop, G., Pesaran, M. H., & Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. Journal of Econometrics, 74(1), 119–147.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. The Annals of Statistics, 17(3), 1217–1241.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, 54(1–3), 159–178.

Maurya, P. K., Bansal, R., & Mishra, A. K. (2025). Sectoral connectedness, volatility spillover and hedging opportunities: Evidence from Indian stock market. Eurasian Economic Review, 15, 1099–1134. https://doi.org/10.1007/s40822-025-00324-z

Mishra, A. K., Anand K, K., & Kappagantula, A. V. (2024). Unveiling asymmetric return spillovers with portfolio implications among Indian stock sectors during Covid-19 pandemic. North American Journal of Economics and Finance, 75, 102297. https://doi.org/10.1016/j.najef.2024.102297

Mishra, P., Montiel, P., & Sengupta, R. (2016). Monetary transmission in developing countries: Evidence from India (IMF Working Paper No. 16/167). International Monetary Fund.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3), 703–708.

Newman, M. E. J. (2002). Assortative mixing in networks. Physical Review Letters, 89(20), 208701. https://doi.org/10.1103/PhysRevLett.89.208701

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Nyakurukwa, K., & Seetharam, Y. (2023). Cross-country categorical economic policy uncertainty spillovers: Evidence from a conditional connectedness TVP-VAR framework. Journal of Financial Economic Policy, 15(2), 164–183. https://doi.org/10.1108/JFEP-10-2022-0256

Panda, P., Anjaly, B., Panda, B., & Mohanty, R. K. (2026). Unconventional and conventional monetary policy spillovers to advanced and emerging stock markets. Journal of Emerging Market Finance, 25(3), 313–337. https://doi.org/10.1177/09726527251408565

Pandit, B. L., Mittal, A., Roy, M., & Ghosh, S. (2006). Transmission of monetary policy and the bank lending channel: Analysis and evidence for India (Development Research Group Study No. 25). Reserve Bank of India.

Pesaran, M. H., & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. Economics Letters, 58(1), 17–29.

Phillips, P. C. B., & Perron, P. (1988). Testing for a unit root in time series regression. Biometrika, 75(2), 335–346.

Prabu, A. E., Bhattacharyya, I., & Ray, P. (2020). Impact of monetary policy on the Indian stock market: Does the devil lie in the detail? Indian Economic Review, 55, 27–50. https://doi.org/10.1007/s41775-020-00078-2

Sahoo, S., & Kumar, S. (2023). Volatility spillover among the sectoral indices of the Indian capital market: Evidence from the COVID period. Indian Journal of Finance, 17(9).

Singh, B. (2011). How asymmetric is the monetary policy transmission to financial markets in India? RBI Occasional Papers, 32(2). Reserve Bank of India.

Tomar, K. S., & Kesharwani, S. (2022). Asymmetric effect of monetary policy on Indian stock market sectors: Do monetary policy stimulus transpire the same effect on all sectors? Cogent Economics & Finance, 10(1), 1999058. https://doi.org/10.1080/23322039.2021.1999058

Uddin, S. M. M., Qamar, F., & Ahmad, A. (2022). Impact of interest and exchange rates on stock returns and volatility: Evidence from Indian private banks. BHU Management Review, 10(2), 45–58.

Appendix A. Reserve Bank of India policy phases

Phases are dated from published RBI policy statements and classified by the Bank's own stated rationale. Crisis phases are flagged by date, not by inspecting the volatility series. The first phase falls entirely inside the first rolling window and so produces no estimates.

Table A1. Policy phases, 15 April 2008 to 15 September 2026

Phase	Start	End	Direction	Crisis	Repo move	Windows
T1 2008	2008-04-15	2008-07-30	Tightening		to 9.00%	0
E1 GFC	2008-10-20	2009-04-21	Easing	Yes	9.00 to 4.75%	42
P1	2009-04-22	2010-03-18	Pause		4.75%	222
T2 2010-11	2010-03-19	2011-10-25	Tightening		4.75 to 8.50%	401
P2	2011-10-26	2012-04-16	Pause		8.50%	117
E2 2012-13	2012-04-17	2013-05-03	Easing		8.50 to 7.25%	262
P3	2013-05-04	2013-09-19	Pause		7.25%	97
T3 2013-14	2013-09-20	2014-01-28	Tightening	Yes	7.25 to 8.00%	89
P4	2014-01-29	2015-01-14	Pause		8.00%	234
E3 2015-17	2015-01-15	2017-08-02	Easing		8.00 to 6.00%	631
P5	2017-08-03	2018-06-05	Pause		6.00%	208
T4 2018	2018-06-06	2018-08-01	Tightening		6.00 to 6.50%	41
P6	2018-08-02	2019-02-06	Pause		6.50%	126
E4 2019	2019-02-07	2019-10-04	Easing		6.50 to 5.15%	160

P7	2019-10-05	2020-03-26	Pause		5.15%	119
E5 COVID	2020-03-27	2020-05-22	Easing	Yes	5.15 to 4.00%	36
P8	2020-05-23	2022-05-03	Pause		4.00%	484
T5 2022-23	2022-05-04	2023-02-08	Tightening		4.00 to 6.50%	194
P9	2023-02-09	2025-02-06	Pause		6.50%	496
E6 2025	2025-02-07	2025-12-05	Easing		6.50 to 5.25%	204
P10	2025-12-06	2026-09-15	Pause		5.25%	191

Notes: Windows count the 200-day rolling estimates falling inside each phase, for a total of 4,354. The core comparison for H2 uses the three non-crisis tightening phases (T2, T4, T5) and the four non-crisis easing phases (E2, E3, E4, E6). Repo rates are the policy repo rate under the liquidity adjustment facility.

MUSICAL SYMMETRY AND PARALLEL MODULATION VIA THE PLR GROUP

JITONG LEI

Abstract. When attempting to modulate from a major to a minor key, music can often sound dissonant. However, by taking advantage of the deep fundamental symmetries found within music, it is often possible to resolve such dissonance. One of the major symmetries underlying Western classical music can be endowed with a group structure. Using the "Happy Birthday" song as an example, we will show how one such group, known as the PLR group, can be used to go from parallel major to minor to minimize such dissonance. This strategy is a useful framework to help classical composers and modern musicians smoothly transition between distant keys using the PLR group.

1. Introduction

Mathematics is often said to connect to everything around us, whether that is finding the total cost of groceries, coding for computers, or medicine. Though it is not the first connection for many people to recognize, math is deeply embedded in music. For instance, classical western music theory describes musical sounds in terms of pitches, grouped into 7-note whole tone or 12-note chromatic scales. Any two pitches that are 12 semitones (i.e. an octave) apart represent the same note. 7-note scales are primarily divided into two types: major and minor. Both types are made with the notes being either a whole tone or semitone apart.

Music is full symmetries. One of the first prominent appearances of symmetry in music consists of musical chords. Chords are several musical notes played simultaneously. The most common type of chord is a triad, which contains three notes. Chords, and therefore also mathematical symmetries, permeate all through music. Another major symmetry reflects how every major key has a corresponding parallel minor key with the same root note. To take the simplest example, for the C major key, the corresponding parallel minor is the C natural minor key.

There is an abundance of other symmetries found in music theory and group theory. Group theory is a branch of mathematics that helps explore the deep connections between the various symmetries within music. However, group theory does not only connect to music theory. In fact, it is everywhere in the natural world. For instance, the hours of a clock form a group under addition. This means that, starting from midnight, every twelve hours form a cycle, after which the clock reaches twelve again. Another such example is the collection of groups of symmetries of regular polygons, known as the dihedral groups. A dihedral group preserves the shape and symmetries of a polygon

and consists of reflections and rotations. With so many symmetries found that connect to the natural world, it is inevitable for researchers and musicians to eventually discover more new symmetries within music.

Two centuries ago, the musical notions of relative and parallel major and minor have led to the development of a particular symmetry group known as the PLR group, or formally, the neo-Riemannian group. Its conception goes back to the late 19th century, when the German musicologist and composer Hugo Riemann developed his notion of "Harmonic Dualism." It is the idea that minor triads are the inverses of major triads. Another way to interpret this is to see minor chords as "mirrored images" of their major counterparts. Riemann was originally inspired by past theorists such as Moritz Hauptmann and Arthur von Oettingen. The two originally saw the minor triad as an altered form of the major triad and not as its own independent form. Riemann wanted to change this conception [1]. He wanted to prove that a minor triad is just as legitimate as a major triad [2].

In harmonic dualism, the specific relationships between pitches are derived from concepts known as "Schritt", "Wechsel" and "Leittonwechsel" [3]. These terms refer to common, parallelism, and leading tones, respectively. In the late 20th century, David Lewin built upon Riemann's theory by mapping the musical relationships to mathematical functions, through a process now known as transformational theory [4]. Later, Richard Cohn and others took Riemann's three terms and relabeled them as P, L, and R. This allowed them to formalize the geometric configuration of the PLR group within the Tonnetz. The Tonnetz, which translates to "the table of tonal relations," shows pitches as vertices of a regular polygon inscribed in a circle. Connecting triplets of vertices along the circle then geometrically represents the corresponding chords [5]. Cohn showed how P, L, and R relate to the group of symmetries of this polygon. Through many music theorists and mathematicians, this eventually led to the notion of the PLR group as we know it today.

1.1. **Motivation.** A major question for musicians and composers is that of composing parallel modulations. For instance, modulating from C major to C minor does not always guarantee that the transition will sound smooth. Instead, it can often sound misplaced and dissonant. In this paper, we will take the famous 'Happy Birthday' melody as a proof of concept for our approach. If we attempt to instantly modulate the piece from its original key (C major) to the corresponding parallel minor, it sounds unpleasant. To ensure the bridge within the modulation sounds musically sophisticated and polished, we will explore an identity from the PLR group that will enable us to mitigate this issue.

2. Prerequisite Information

Below, we first introduce the necessary knowledge relevant to group theory, which will be essential to the remainder of the paper. We refer to [6] for the basic group theoretic definitions.

Definition 2.1. A *group* is a set G with a binary operation $*$ satisfying the following axioms:

(1) Closure:

$$\forall\, g_1, g_2 \in G,\ g_1 * g_2 \in G.$$

(2) Associativity:

$$\forall a, b, c \in G, (a * b) * c = a * (b * c).$$

(3) Identity:

$$\exists\, e \in G \text{ such that } \forall\, g \in G,\ e * g = g * e = g.$$

(4) Inverses:

$$\forall\, g \in G,\ \exists\, h \in G \text{ such that } g * h = h * g = e, \text{ where } h = g^{-1}.$$

We call h the *inverse* of g.

Definition 2.2. A *subgroup* H of a group G is a subset of G that is itself a group under the operation of G.

Example 2.3. Consider the set $G = \mathbb{Z}/8\mathbb{Z}$. This set is given by:

$$\{0, 1, 2, 3, 4, 5, 6, 7\}.$$

It is well-known that G is a group with the group operation being addition modulo 8. If we now take the subset

$$H = \{0, 2, 4, 6\},$$

we may in fact show that H is actually a subgroup of G. We need to test the subset for the four subgroup properties following the four axioms specified above:

(1) Closure:

To prove closure, we need to check that the sum of any two elements in the subset modulo 8 stays within the subset.

$$2 + 2 \equiv 4 \in H$$

$$2 + 4 \equiv 6 \in H$$

$$2 + 6 \equiv 8 \equiv 0 \in H$$

$$4 + 6 \equiv 10 \equiv 2 \in H$$

(2) Associativity:
Addition is associative, and hence the group operation of H is associative as well.

(3) Identity:
To prove the existence of the identity element, we must find an element $e \in H$ such that $e + g = g + e = g$ for all $g \in H$. We claim that 0 is the identity:

$$\forall\, g \in \mathbb{Z}/8\mathbb{Z},\ 0 + g \equiv g + 0 \equiv g \pmod 8.$$

(4) Inverses:
According to the definition of an inverse, for each element $g \in H$, we want to find another element $h \in H$ such that $g + h = h + g = 0$.

$$0 + 0 \equiv 0 \pmod 8$$
$$2 + 6 \equiv 8 \equiv 0 \pmod 8$$
$$4 + 4 \equiv 8 \equiv 0 \pmod 8$$
$$6 + 2 \equiv 8 \equiv 0 \pmod 8$$

From the equations, we can see that the inverse of 0 is 0, the inverse of 2 is 6, the inverse of 4 is 4, and the inverse of 6 is 2.

As we have checked all four axioms, we may conclude that H is indeed a subgroup of G.

Definition 2.4. *Order of a Group*
The order of a group G gives the number of elements in G.

- Example:
  Consider the group

$$G = \{0, 1, 2, 3, 4, 5\} \quad \text{under addition mod 6.}$$

  Since there are 6 elements in the group, the order of the group is simply 6.

Definition 2.5. *Order of an Element*

- The order of an element $a \in G$ is the smallest positive integer n such that $a^n = e$.

Definition 2.6. *Abelian Group*

- A group G has the property $ab = ba$ for every pair of elements $a, b \in G$.

Definition 2.7. *Cyclic Groups*
We say that a group G is *cyclic* if it takes the form

$$G = \{a^n \mid n \in \mathbb{Z}\},$$

where the element a is called a *generator* of G.

Definition 2.8. *Dihedral Groups*
Also known as the dihedral group of order $2n$, such a group is denoted as D_{2n}. It is the group of symmetries of the vertices of a regular n-gon ($n \geq 3$).

- **Example:** In an octagon, there are 8 rotations and 8 reflections. Thus, the order of D_{16} is

$$16 = 2 \cdot 8.$$

2.1. **Music Theory.** We now also introduce the relevant musical theoretic definitions.

Definition 2.9. *Major and Minor Scales*
The major and minor scales are 7-note scales that are made up of notes being either a tone or semitone away from each other. The major scale has a pattern of Whole-Whole-Half-Whole-Whole-Whole-Half between each note. The minor scale has a pattern of Whole-Half-Whole-Whole-Half-Whole-Whole. The major third scale degree is where the two scales differ. In a minor scale, there is a minor third interval above the root note, while in a major scale, there is a major third interval above the root note.

Definition 2.10. *Triads*

A chord is a combination of three notes or more, built off of a single note, the root. A chord is known as a *triad* when it consists of exactly three notes. A major triad consists of a major third and a perfect fifth from the root note [7]. Triads can be categorized as major, minor, augmented and diminished [8].

Example 2.11. In the C major triad, built from the notes of the C major key, the "C" is the root note, with an "E" and a "G" being a major third and perfect fifth above, respectively. In C major, which does not include any sharps or flats, accidentals do not have to be accounted for [7]. Alternatively, an example of a minor chord, built from the notes of the key C minor, can be the C minor triad. Starting from the root note C, three semitones are counted up, landing on the note $E\flat$. Then, from the root note again, seven semitones are counted up to land on the final note, G.

Definition 2.12. *Symmetry in Music*
In Western music, a chromatic scale is a 12-note ascending scale which increases by a semitone with each step [7]. The distances between the notes in the twelve-note scale, as measured in semitones, turn out to be equivalent to the addition in the group $\mathbb{Z}/12\mathbb{Z}$. For example, in the group $\mathbb{Z}/12\mathbb{Z}$, starting from 0 to 11, each number represents a note in the twelve-note chromatic scale. If we start from C, which is equivalent to 0, going up by two semitones would equal to D. In the group $\mathbb{Z}/12\mathbb{Z}$, D is equivalent to the element 2. Thus, the sentence

"going up from C by two semitones leads you to D"

is equivalent to the equation

$$0 + 2 = 2.$$

One soon realizes that music is full of symmetries. For instance, one may transpose a piece by shifting all of its notes in the same direction by the same number of semitones. In group theory, this can be demonstrated through translational symmetry in the group $\mathbb{Z}/12\mathbb{Z}$, mathematically expressed as:

$$T_k(x) = (x + k) \bmod 12,$$

where x represents the pitch, while k represents the shift (measured in semitones). For example, the C major triad, consisting of the notes $C - E - G$, yields the E major triad, $E - G\# - B$, when transposed up by four semitones.

From the twelve chromatic notes, one can also build major and minor triads. There are a total of twelve major and twelve minor triads. Adding them together results in 24 total triads. These triads and the relationships between them can be mathematically represented using a regular dodecagon and its associated group of symmetries, the dihedral group D_{24}.

2.2. **PLR Group.** We now define a fundamental concept of our report, which is the PLR group.

Definition 2.13. The *PLR group* is a group generated by three elements P, L, and R. Each letter of the PLR group represents an action on the 24 triads. The letter P represents a parallel modulation. P has order 2 and sends a parallel major chord to its minor and vice versa.

Example: C major $\to$ C minor or $E\flat$ minor $\to$ $E\flat$ major.

The letter L represents the exchange of leading tones. L has order 2 and exchanges the leading tone by bringing the root one semitone down going from major to minor.

Example: C major $\to$ E minor.

However, if the chord is minor, the fifth is raised by one semitone up going from minor to major.

Example: A minor $\to$ F major.

The letter R stands for the word relative. R has order 2 and sends a major/minor chord to its relative minor/major by either bringing the fifth up by two semitones (major to minor) or bringing the root down by two semitones (minor to major).

Example: C major $\to$ A minor or E minor $\to$ G major.

It turns out that there is a direct connection between the PLR group and the 24 major and minor chords.

Proposition 2.14. *The PLR group is isomorphic to the dihedral group D_{24}.*

Proof. For a proof of a Proposition 2.15, see Theorem 5.1 in [9].

□

Lemma 2.15. *One has the identity $P = R(LR)^3$.*

Proof. See Figure 1 for the verification of the identity using a sequence of modulations from C major to C minor. □

We now show how we can apply these concepts to our chosen piece.

3. Results

The piece 'Happy Birthday,' arranged for the piano, consists of a melody in the treble clef and a harmony consisting of a sequence of chords in the bass clef. Our goal is to apply the identity $P = R(LR)^3$, proven in the previous section, in order to construct a consonant sequence of stepwise modulations.

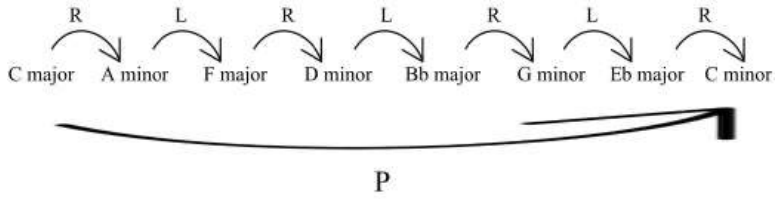


Figure 1. From C major $P = R(LR)^3$ applied to change the keys

In Figure 1, starting on C major, R is applied to change C major to its relative minor, which is A minor. Then, L is applied to A minor to change A minor to its leading tone relative, F major. R and L is continuously applied in the pattern of *RLRLRLR,* ultimately arriving at the C minor key. There is a large arrow in Figure 1 going from C major to C minor, used to indicate that the pattern *RLRLRLR* is equivalent to a single parallel modulation P from C major to C minor. In other words, we can replace a single parallel modulation with a sequence of interwoven relative and leading tone modulations.

A well-known technical tool that can be used to analyze the structure of musical compositions is known as Roman numeral analysis. Prior to a Roman numeral analysis, we have to understand the relationships between major, minor, and diminished chords within a single scale. The uppercase Roman numerals indicate major chords, while the lowercase Roman numerals reflect minor chords. For a major key, the corresponding pattern of chords within the scale is I, ii, iii, IV, V, vi, vii° (diminished). For keys that

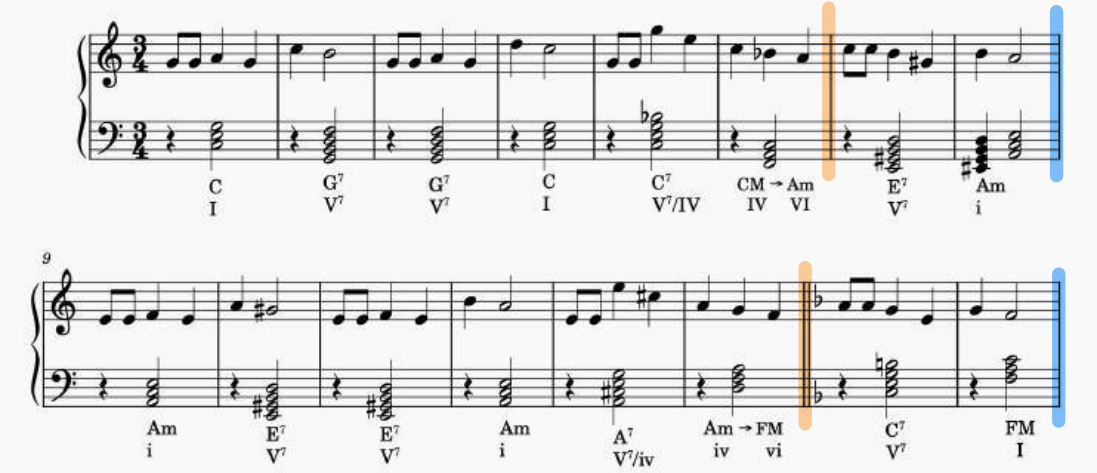



FIGURE 2. Roman Numeral Analysis for Happy Birthday in Major and Minor

are minor, the sequence is i, ii, III, iv, v, VI, VII instead. Now that we have defined the Roman numerals, we will use them to analyze the 'Happy Birthday' song. The first two transpositions of 'Happy Birthday' into the C major and A minor keys are shown in Figure 2, along with the corresponding Roman numeral analyses. The fact that the Roman numerals are the same in both the major and minor keys highlights that while the key is being changed, the underlying melodic and harmonic structure is preserved. We use this strategy to determine which chords are to be played in each major and minor key as we use the $P = R(LR)^3$ identity to modulate throughout the piece. The orange lines in Figure 2 explicitly show the locations of the key changes, while the blue lines indicate that the eight-bar melody has ended and needs to restart in the new key. The first four bars in each key consist of a $I - V - V - I$. The goal is to establish the melody in the new key right after the preceding modulation. At the same time, the end of each melody consists of a $V - I$ resolution. An example is shown in Figure 2, where bars seven and eight indicate a strong resolution to the new key. This indicates the end of a melody and allows us to transition to the melody in the new, modulated key.

Figure 3 depicts the sheet music for the entire piece, in which the identity $P = R(LR)^3$ is successively applied to the 'Happy Birthday' piece. The chosen sequence of transitions sounds consonant and smooth. To ensure that the modulations are not too sudden, two bars prior to each new melody, the chord progression pivots to establish the new key. Pivoting is a technique used to create smooth musical transitions by finding a chord that is both in the old and new keys, known as a pivot chord.

We also add some final changes after initially composing the given transition. At the end of each melody, i.e. every eight bars, instead of a flat 7th chord, the seventh note in a major scale is raised by a half step to obtain a dominant 7th chord. We change this to create a stronger resolution at the end of each section and produce more tension in the piece. This allows for a distinct start and finish and clearly separates the modulated

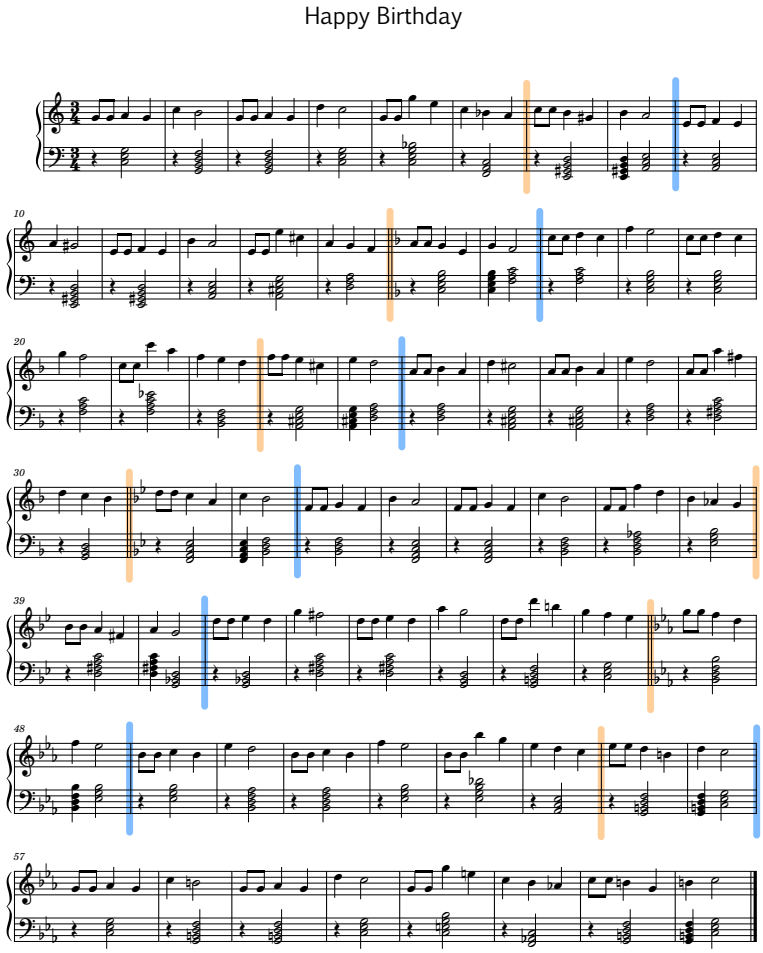


FIGURE 3. Happy Birthday after applying $P = R(LR)^3$

keys. We also shift the melody up or down by an octave where necessary in order to simplify the sheet music notation and to stay within the range of the original melody. These minor changes help us create more structure and cohesion within the piece.

4. Discussion

In the real world, both classical musicians and pop artists can use this structure to write their own music. The underlying framework helps create smooth modulations from a major key to its parallel minor without it ever sounding too sharp or sudden.

The structure we propose has multiple benefits and drawbacks. One con of the strategy is the fact that it uses up to seven transitions before it reaches the parallel minor. This configuration isn't recommended for pop artists because if used, a song might get too long. In today's society, pop songs are usually between two to four minutes long in order to keep the listeners engaged. Using the above series of modulations would

likely make the song longer than four minutes. Repeating the melody several times could also run the risk of listeners getting bored of the music.

On the other hand, we highly recommend this structure for classical composers and musicians. In long and complex concertos (such as Brahms's Piano Concerto No. 2, which is nearly an hour long) composers can take advantage of the underlying mathematical structure within our proposed sequence of transitions. It wouldn't harm a piece if an extra five minutes were added to something that is already fifty minutes long. Moreover, classical music is always described as a story. Through the use of multiple melodies and key shifts, a consecutive sequence of modulations can better portray the story that is trying to be told. Hence, adding $P = R(LR)^3$ would be at much lower risk of making the music sound repetitive, as it would only enhance the story trying to be communicated. This configuration thus provides another approach which allows composers to produce music that isn't harsh to the ear, while also successfully changing between two distant keys.

4.1. **Recommendations for Future Research.** Throughout this paper, only the modulation from major to minor was explored. If more time was allowed, the transition from minor to major can also be explored. In particular, instead of going from C major to C minor, the song can start from an arbitrary minor key and end on its parallel major key. Additionally, there are other ways that music theory and group theory interlink - for instance, other variants of the PLR group exist, such as the TI group and the SP group [9]. Instead of using $P = R(LR)^3$ as the basis for our modulation sequence, one might also attempt to explore the elements and identities within other groups with similar underlying mathematical structure.

5. Conclusion

In this paper, we introduce the fundamental group theory knowledge necessary to define and identify key properties of the PLR group. Using this, we address the key problem of dissonance present within modulations between parallel major and minor keys. From there, using the simple tune of 'Happy Birthday,' the identity $P = R(LR)^3$ is utilized to develop a better way to modulate between such major and minor keys. In the end, this enables us to smoothly modulate in 'Happy Birthday' from C major to C minor.

The understanding of group theory allows us to recognize the symmetry behind $P = R(LR)^3$. The consecutive actions of L and R generate the series of modulations that is equivalent to a single parallel modulation from C major to C minor. In a greater perspective, group theory allows us to understand the underlying relationship between the tonal symmetries within our piece.

More broadly, through the lens of music, music can also help us better understand various mathematical concepts. Though this relationship may not be immediately

apparent, there are a lot of similarities between math and music. For instance, musical consonance and dissonance are properties that turn out to be intimately connected to mathematical ratios. The minor third, major third, perfect fourth, perfect fifth, minor sixth, major sixth, and the octave turn out to be the consonant intervals [10]. These are the intervals that we perceive as sounding "good" or "smooth". On the other hand, the remaining intervals, that is, the minor second, the major second, the minor seventh, the major seventh, and the tritone are classified as dissonant. These are the intervals that sound "unpleasant".

All musical pitches can be represented using the corresponding frequencies of air pressure waves, as measured in the number of vibrations per second [11]. This unit of measurement is known as a Hertz (Hz). When notes are played simultaneously, their sinusoidal waveforms are combined into a new, more complex waveform. For instance, when an interval is in unison, the corresponding ratio of sinusoidal frequencies is $1:1$. An octave is represented through the ratio of $2:1$, while when an interval is a fifth, the ratio is $3:2$ [12]. The middle *A* note, A4, is set to be exactly 440 Hertz, whereas in octave 5, the frequency of A5 is double that of A4, i.e. exactly 880 Hertz. The two notes being an octave apart thus confirms the $2:1$ ratio of frequencies. Anecdotally, many have found a strong causative correlation between mathematical and musical ability. To this end, neuro-scientific research has shown that when doing math and playing music, there turns out to be an overlap in brain activation and working memory. Many of the same parts of the brain are activated during the tasks of calculations and music transpositions [13]. Research has also shown that real world tasks of music or math, usage of the working memory despite the task will still be high. There remains much to be done in exploring the numerous ways in which music and mathematics are interconnected, and our approach highlights just one of the many deep connections that are yet to have been discovered.

References

[1] David L. Forrest. Pl voice leading and the uncanny in pop music. *Music Theory Online*, 23(4), December 2017.

[2] Alexander Rehding. *Hugo Riemann and the Birth of Modern Musical Thought*. Cambridge University Press, 2003. Sample PDF available at https://assets.cambridge.org/97805218/20738/sample/9780521820738ws.pdf, accessed 2025-11-16.

[3] Henry Klumpenhouwer. Dualist tonal space and transformation in nineteenth-century musical thought. In Thomas J. Christensen, editor, *The Cambridge History of Western Music Theory*, page 456–476. 2008. https://music.arts.uci.edu/abauer/5.2/readings/Klumpenhouwer_on_Dualism_CHWMT.pdf?utm_, accessed 2025-11-16.

[4] Richard Cohn. Neo-riemannian operations, parsimonious trichords, and their "tonnetz" representations. *Journal of Music Theory*, 41(1):1–66, 1997. https://music.arts.uci.edu/abauer/3.1/notes/Cohn_Neo-Riemannian%20Operations97.pdf, accessed 2025-11-16.

[5] Alissa S. Crans, Thomas M. Fiore, and Ramon Satyendra. Musical actions of dihedral groups. *arXiv preprint arXiv:0711.1873*, 2007.

[6] Joseph A. Gallian. *Contemporary Abstract Algebra*. Cengage Learning, 9 edition, 2017.
[7] Gemma Crowe. Group theory in music. https://web.socem.plymouth.ac.uk/wim/posters/PhD%2011%20-%20Gemma%20Crowe.pdf. Accessed: 2025-11-16.
[8] MusicTheory.net contributors. Lessons. https://www.musictheory.net/lessons. Accessed: 16 November 2025.
[9] Alissa S. Crans, Thomas M. Fiore, and Ramon Satyendra. Musical actions of dihedral groups. *American Mathematical Monthly*, 116(6):479–495, 2009. Accessed: 2026-01-26.
[10] EarMaster. Consonance and dissonance, 2013. EarMaster Music Theory Online.
[11] MixButton. Music note to frequency chart, n.d. Accessed: 2026-03-01.
[12] Lluis L. Trulla, Nicola Di Stefano, and Alessandro Giuliani. Computational approach to musical consonance and dissonance. *Frontiers in Psychology*, 9:381, 2018.
[13] Elian ten Hoopen et al. Neural mechanisms of rhythm and pattern perception in the human brain. *Frontiers in Human Neuroscience*, 16:866256, 2022.

Screening of Selected Clay Minerals for RNA Compatibility: Implications for Prebiotic Chemistry on Early Earth and Mars By Nidheesh Prathi

Bal Bharati Public School, Pitampura, New Delhi, India
Global R-Hub, India

Abstract

Clay minerals are present everywhere on Earth. They are suggested as substrates for the protection, concentration, and catalysis of RNA in the origin of life hypothesis. This study introduces secondary data analysis that utilizes the oxide composition of three mineral clays, namely STx-1 montmorillonite, NG-1 nontronite, and IMt-1/IMt-2 illite, to forecast their comparative capacity for RNA preservation from the Clay Mineral Society
To compare with the existing experimental literature available review a two-dimensional ratio space was prepared regarding RNA stability, nucleotide adsorption, and catalysis on clay minerals using Jamovi software. The results indicated that basic chemical ratio such as Al_2O_3/Fe_2O_3 and SiO_2/Al_2O_3 can effectively categorize clay minerals as RNA-friendly or RNA-poor preservation by clay minerals, offering a realistic framework for future experiments to be conducted on Clay -RNA interactions.

Introduction

Earth was roughly formed 4.54 billion years ago. Early conditions on Earth is believed to be harsh and devoid of oxygen for any human being to exist was out of question at that time.Later, in 1960, work on molecular biology raised a fundamental question -the chicken and egg question: How could a complex network of nucleic acid and proteins arise from simple prebiotic chemistry? Work carried out by three researchers, namely Crick, Orgel and Woese showed a contribution to the prebiotic stages of life. They proposed that life on earth might have begun through RNA, which could catalyze its own replication. So mineral surfaces have been essential for preserving nucleic acids present in RNA.

Clay minerals play an important role in holding various kind of biomolecules that exists in nature. Early experimental and theoretical studies from 2000's shows that clays especially montmorillonite which can strongly adsorb nucleotides and can catalyze RNA oligomer formation, supporting the RNA world hypothesis (Ferris, "Montmorillonite-Catalysed Formation"; Franchi and Gallori). Nontronite and Illite are also believed to adsorb biomolecules. Clay minerals might have served as vessels for purines, pyrimidines, and other biomolecules such as uracil.

Recent astromaterials that have come to Earth highlight the importance of organic and mineral interactions in primordial chemistry. The samples retrieved by NASA's OSIRIS-REx mission from the carbonaceous asteroid Bennu include bio-essential sugars, such as ribose—the sugar backbone of RNA—alongside glucose and other pentoses and hexoses, in addition to amino acids and nucleobases (Furukawa et al.).The findings indicated by researchers show that

essential nucliec acid present in RNA can originate in alien conditions and can be transported to planetary surfaces, suggesting clay-rich minerals might have acted as substrate or RNA might have been preserved on clay-rich minerals.

This study aimed to analyze the chemical ratios of three clay minerals -montmorillonite, illite, and nontronite—as potential candidates for RNA-preserving, moderately preserving, or poor-preserving clays, informed by published evidence on their surface reactivity and catalytic behavior.

Methodology

Three clays were employed in this work, along with chemical ratio, bulk oxide and some physical attributes of the clay were pulled out from "source Clays" as specified by clay minerals society (Clay Minerals Society; Mermut and Faz Cano; van Olphen and Fripiat), which were used to compute SiO_2/Al_2O_3 and Al_2O_3/Fe_2O_3 ratios for analysis.

The ratio based method was adopted as first order screening to identify mineral properties important for RNA adsorption and their stability.These properties mainly include Cation Exchange Capacity (CEC), surface charge characteristics and redox behaviour. A comparison and Bar graph were prepared using the software Jamovi.

This study is secondary study which does not include direct lab validation. Instead it offers a hypothesis frame work that can be used to rank mineral clays for future experiments in astrobiology field.

Results

Table 1. Elemental Compositions of Three Clays, Adapted from the Clay Minerals Society

Oxide / Property	**STx-1 Montmorillonite (Texas, USA)**	**NG-1 Nontronite (Hohen Hagen, Germany)**	**IMt-1/IMt-2 Illite (Silver Hill, Montana, USA)**
SiO_2 (%)	70.10	45.80	49.30
Al_2O_3 (%)	16.00	5.93	24.25
TiO_2 (%)	0.22	0.05	0.55
Fe_2O_3 (%)	0.65	32.20	7.32
FeO (%)	0.15	0.00	0.55
MnO (%)	0.009	0.01	0.03
MgO (%)	3.69	1.02	2.56
CaO (%)	1.59	1.95	0.43
Na_2O (%)	0.27	0.03	0.00

Oxide / Property	STx-1 Montmorillonite (Texas, USA)	NG-1 Nontronite (Hohen Hagen, Germany)	IMt-1/IMt-2 Illite (Silver Hill, Montana, USA)
K_2O (%)	0.078	0.13	7.83
P_2O_5 (%)	0.026	0.02	0.08
S (%)	0.04	–	–
F (%)	0.084	–	–

Table 2. Calculated Major Oxide Ratios (SiO_2/Al_2O_3 and Al_2O_3/Fe_2O_3) for Three Clay Minerals Society Source.

Clay	SiO_2/Al_2O_3	Al_2O_3/Fe_2O_3
STx-1	70.10 / 16.00 ≈ 4.38	16.00 / 0.65 ≈ 24.62
NG-1	45.80 / 5.93 ≈ 7.72	5.93 / 32.20 ≈ 0.18
Illite	49.30 / 24.25 ≈ 2.03	24.25 / 7.32 ≈ 3.31

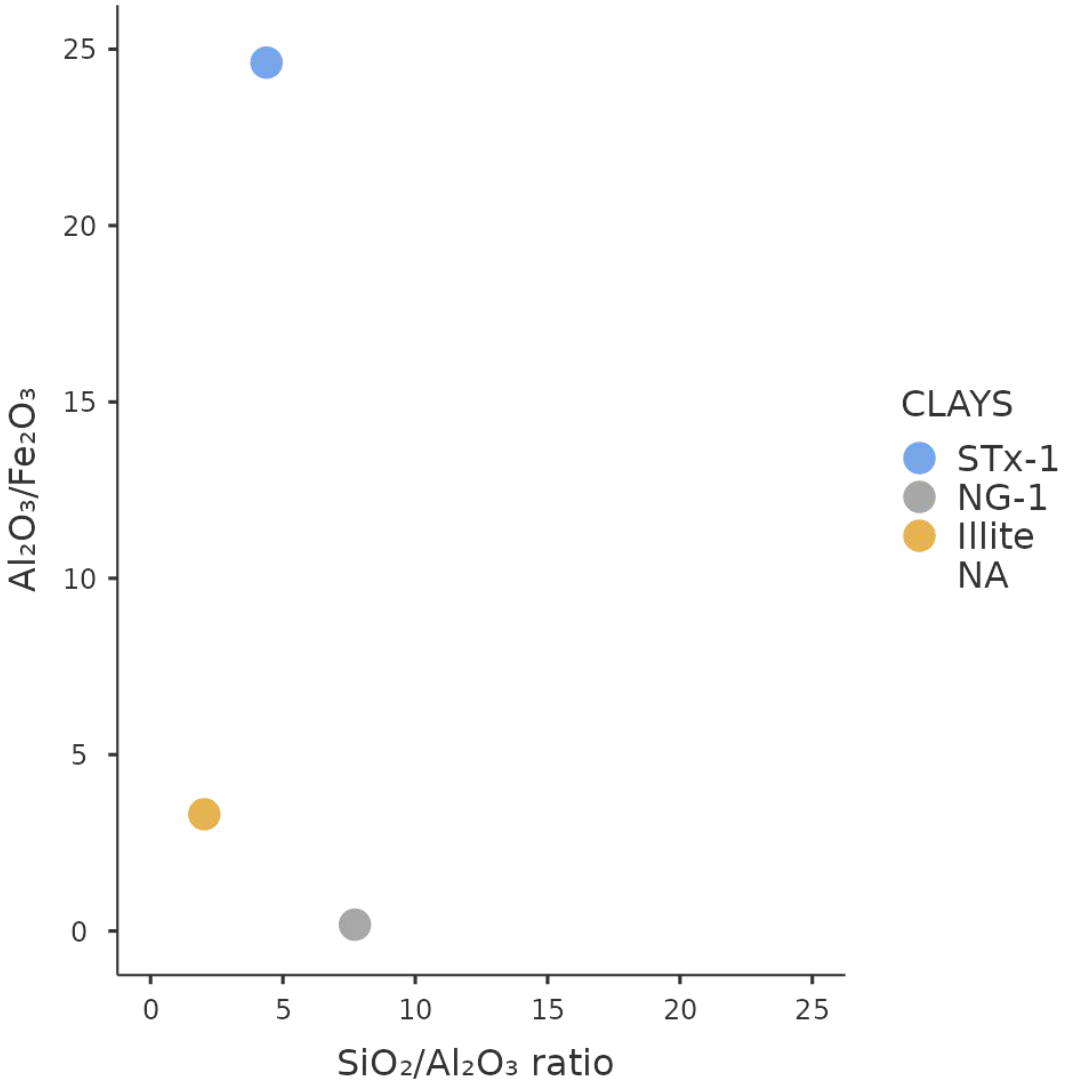


Fig. 1. Oxide ratio scatter plot: two-dimensional oxide ratio space for screening RNA–clay interactions.

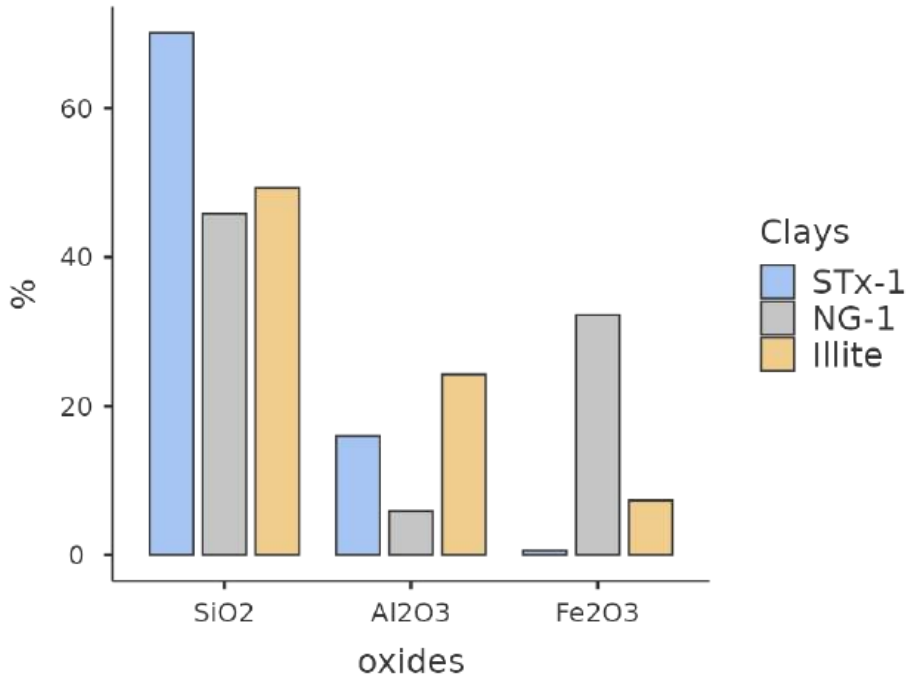


Figure 2: Bar graph showing the proportion of major oxides in percentage of STx-1 montmorillonite, NG-1 nontronite andIMt-1/IMt-2 illite,

Discussion

The above results, especially in Table 2, show oxide ratios generated to evaluate three potential clay mineral candidates: STx-1 montmorillonite, IMt-1/IMt-2 illite, and NG-1 nontronite for their capability to act as RNA -friendly or RNA-poor clay minerals in origin of life settings. Figure 1 shows an analysis of the ratio between SiO_2/Al_2O_3 against Al_2O_3/Fe_2O_3 showed different compositional space in 2D ratio space.

The ratio-based method revealed that even secondary data can also help to generate hypotheses regarding which clay minerals might have played an important role in preserving RNA to support prebiotic chemistry on Earth and Mars. The most significant aspect in this paper is Figure 1, showing the ratio plot positioning of STx-1 montmorillonite. STx-1 demonstrates a moderate SiO_2/Al_2O_3 ratio (≈4.38) and a significantly elevated Al_2O_3/Fe_2O_3 ratio (≈24.62), signifying an aluminum-rich, iron-poor composition. This positions STx-1 in the upper segment of the ratio spectrum, distinctly separated from the other clays along the Al_2O_3/Fe_2O_3 axis. Montmorillonite is already known in literature for origin-of-life research for its ability to adsorb activated nucleotides and facilitate the synthesis of RNA oligomers, as well as to concentrate RNA on its charged surfaces (Ferris, “Montmorillonite-Catalysed Formation”; Yu et al.).

On the ratio plot, if we see Figure 1, illite is situated in the lower left quadrant, nearer to the origin but distinctively different from nontronite and montmorillonite. Mineralogically, illite is mica-like clay characterized by significant K_2O and a reduced cation-exchange capacity compared to smectites; a substantial portion of its K^+ is retained within the inside framework

rather than being readily exchangeable and soluble. These features suggest that illite can bind RNA and other biomolecules, but with less interlayer flexibility and fewer exchange sites than montmorillonite. Researchers have found that nucleic acid and protein adsorption indicate that illite is involved in biomolecule binding, but less strongly than high-CEC smectites (Yu et al.). Its intermediate position in the ratio space, therefore, supports its qualitative classification as "moderate" for RNA preservation—neither strongly RNA-friendly nor clearly RNA-destructive, while montmorillonite is suggested as RNA -friendly.

An important outcome of this piece of study was that very simple geochemical descriptors—SiO_2/Al_2O_3 and Al_2O_3/Fe_2O_3 were sufficient to reproduce the ordering of three clays ranging from RNA-friendly to RNA-Poor gradient.Before conducting any prior experimental adsorption experiments in the lab, the spatial distribution of clays may be helpful as a preliminary screening source.

The present study has some limitations, such as it overlooks environmental variables, such as pH, temperature, ionic strength, and the presence of divalent cations. These gaps indicate clear directions for future research. Although theoretical and limited in scope, this secondary-data screening demonstrates how simple geochemical metrics can bridge mineralogy and prebiotic chemistry, guiding future laboratory experiments and projects that explore how clay minerals might have helped RNA multiply and preserve on early Earth and Mars (Ertem; Ferris; Kloprogge and Hartman; NASA Astrobiology Program; Rensselaer Polytechnic Institute NAI Team).

Conclusion

This study can function as a reliable first-order proxy for evaluating the RNA interaction potential of clay minerals, basically the oxide ratios mentioned in the study. Montmorillonite, distinguished by its aluminum-rich and iron-poor composition, functions as the optimal or RNA-friendly category substrate for RNA adsorption and storage, while iron-rich clays like nontronite are relatively inhospitable to RNA. Illite represents an intermediate condition, promoting limited RNA stability. These findings support the utilization of geochemical screening to prioritize prospective minerals for experimental investigations into the origins of life and astrobiology on early Earth and Mars.

Acknowledgements

The author sincerely acknowledges Global R-Hub, India for providing guidance and support throughout this research. The author also extends heartfelt gratitude to his mentor, Dr. Pooja Bachani, for her continuous guidance, encouragement, and invaluable support throughout the entire research process.

Works Cited

Clay Minerals Society. “Physical and Chemical Data of Source Clays.” Clay Minerals Society, 2020, www.clays.org/sourceclays_data/.

Crick, Francis H. C. “The Origin of the Genetic Code.” Journal of Molecular Biology, vol. 38, 1968, pp. 67–79.

Ertem, Gözen. “The Role of Minerals in Events That Led to the Origin of Life.” Astrobiology, vol. 21, no. 2, 2021, pp. 137–150.

Ferris, James P. “Catalysis and Prebiotic Synthesis.” Reviews in Mineralogy and Geochemistry, vol. 59, no. 1, 2005, pp. 187–210.

Ferris, James P. “Montmorillonite-Catalysed Formation of RNA Oligomers: The Possible Role of Catalysis in the Origins of Life.” Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 361, no. 1474, 2006, pp. 1777–1786.

Franchi, Marco, and Enzo Gallori. “A Surface-Mediated Origin of the RNA World: Biogenic Activities of Clay-Adsorbed RNA Molecules.” Gene, vol. 346, 2005, pp. 205–214.

Furukawa, Yoshihiro, et al. “Bio-Essential Sugars in Samples from Asteroid Bennu.” Nature Geoscience, 2025, pp. 1–6.

Kloprogge, Jacob Teunis, and Hyman Hartman. “Clays and the Origin of Life: The Experiments.” Life, vol. 12, no. 2, 2022, p. 259.

Mermut, Ahmet R., and Angel Faz Cano. “Baseline Studies of the Clay Minerals Society Source Clays: Chemical Analyses of Major Elements.” Clays and Clay Minerals, vol. 49, no. 5, 2001, pp. 381–386.

NASA Astrobiology Program. “RNA Origins in Sheets of Clay.” NASA Astrobiology, 19 Feb. 2012, astrobiology.nasa.gov/news/rna-origins-in-sheets-of-clay/.

Orgel, Leslie E. “Evolution of the Genetic Apparatus.” Journal of Molecular Biology, vol. 38, no. 3, 1968, pp. 381–393.

Rensselaer Polytechnic Institute NAI Team. “Project 7: Prebiotic Chemical Catalysis on Early Earth and Mars.” NASA Astrobiology Institute Annual Report, 2011.

van Olphen, Hendrik, and J. J. Fripiat. “Data Handbook for Clay Materials and Other Non-Metallic Minerals.” Soil Science, vol. 131, no. 1, 1981, p. 62.

Woese, Carl R. The Genetic Code: The Molecular Basis for Genetic Expression. Harper & Row, 1967.

Yu, Wei Hua, et al. “Adsorption of Proteins and Nucleic Acids on Clay Minerals and Their Interactions: A Review.” Applied Clay Science, vol. 80, 2013, pp. 443–452.

Preserving Little Saigon: Economic Stability and Cultural Continuity in Westminster, California By Andy Cao

Introduction

Little Saigon in Orange County is the largest and most institutionally developed Vietnamese commercial and cultural center in America. According to the 2020 Census, there are 241,557 people identified as Vietnamese living in Orange County, which has the largest Vietnamese population of any U.S county. A 2024 economic profile identified 99,585 Vietnamese residents within a broader 41-census-tract Little Saigon study area (U.S. Census Bureau, 2023; Puri, Popp, & Fleissig, 2024). The extending district into neighboring cities centered on Bolsa Avenue in Westminster, containing restaurants, grocery stores, medical and professional services, media organization, religious institutions, and civic spaces built through decades of refugee settlement and entrepreneurship (National Park Service, 2017; University of California, Irvine, Vietnamese American Oral History Project, 2015). This paper focuses on answering the following question of how Little Saigon can preserve its cultural identity while strengthening the economic conditions that allow immigrant-owned business, workers, renters and future generations to remain in the district.

The district is not just a symbolic homeland but it is also a dense economy with small businesses. There are 11,252 establishments, 49,896 workers and approximately $2.02 billion in annualized payroll in the second quarter of the year 2023 (Puri, Popp, & Fleissig, 2024). The average business employed 4.4 employees, indicating that the area's economic identity replies heavily on small firms rather than a few large employers (Puri, Popp, & Fleissig, 2024). The narrower 1.25-mile Bolsa corridor examined by Westminster contained more than 700 businesses and generated an estimated $938.6 million in annual sales, demonstrating that a relatively compact commercial strip carries substantial regional economic weight (City of Westminster, 2021). These figures make business continuity, commercial occupancy, labor conditions, and customer access central preservation issues rather than secondary policy concerns.

Little Saigon's success also comes with vulnerabilities. Due to its structure of being an cluster of small establishments, it has limited capacity to absorb rent increases, digitize operations, plan for succession, or just survive a temporary decline in customer traffic. The broader area had an average annual payroll of $40,519 per worker in 2023, including both full- and part-time employment, compared with $73,564 across Orange County; this comparison does not measure identical job mixes, but it signals a lower-wage local economy with less household capacity to absorb rising costs (Puri, Popp, & Fleissig, 2024). Housing conditions continue to intensify the pressure on businesses, in 2022, 52.6% of renter households in the study area spent more than 35% of their income on gross rent, compared to 46.1% countywide (Puri, Popp, & Fleissig, 2024). Cultural continuity is therefore threatened not only by assimilation but also by the possibility that the people who operate, staff, and patronize the district will find it increasingly difficult to remain nearby.

The analysis uses a qualitative comparative case-study approach supported by descriptive economic evidence. It combines Census data, government plans, state employment statistics compiled by the California State University, Fullerton, public records, and community-organized materials. This paper focuses on policy comparison among Little Tokyo, Calle 24, and Sawtelle Japantown as each illustrates a different relationship among cultural recognition, land-use regulation, affordability, business support, and public-realm investment (Los Angeles City Planning, 2020; San Francisco Planning Department, 2017; Sawtelle Japantown Association, n.d.). It evaluates whether proposed tools match Little Saigon's documented risk, whether responsibilities and funding are identified, and whether outcomes can be measured over time.

This paper argues that that Little Saigon's cultural preservation depends on economic preservation. The district's restaurants, markets, services, religious institutions, media, and public rituals survive only when entrepreneurs can transfer or renew businesses, workers and customers can reach the corridor, and residents can withstand housing and commercial-cost pressures. Westminster's current revitalization strategy is valuable but incomplete because it emphasizes branding and physical improvements more clearly than affordability, business succession, ownership concentration, and measurable protection against displacement. The paper first explains how refugee settlement and entrepreneurship created the district, then analyzes its present economic structure and vulnerabilities, compares policy tools used in three other ethnic commercial districts, evaluates Westminster's Little Saigon Blueprint, and proposes an integrated preservation strategy (City of Westminster, 2021; Puri, Popp, & Fleissig, 2024).

Background

The origins of Little Saigon are rooted in the fall of Saigon on April 30, 1975, when more than a hundred thousand Vietnamese refugees left their country and resettled in the United States during the first major Vietnamese immigration wave. A large portion of the refugees initially passed through Camp Pendleton in Southern California, which placed many refugees close to Orange County. In addition, in the 1970s, Orange County offered many assembly line jobs in the computer, electronics, and defense industries that required few language or technical skills, which provided an economic foothold for the Vietnamese refugees who are still currently learning English and navigating in an unfamiliar environment. Eventually, this led to the second immigration wave of Vietnamese who had initially settled in other states to relocate to Orange County, because they were drawn by the existing community there.

The emergence of Little Saigon was shaped by economic geography as well as cultural affinity. Camp Pendleton's role in receiving roughly 50,000 Vietnamese refugees placed many newcomers near Orange County, while local churches, relatively affordable housing, available jobs, and undeveloped or inexpensive commercial land encouraged settlement in Westminster and Garden Grove (University of California, Irvine, Vietnamese American Oral History Project, 2015). These conditions lowered the cost of rebuilding households and starting firms. Once Vietnamese markets, restaurants, professional services, newspapers, and religious institutions appeared, they created a reinforcing market: new arrivals could find familiar goods, information, credit relationships, employment, and customers in the same area. The district therefore grew

through agglomeration, in which proximity reduced search and transaction costs for both businesses and consumers.

Commercial property development helped give this refugee economy a durable physical center, but it also concentrated influence over the district's future. Bridgecreek Development, associated with Frank Jao, historically developed more than one-third of the commercial district, and contemporary reporting in 1997 estimated that roughly half of Little Saigon's retail and office space then paid rent to Jao (Orange County Board of Supervisors archive, n.d.; Los Angeles Times, 1997). Those figures should not be presented as a current ownership map, yet they reveal an important structural issue: cultural districts can depend on decisions made by a relatively small number of property owners. Preservation policy must therefore consider not only who operates businesses but also who controls leases, redevelopment, maintenance, and the timing of property transitions.

The formal recognition of Little Saigon as a distinct cultural district also came through years of community advocacy. On February 10, 1988, the Westminster Redevelopment Agency designated a 1.5 mile section of Bolsa Avenue as the "Little Saigon Tourist Commercial District" which ensured the area's official cultural recognition. In addition, the opening of the Asian Garden Mall (also known in Vietnamese as Phước Lộc Thọ which symbolizes fortune, prosperity, and longevity) became the most visible symbol of this cultural claim. This history further explains why cultural and economic preservation must come hand-in-hand. Vietnamese entrepreneurs did not just simply sell merchandise, but they converted refugee networks, language skills, family labor, and community trust into a functioning commercial ecosystem. The district's official recognition in 1988 gave them a public name but it could not guarantee that its firms, employees, and residents would remain economically secure (National Park Service, 2017). The relevant question today is whether the conditions that once supported entry - affordable space, accessible customers, dense social networks, and opportunities for small-scale enterprise - can be renewed for a younger and more economically diverse generation.

Economic Structure of Little Saigon

How Little Saigon is defined affects its economic claims. Westminster's 2021 Blueprint studies a narrow 1.25-mile stretch of Bolsa Avenue. Meanwhile, according to the 2024 California State University, Fullerton profile defines a wider 41-census-tract area in which at least 30% of residents identify as Vietnamese or in combination (City of Westminster, 2021; Puri, Popp, & Fleissig, 2024). The narrower definition is useful for evaluating storefronts, traffic, public space, and municipal interventions. The broader definition is more applicable for understanding households, employment, and industry composition. Therefore, the figures are more complementary of each other rather than interchangeable, and whenever a statistic is presented, the relevant geography should be included in the revised paper.

There are 11,252 establishments and 49,896 workers in the second quarter of 2023 in the broader area of Little Saigon, with annualized payroll of around $2.02 billion (Puri, Popp, & Fleissig, 2024). Dividing employment by establishments yields an average of 4.4 workers per

establishment, confirming that very small firms are a defining feature of the local economy (Puri, Popp, & Fleissig, 2024). Health care and social assistance was the largest sector by establishment count, while retail, food services, personal services, and professional activities also formed visible parts of the district's commercial base (Puri, Popp, & Fleissig, 2024). This mix matters because preservation policy designed only for restaurants and tourism would overlook medical offices, care providers, accountants, media, beauty services, and other firms that serve residents throughout the life cycle.

The narrower Bolsa corridor demonstrates unusually high commercial concentration. Westminster estimated that more than 700 businesses along the 1.25-mile study area generated $938.6 million in annual sales (City of Westminster, 2021). Although sales are not the same as local income or profit, the figure shows that the corridor attracts spending beyond its immediate residential population. Little Saigon functions simultaneously as a neighborhood service center, an ethnic marketplace, and a regional destination. Policy should protect all three roles: daily necessities sustain local households, specialized Vietnamese goods draw a dispersed diaspora, and festivals or cultural tourism introduce new customers. Overreliance on any one market would make the district more vulnerable to demographic or consumer change.

The uniformly high earnings does not correspond with the district’s scale . Average annual payroll per employee was $40,519 in the broader area in 2023, compared with $73,564 across Orange County; due to the measurement including both full-time and part-time workers and reflecting a mix in different industries, it should not be interpreted as a direct wage comparison, but rather as a warning indicator (Puri, Popp, & Fleissig, 2024). Housing costs further emphasize the concern. Even though the area's typical rent of $1,900 in 2022 was noticeably lower than the county average of $2,248, locals were still feeling the squeeze. More than half (52.6%) of renter households had to spend over 35% of their income just to cover rent (Puri, Popp, & Fleissig, 2024). Lower absolute rent therefore does not necessarily mean affordability when household incomes are also lower.

Little Saigon's economic system is also highly automobile-dependent. According to the City of Westminster (2021), 97% of youth survey participants mentioned accessing the corridor by automobile. Additionally, daily traffic volumes were recorded to be between 32,000 and 34,000 vehicles, physical obstacles that hinder their walking between local destinations were also identified. 4.6% of households had no vehicle in the whole county while it is 6.8% in the broader study area, suggesting that unequal access is present due to the car-oriented layout even within a car-dependent region (Puri, Popp, & Fleissig, 2024). Dense traffic, poor pedestrian crossings, weak inter-parcel connections and limited amenities can result in the reduction of the number of stores a customer visits per trip. Hence, transportation should be evaluated as a tool that can improve the economy as it changes customer circulation,worker access, and business visibility.

Economic Vulnerabilities and Cultural Continuity

Little Saigon's central challenge is not whether its culture is meaningful; it is whether the economic platform that carries that culture can adapt without excluding the people who created

it. The district depends on small establishments, lower-paid service work, automobile access, and a renter population facing substantial cost burdens (Puri, Popp, & Fleissig, 2024). At the same time, many founding entrepreneurs are approaching retirement, younger Vietnamese Americans use the district differently, and public or private reinvestment may raise property values without automatically protecting legacy firms. These pressures are connected. A retiring owner who lacks a successor is more likely to close; a vacant or repositioned storefront may attract a higher-rent tenant; and a worker facing high housing and transportation costs may leave the area even if the cultural destination itself remains popular.

The area's very small to average establishment size creates both advantages and disadvantages. Small firms can respond quickly to community demand while preserving specialized knowledge, but they usually have limited time and resources for digital marketing, online ordering, financial planning, legal review, accessibility upgrades, or emergency reserves. Westminster's Blueprint specifically identifies business valuation, legal support, succession planning, and digital capacity as critical needs for corridor firms (City of Westminster, 2021). The economic goal should not be to replace smaller firms with larger operators. The aim should be to lower the fixed costs of adaptation by providing shared technical assistance, trusted bilingual advice, and coordinated marketing that no single small firm could afford on its own.

The clearest point at which generational change becomes an economic mechanism is the business succession. Even with strong consumer demands, legacy businesses frequently face closure due to the lack of an heir. The operational vulnerabilities are caused by a lack of formal financial records, the lease cannot be transferred, or simply the successors do not want to take on the workload. According to the City of Westminster (2021), Westminster's Blueprint noticed this issue but does not provide a public accounting of at-risk firms or measurable succession outcomes . The revised paper should differentiate succession from nostalgia: the policy aim is not to freeze every business in aspic, but to preserve viable firms, knowledge, supplier relationships, and culturally significant uses whilst allowing new owners and business models to emerge.

Evidence of youth disengagement is suggestive but limited. Westminster's Blueprint reports a survey of 147 young respondents distributed through the city's youth committee and two local high schools; 45% said they visited Little Saigon never or only a few times, 44% identified inconvenient access, 28% cited the tenant mix, and 17% cited the lack of an anchor destination (City of Westminster, 2021). Because this was not a representative survey of all young Vietnamese Americans, it should not support a broad claim that an entire generation rejects the district. It does, however, identify testable barriers. A pilot program can measure whether safer walking routes, evening events, shared public space, youth-led pop-ups, or a broader mix of affordable activities changes visit frequency and spending.

Real housing affordability is not simply about the rental costs, but rather about its relativity to income. In 2022, the median gross rent in the broader Little Saigon area was $1,900, below Orange County’s median of $2,248, but 52.6% of renters paid more than 35% of their household income on gross rent, compared with 46.1% countywide (Puri, Popp, & Fleissig,

2024). It is suggested that despite some retained lower-cost housing in the area, it still places a heavier burden on local renters. While residents displaced by rising costs may occasionally visit the area, their geographical distance ultimate diminishes daily retail patronage, shrinks local labor supply, and disrupts both community care networks and civic engagement. Housing stability is therefore part of commercial and cultural stability.

The Green Lantern mobile-home-park dispute showed redevelopment pressure without proving completed displacement. A proposed conversion threatened approximately 130 mobile-home spaces occupied largely by elderly, disabled, low-income, and Vietnamese residents, but the owner terminated the conversion application on January 16, 2020, after resident organizing and legal advocacy (Disability Rights California, 2020; Los Angeles Times, 2020). The case is useful because it shows how quickly vulnerable residents can face uncertainty when land values and redevelopment plans change. It also shows that organized residents, legal protections, and public scrutiny can alter an outcome. The paper should present the case as a warning and a governance lesson, not as an example of residents who were actually removed.

Commercial displacement risk cannot be evaluated without better information about ownership and leases. Historical evidence shows that large developers played an important role in assembling and building the district, but the available sources do not provide a current parcel-level picture of ownership concentration or commercial-rent trends (Orange County Board of Supervisors archive, n.d.; Los Angeles Times, 1997). This is an evidence gap, not permission to assume either stability or crisis. Westminster should create a voluntary and confidential commercial conditions survey that tracks asking rents, lease lengths, vacancy, ownership changes, and business closures in aggregate. Public reporting can reveal trends without publishing individual firms' proprietary lease terms.

The corridor's physical layout affects the distribution of economic value. Heavy traffic may lead customers to the district, but wide roads, fractured parking lots, and difficult crossings discourage visitors from walking among multiple businesses. According to the City of Westminster in 2021, the Blueprint showed 32,000 to 34,000 daily vehicles and noted obstacles to pedestrian mobility, while 97% of surveyed youth reported arriving by car. A visitor who drives to one destination and leaves produces less spillover demand than a visitor who comfortably walks to several stores. Streetscape investments should therefore be evaluated through measures such as pedestrian counts, cross-shopping, dwell time, collision data, and sales or vacancy trends, not only through photographs of new landscaping.

Taken together, the risks are cumulative rather than isolated. Small-business capacity affects succession; succession affects storefront continuity; housing costs affect the workforce and customer base; ownership patterns affect rent and redevelopment; and transportation design affects who can reach and circulate through the district. The comparative cases should therefore be judged by a consistent question: Which policies change these underlying economic conditions, and which policies mainly improve cultural visibility without changing who can afford to remain?

Comparative case studies

The three comparison districts are not templates to copy wholesale. Little Tokyo illustrates the pairing of cultural institutions with affordable housing and organized community development; Calle 24 demonstrates how land-use controls and legacy-business assistance can address commercial change; and Sawtelle Japantown shows the visibility created by naming, branding, and a renewed restaurant economy (Los Angeles City Planning, 2020; San Francisco Planning Department, 2017; Sawtelle Japantown Association, n.d.). Each case operates under different land markets, municipal powers, and community institutions. The relevant lesson is therefore a policy mechanism, not a claim that identical results will follow in Westminster.

1. Little Tokyo

The case of Little Tokyo demonstrates that cultural recognition does not ease land pressure. The National Trust for Historic Preservation (2024) has documented threats from development and increasing rents even in a district with nationally recognized institutions and a strong visitor identity. Los Angeles has also adopted community design guidelines that describe how new development should relate to the scale and cultural character of the neighborhood (Los Angeles City Planning, 2020). These tools protect visibility and form, but they are most effective when combined with institutions that are capable of acquiring land, producing housing, or negotiating community benefits.

The First Street North project illustrates this institutional role. Current Little Tokyo Service Center materials describe 165 affordable rental units and 80 permanent supportive-housing units, for a total of 245 homes, alongside community-serving development (Little Tokyo Service Center, n.d.). The lesson for Little Saigon is not that one mixed-use project can preserve an entire district. It is that a mission-driven intermediary can connect cultural planning to land control and long-term affordability. Westminster should therefore identify whether an existing nonprofit, public agency, or new community development corporation could assemble financing, hold property, and protect culturally significant uses over time.

2. Calle 24 Latino Cultural District (Mission district)

Calle 24 adopted targeted land-use rules after measurable demographic and commercial change. According to the San Francisco Planning Department (2017), the analysis showed that the share of households at or lower than 120% of area median income fell by 10 percentage points between 2000 and 2016, meanwhile the share above 200% rose by 9 points, and the Hispanic population share drops by 11 points in the study area. The Calle 24 Special Use District and design guidelines respond to this by regulating aspects of storefront size, use, and physical character rather than relying on branding alone (San Francisco Planning Department, 2017; San Francisco Planning Department, 2020). Even though these policies cannot control all rents, it does contribute to slowing down the physical consolidation that favors larger tenants and can preserve a fine-grained commercial pattern compatible with small firms.

San Francisco's Legacy Business Program adds a direct financial instrument. From fiscal year 2016-2017 through 2021-2022, the program reported 178 rent-stabilization grants totaling approximately $2.739 million, an average of $15,390 per grant, within a program whose annual

budget was $1 million (City and County of San Francisco, 2023). Westminster cannot assume access to the same fiscal resources, but it can adapt the mechanism at a smaller scale: define eligibility transparently, require a lease extension or other preservation commitment, cap assistance, and publish aggregate outcomes. The essential lesson is that heritage designation becomes economically meaningful when it changes the cost or security of remaining in place.

3. Sawtelle Japantown

Similar to Little Tokyo, Sawtelle Japantown also went through decades of redevelopment pressure. Rising land values have fueled the replacement of historic nurseries, small businesses, and older homes with new commercial and residential developments. By the early 1990s, many longtime residents expressed concern that the neighborhood was gradually losing the character that had defined it for generations. The number of Japanese-owned nurseries declined dramatically as developers increasingly viewed the area as valuable real estate. Sawtelle received official recognition as Sawtelle Japantown in 2015 after a community campaign in order to change the public image of a district that was once a Japanese American nurseries and services into a contemporary restaurants and retail destination (City of Los Angeles, 2015; Sawtelle Japantown Association, n.d.; UCLA Newsroom, 2015). These elements are relevant to Little Saigon as a distinct neighborhood identity, together with strong visibility and easy navigation can directly drive up commercial value. At the same time, naming and popularity do not constitute an affordability strategy. Westminster can borrow the emphasis on gateways, signage, and contemporary businesses, but it should pair those tools with commercial and housing protections rather than infer that successful branding automatically preserves community control.

The cases reveal three different layers of preservation. Cultural designation and design standards make identity visible; business and land-use tools influence the survival of small storefronts; and affordable housing or mission-controlled property affects who can remain near the district. Little Saigon currently has the strongest momentum in the first layer through branding and streetscape planning. Its revision strategy should add the second and third layers while recognizing Westminster's more limited budget and institutional capacity. This framework provides a fairer basis for evaluating the Little Saigon Blueprint than asking whether the city has copied another neighborhood.

Evaluation of the Little Saigon Blueprint

Westminster's 2021 Little Saigon Blueprint should not be considered as the final preservation plan but rather as a useful agenda. It assigns rough cost categories of less than $50,000, $50,000 to $200,000, or more than $200,000 to different actions and identifies physical improvements, branding, events, online resources, business assistance, leadership, and possible management structures (City of Westminster, 2021). The document also recognizes succession planning, business valuation, legal assistance, and digital capacity as economic needs (City of Westminster, 2021). Its main weakness is not a lack of ideas. It is the absence of a single implementation framework showing which entity is responsible, how actions will be funded, and how the city will determine whether businesses and residents are actually becoming more secure.

Recent city actions show that implementation has started. Westminster received a $250,000 California Jobs First grant in 2025 and launched a Little Saigon streetscape feasibility study in April 2026 covering lighting, public art, benches, bus shelters, bicycle stands, and related improvements (City of Westminster, 2026a). A city project update also reports that an outreach coordinator was hired, the Beautify Bolsa effort served more than 50 businesses and property owners, and branding, gateway, and streetscape work remained active (City of Westminster, 2026b). These are concrete outputs, but evidence of longer-term results such as lower vacancy, successful business transitions, improved pedestrian circulation, stable commercial space costs, or less pressure to displace people are yet to be seen.

The Blueprint's youth findings are useful for program design but should only be interpreted within the limitations of their sampling. The survey received 147 responses via a city youth committee and Westminster and La Quinta high schools; it showed 45% visiting never or only a few times, 44% said inconvenient access, 28% cited the tenant mix, and 17% cited the absence of an anchor destination (City of Westminster, 2021). These results do not give conclusions for the whole population, and only justify the experiments. The city should repeat the survey with disclosed demographics, broader recruitment, and consistent questions before and after youth-oriented pilots.

The Blueprint is least developed where preservation requires continuing economic protection. It does not establish a commercial-rent or vacancy monitoring system, a funded succession pipeline, a legacy-business definition tied to benefits, a housing-stability program, or a current ownership map. It refers to a management entity or business improvement district, however, an assessment-based structure must address the possibility that small firms or property owners experience fees differently (City of Westminster, 2021). Prior to adopting a mandatory assessment, the city should make public the expected costs, voting rules, eligible services, exemptions or graduated contributions, representation of tenants and property owners.

The Blueprint should therefore be judged as a strong first-stage revitalization plan with an incomplete preservation architecture. It provides a credible foundation for public-realm improvements, identity, outreach, and business services, and recent grants show that the city can convert some recommendations into action (City of Westminster, 2026a; City of Westminster, 2026b). Its next phase should move from a list of projects to a coordinated economic strategy with baselines, annual targets, named implementing partners, and protections that reach businesses, workers, and residents.

Analytical Framework: Three Levels of Cultural Persistence

Discussions of ethnic commercial and cultural districts often treat cultural presence, such as a recognized name, a distinctive streetscape, or a concentration of restaurants and festivals, as evidence that a district is being preserved. The literature suggests this inference is unsafe. Hackworth and Rekers (2005) argue that in four Toronto business-improvement areas (Corso Italia, Little Italy, India Bazaar, and Greektown), ethnically defined commercial strips functioned as a branding mechanism that valorized nearby residential real estate. Some of these

neighborhoods were marketed as ethnic even though they were no longer home to significant numbers of people from those groups. Zukin (2010) likewise argues that the demand for "authentic" urban places pushes up prices and therefore drive out the immigrants, workers, and artists who first gave a neighborhood its character. To avoid conflating visibility with persistence, the author aims to form a framework to analyze each district at three levels: **(1) cultural presence**, **(2) economic continuity**, and **(3) the capacity of the community to remain present over the long term**. Level 1 is necessary but not sufficient. It endures only if supported by Level 2, which in turn depends on whether the community can stay and retain control of space (Level 3).

The three levels are analytically distinct but interdependent. Cultural presence can exist without ensuring that the original community can remain in the distinct. Therefore, the framework does not treat the three levels as a linear sequence or as a single aggregate measure of preservation. Instead, it examines the gaps and linkages between them, allowing a district to be assessed not only by how visible its cultural identity is, but also by whether the economic and community foundations required to reproduce that identity overtime remain intact.

This framework is especially useful for Little Saigon because the district is not only a concentration of Vietnamese businesses but also a cultural, institutional, and symbolic center of the Vietnamese American community in Orange County. Previous studies has framed Little Saigon as a spatial community that has been formed through the confluence of economic opportunities, community formation, cultural institutions, and political processes (Vo, 2008; Lieu, 2011). In particular, Mazumdar et al. (2000) contend that Little Saigon's sense of place is created by the combination of its built environment, daily social and commercial activities, and public ritual events that produce community identity and social bonds. These research findings indicate that Little Saigon ought to be viewed not just as a simple gathering of Vietnamese enterprises or cultural markers, but as a lively cultural and social setting where identity is constantly formed through the interactions between individuals, organizations, economic functions, and the environment.

The proposed approach builds upon this knowledge by establishing a general analytical framework to examine the factors that sustain cultural presence over the long term, and to facilitate comparisons across ethnic cultural regions.

Level 1: Cultural Presence

The first level of the framework, the level of cultural presence, concerns the degree to which the cultural identity of an ethnic community is integrated and articulated in a given space. Cultural presence is not only the visibility of ethnic symbols, but a combination of overt expressions, cultural institutions, ongoing cultural practices and official recognition, which together create the cultural characteristic of an area. This is consistent with studies of Little Saigon in Orange County that emphasize the interaction between the built environment, daily social and commercial activities, and public cultural events in creating a sense of place and community identity. It also draws on research into the function of cultural institutions and community organizations in constructing Vietnamese-American identity.

Table 1
Dimensions of Cultural Presence

Source: Created by the author.

Dimension	Meaning	Example of indicators
C1. Physical & Symbolic Presence	The extent to which cultural identity is embedded in the physical and visual environment of the district.	Architecture, ethnic-language signage, monuments, murals, public art, streetscape, culturally significant commercial spaces.
C2. Institutional presence	The existence and number of institutions that maintain, transmit, and represent the community's cultural identity.	Religious institutions (temples and churches), museums, cultural centers, community organizations, ethnic media, language/education programs.
C3. Cultural Practices	The extent to which culture is actively practiced and experienced by community members rather than merely displayed.	Festivals, religious ceremonies, community gatherings, language use, food traditions, performances, recurring cultural activities.
C4. Cultural Recognition	The extent to which the cultural significance of the district is formally recognized within planning and governance systems.	Cultural-district designation, historic recognition, special zoning, cultural overlays, design guidelines, municipal recognition.

Together, these four dimensions help distinguish between a region where culture is merely visible and one where cultural aspect is embedded, institutionalized, practiced, and recognized apparently. It is vital to emphasize this distinction as the presence of cultural assets alone does not necessarily indicate that the community has the economic and social conditions required to sustain its cultural identity over time. Therefore, the cultural presence forms the basis for the other levels of the framework, which will explore whether the viability of the economic activities and community structures sustaining the cultural presence in the long term.

Level 2: Economic Continuity

The second level, Economic Continuity, measures the ability of an economic system underpinning a cultural zone to stay healthy, stable and adaptable over time. Rather than measuring economic performance solely through aggregate indicators such as revenue or employment, this level focuses on the conditions that allow culturally significant businesses and commercial spaces to remain part of the district, without losing the culturally anchored used that define the district.

The idea is rooted in the research on immigrant and ethnic entrepreneurship, which acknowledges that immigrant-owned businesses can create economic opportunities and provide culturally specific goods, services, and social spaces in ethnic communities (Wilson & Portes, 1980). But such businesses can't be explained by co-ethnic networks alone. The "mixed-embeddedness" perspective highlights that immigrant firms are embedded not only in ethnic social networks, but also in the broader economic and institutional environment, such as markets, customers, suppliers, regulations and public institutions (Kloosterman et al., 1999). This perspective is particularly important for a cultural district, because the ability of culturally anchored businesses to stay in business relies on community demand as well as access to capital, commercial space, institutional support, and overall market conditions.

Nevertheless, economic continuity requires the district to adjust to shifting market conditions, while not allowing economic change to corrode its cultural purpose. Studies of commercial districts show that local economies can undergo major transformations in the mix of businesses, as locally-based stores and services are replaced by businesses serving new consumer groups and higher-income markets (Zukin et al., 2009). Thus economic growth and economic continuity are not necessarily the same: a district may remain commercially successful, but gradually lose the businesses, ownership structures and services that have historically connected its economy to the cultural community. This concern is especially relevant to Little Saigon, where the commercial district has grown and changed along with demographic and generational change. Research and local planning documents have identified the need to retain businesses and economic activities while adapting Little Saigon to changing consumer preferences and a changing regional economy.

Five criteria are therefore considered: (1) Business survival, (2) Ownership and lease security, (3) Access to financial resources, (4) Business adaptation and succession, and (5) Commercial diversity.

Table 2

Dimensions of Economic Continuity

Source: Created by the author.

Dimension	**Meaning**	**Example of indicators**
E1. Business Survival	Whether culturally significant businesses can remain operational and retain their presence in the district over time.	Business openings/closures, vacancy, business longevity, legacy businesses.
E2. Lease Security and Ownership	Whether businesses have have steady and affordable access to the commercial spaces they depend on.	Ownership, stability of lease, pressure of rent, length of lease.
E3. Availability of	The district and its businesses' ability	Grants, loans, CDFIs, tax

Funds	to access funding needed to support operations, investment and resilience.	incentives, public/private investment.
E4. Succession & Adaptability of business	Ability of businesses to adapt to changes in the market and to transfer ownership, knowledge and operations between generations.	Digitalization, new products/services, market adaptation, succession, second-generation ownership.
E5. Commercial Diversity	Whether the district maintains a sufficiently diverse economic base to support both everyday community needs and cultural functions.	Restaurants, retail, professional services, cultural businesses, service businesses, mix of business types.

Economic presence is not the same as economic continuity according to these criteria. A district may contain a large number of businesses at a given point in time, yet still face weak economic continuity if businesses are frequently displaced, commercial rents become unaffordable, access to capital is limited, or businesses lack mechanisms for adaptation and succession. By contrast, an economy is better off when companies can stay put, buy what they need, adjust to new conditions and have a varied portfolio of businesses. This approach fits well with research that shows immigrant-owned businesses are economically and culturally beneficial to neighbourhoods. But they also face the challenge of being displaced from commercial areas, rising costs and trouble getting the money they need. (Gandhi & Minner, 2017; Alvarez et al., 2026).

Level 3: Long-term Community Presence

The third level evaluates the community's social, demographic, institutional and intergenerational potential to maintain its association with the district over the long-term. Displacement theory sharpens what should be measured. Marcuse (1985) describes exclusionary displacement as the situation in which households that could formerly have moved into a unit can no longer do so. His typology runs from the most observable form, last-resident displacement, to less measurable forms such as exclusionary displacement and displacement pressure, so an assessment based only on who has left will understate the problem. Newman and Wyly (2006) find that displacement is limited in scale yet a crucial indicator of class polarization, and that the main buffers, public housing and rent regulation, are the very interventions under attack. Mirabal (2009) shows how heavily the Mission District, home of Calle 24, was gentrified in the 1990s and early 2000s. The temporal dimension draws on Throsby's (2014) notion of culturally sustainable development, which requires intergenerational and intragenerational equity. As for tools, Choi et al. (2018), using national data on community land trusts, conclude that these trusts may stabilize neighborhoods at risk of gentrification.

This third level is beyond the physical existence of cultural institutions and the economic viability of business to question if the community can continue to recreate its identity, social

networks and collective attachment to place. This divide is especially salient for Little Saigon, which has developed into more than a concentration of Vietnamese-owned enterprises. It has become a site of community-building, political activism, cultural expression, and collective memory. Studies of Little Saigon point to the importance of place in the formation and maintenance of Vietnamese American community and identity, and suggest that cultural institutions and collective activities have helped to transform the district into a broader social and symbolic center of Vietnamese American life (Aguilar-San Juan, 2009; Lieu, 2011; Mazumdar et al., 2000).

This level is operationalized through five dimensions: (1) Residential Continuity, which measures the stability of the community's residential presence; (2) Housing Affordability, which considers the capacity of community residents to afford to remain in the area; (3) Intergenerational Renewal, which looks at the continued participation of younger generations; (4) Social Infrastructure, which captures the institutions and spaces that support community life; and (5) Community Control and Governance, which considers the community's capacity to influence the ownership, management and development of resources that affect its future.

Table 3
Dimensions of Long-Term Community Presence
Source: Created by the author.

Dimension	**Meaning**	**Example of indicators**
L1. Residential Continuity	The degree to which the community sustains a stable residential presence and geographic link to the district over time.	Population, ethnic concentration, migration.
L2. Housing Affordability	The capacity of community members to continue to afford housing in or around the district	Rent burden, affordable housing units, displacement.
L3. Intergenerational Renewal	Whether younger generations continue to participate in and reproduce the community's social, cultural, and economic life.	Youth participation, second-/third-generation businesses.
L4. Social Infrastructure	The degree that institutions and physical locations provide services, networks, and activities that support community life	Schools, churches, temples, community centers.
L5. Community	Whether the community has	Nonprofit ownership,

Control & Governance	meaningful capacity to influence the ownership, management, and development of resources that impact the community's future.	community development corporations (CDCs), community land trusts (CLTs), community representation.

Comparative contribution using 4 case studies

Based on the aforementioned framework, this paper's main comparative contribution is to apply the same analytical framework to four distinctive cultural districts: Little Saigon in Orange County, Little Tokyo in Los Angeles, Calle 24 Latino Cultural District in San Francisco, and Sawtelle Japantown in Los Angeles. The idea of applying this three-level structure (Cultural Presence, Economic Continuity, and Long-term Community Presence) allows the comparison to take a deeper look at whether the economic and community conditions supporting that cultural visibility can be maintained over time.

(1) Level 1: Cultural Presence

Little Saigon

Little Saigon's cultural presence is represented in the visible commercial and symbolic features, community institutions, recurrent cultural events and formal acknowledgment. Its identity is expressed in a physical concentration around Bolsa Avenue and related commercial establishments, while restaurants, grocery stores, professional services, religious institutions, media organizations and community spaces make up its institutional and quotidian cultural presence. In 1988, Westminster officially designated the district as the "Little Saigon Tourist Commercial District," an example of how community identity was embedded into municipal status.

Little Tokyo

The comparison with Little Tokyo illustrates a somewhat different relationship between cultural identity and the built environment. Little Tokyo combines cultural institutions and community-serving development with formal design guidance. Los Angeles' Little Tokyo Community Design Overlay establishes design standards intended to relate new development to the neighborhood's cultural character, while the First Street North project connects cultural preservation with affordable and supportive housing. Thus, the Little Tokyo case demonstrates how C1 and C4 can be connected to institutional mechanisms that also influence the economic and residential conditions of the district.

Calle 24

Calle 24 offers a different arrangement of cultural acknowledgment. Visible ethnic businesses and community activities help sustain its cultural identity, as do the Calle 24 Special Use District and design requirements that restrict such things as storefront size, use, and physical character. This shows how C4 can be more than symbolic recognition, and be embedded in land-use and design legislation.

Sawtelle Japantown

Sawtelle Japantown, in contrast, illustrates the importance of awareness and branding in the reinvention of a cultural district's public image. In 2015, community effort led to the district's formal designation as Sawtelle Japantown, establishing a modern identity around Japanese American history, restaurants and retail. The story also, however, highlights a significant shortcoming of C1 and C4: increasing visibility and recognition do not establish affordability or guaranty the continuous presence of the original group.

(2) Level 2: Economic Continuity

Little Saigon

Little Saigon comprises 11,252 establishments, mostly small businesses with an average of 4.4 employees, which makes it structurally vulnerable to changes in demand by its customers and hikes in rent.

Regarding Lease Security and Ownership, the corridor has a long history of commercial property ownership, including Frank Jao's Bridgecreek Development, but there is limited tracking of commercial leases at the parcel level, and there are no mechanisms for commercial rent stability. Capital assistance is dispersed, too. Recent public funding includes a $250,000 California Jobs First award in 2025 for streetscape studies, but the area lacks ongoing dedicated CDFI financing or historic business grant programs.

First-generation firm owners are aging and face a generational succession dilemma, many without succession plans, formal accounting systems, lease-transferability arrangements, or the capacity for digital transformation. Despite these weaknesses, the economic ecology retains a significant commercial diversity, including retail, restaurants, professional services, and health care and social assistance, which cater to both ordinary community needs and tourism-related demand.

Little Tokyo

Little Tokyo is an example of how an organization with a mission may preserve economic continuity through the combination of property control, funding, and business support. Rising rents and real estate pressures threaten small cultural businesses, and community organizations purchase and control property to offer greater lease security.

This includes a diverse financing structure of municipal monies, philanthropic resources and LIHTC and CDFI investment. Community nonprofits also provide technical assistance and succession support to legacy firms to help them navigate market and generational shifts. Zoning laws that help balance historic businesses with civic and community-serving uses, along with these, help keep the neighborhood economically vibrant without converting it into a tourism-oriented commercial corridor.

Calle 24

Calle 24 is a model for how specific government actions can combine cultural conservation with economic integrity. San Francisco's Legacy Business Program provides financial and technical assistance to long-standing businesses, and its Special Use District and architectural rules help protect culturally significant storefronts and commercial uses. For instance, rent-stabilization grants allow landlords to offer longer leases, contributing to Mission

District commercial stability in the face of gentrification. Between 2016–17 and 2021–22, the program provided $2.739 million in rent-stabilization subsidies to 178 The Office of local Business and community liaisons helps legacy merchants with succession and regulatory obligations. Formula retail and commercial parcel merger limits protect local neighborhood businesses. These methods include regulation of land use, financial support and business assistance to prevent commercial displacement and preserve the culture and economy of the district.

Sawtelle Japantown

Sawtelle Japantown is a case in point of the difference between commercial regeneration and economic continuity. Recognition of the district helped establish a contemporary restaurant and retail identity, but redevelopment also displaced many traditional Japanese-owned businesses, particularly plant nurseries and early services.

Unlike Calle 24, Sawtelle has limited mechanisms for commercial rent stabilization or collective land control, leaving businesses exposed to market rents and redevelopment pressures. Economic adaptation has therefore occurred largely through commercial turnover rather than planned succession: newer food, Asian-fusion, and contemporary retail concepts have replaced many legacy operators. As a result, the district's commercial mix has become increasingly concentrated in food and beverage, reducing the range of everyday services available to longtime residents.

Sawtelle thus shows that commercial growth and economic continuity are not necessarily equivalent: a district can develop a successful new commercial identity while losing ownership continuity, legacy businesses, and parts of its community-serving economic base.

(3) Level 3: Long-Term Community Presence

Little Saigon

The sizeable Vietnamese population in Little Saigon contributes to maintaining the community's long-standing presence, although it still faces those challenges such as housing affordability and intergenerational participation. The 240,000 Vietnamese in Orange County, mostly in Little Saigon, offer a strong residential and social base. Median rents are lower than the median for Orange County, but 52.6% of renter households spend more than 35% of their income on gross rent, signaling housing cost pressure. There were not places for the young people to go, thus 45% of the youth in Westminster rarely or never went to Little Saigon. Such features suggest that housing stability and meaningful future engagement could renew communities far beyond demographic retention.

Little Tokyo

Little Tokyo is a more institutionalized approach of long-term community presence, particularly relating to the combination of inexpensive housing, social infrastructure, and community control. A project called First Street North, developed by the Little Tokyo Service Center and Go For Broke National Education Center, will include 248 units of affordable housing, 80 units of permanent supportive housing, and ground-floor space for legacy businesses and community-based cultural institutions. The initiative evolved from a longtime community

movement for better control of land and local development. This is an example of how community organizations can combine housing provision, social infrastructure and property control to minimize residential displacement and provide circumstances for intergenerational continuity.

Calle 24

Calle 24 illustrates the clash between community organizing and residential relocation. Between 2000 and 2016, the Mission District's Hispanic population fell from 50% to 39%, while the share of households earning at or below 120% of area median income dropped from 74.6% to 64.4%. These developments demonstrate that the cultural and civic infrastructure of the ancient area can withstand socioeconomic changes. Calle 24's grassroots council includes residents, merchants, kids, artists, nonprofits, renters and homeowners, and its land use and housing initiatives tackle displacement and affordability. The tale suggests that community control and social infrastructure might endure despite residential demand, making governance key to community longevity.

Sawtelle Japantown

Sawtelle Japantown demonstrates a different pattern of retention, where the historic residential and institutional base is reduced but the cultural character is kept. Historically, the district was supported by Japanese American homes, nurseries, churches, language schools, and community groups, but many of these businesses and institutions have dwindled or disappeared as redevelopment and demographic shift transformed the region. Contemporary Sawtelle is attracting a younger and more diversified clientele, but this is a different kind of rejuvenation than intergenerational community reproduction. Meanwhile, the Sawtelle Japantown Association has evolved into a grass-roots organization working to preserve and promote Japanese American heritage. The story consequently shows that cultural regeneration can take place without full continuity of residence or institution, and that the contrast between cultural visibility and long-term community presence is therefore of particular importance.

The comparative contribution is therefore not to identify a single model that Little Saigon should reproduce. Instead, it identifies which mechanisms address which dimensions of cultural persistence. The comparison suggests that cultural recognition and physical visibility primarily address C1 and C4; institutional and community organizations contribute to C2, C3, and L4; land-use and legacy-business interventions address E1, E2, and E5; financial and property-based mechanisms contribute to E2 and E3; while affordable housing and community-controlled development are particularly relevant to L1, L2, and L5. This mapping allows policy instruments to be matched to the specific vulnerability they are intended to address.

For Little Saigon, this comparison changes the interpretation of the existing Blueprint. Westminster already has substantial activity in the areas of cultural visibility, branding, streetscape improvement, events, and business assistance. The comparative framework shows that these interventions primarily operate within the Cultural Presence level and selected aspects of Economic Continuity. By contrast, the paper identifies less-developed mechanisms for

commercial-rent monitoring, succession outcomes, housing stability, long-term property control, and representative governance.

Recommendations

- Establish a baseline and public scorecard

Westminster should begin with a Little Saigon economic baseline and update it annually. The scorecard should report storefront vacancy, business openings and closures, firms receiving technical assistance, completed ownership or family-successions, average reported asking rent by subarea where data are sufficient, pedestrian counts, transit boardings, reported collisions, youth visit frequency, and the number of affordable homes preserved or created within the study area. The city should publish methods and geographic boundaries so that changes are comparable over time. This recommendation is feasible without an econometric model: it requires consistent descriptive measures and clear definitions.

- Create a bilingual succession and modernization program

The city and a trusted community partner should create a voluntary bilingual business-succession and modernization program. Participating firms would receive a confidential diagnostic covering ownership goals, financial records, valuation, lease transferability, licensing, legal needs, digital presence, and potential successors. Accountants, lawyers, lenders and digital-service providers should be vetted and ready to help. Small grants should be tied to specific milestones like completing a succession plan, adopting bookkeeping software, or launching an accessible ordering system as encouragement. The Blueprint already identifies these needs; the revision should convert them into a pipeline with annual targets and reported completion rates (City of Westminster, 2021).

- Monitor commercial stability and support legacy firms

Westminster should pair a voluntary commercial-conditions survey with a narrowly designed legacy-business pilot. Eligibility might require a minimum operating history, demonstrated cultural or community significance, good standing, and a credible continuation plan. A grant should produce a preservation commitment, such as a multi-year lease extension, a documented transfer to a successor, or completion of essential code and accessibility work. It is evident that recognition can be connected to a measurable economic benefit, and also how important rules for eligibility and budget caps are from Calle 24's experience and San Francisco's rent-stabilization grants (San Francisco Planning Department, 2017; City and County of San Francisco, 2023).

- Connect cultural preservation to housing stability

The city should treat housing stability in and around Little Saigon as economic infrastructure. Immediate actions could include multilingual outreach on tenants’rights, referrals for eviction prevention, preservation of mobile-home and other lower-cost housing, and an inventory of publicly owned or underused sites that could be used for mixed-income or affordable housing. Westminster should explore the legal tools available to the city for replacement housing, relocation assistance, density incentives and community benefits for larger redevelopment proposals. It can be seen from Little Tokyo's experience that long-term

affordability is more sustainable when a mission-driven organization is able to acquire or control land, hence, the city should consider partnering with an existing nonprofit or establishing a community development intermediary (Little Tokyo Service Center, n.d.).

- Run a measurable youth and customer-access pilot

Rather than build an expensive anchor immediately, Westminster should run a 12-month youth and customer-access pilot in one clearly defined corridor segment. The pilot could combine evening or weekend programming, youth-led pop-up businesses, small public seating areas, coordinated promotions, safer crossings, and wayfinding among participating merchants. Before launch, the city should record pedestrian counts, participating-store sales indices where businesses consent, visit frequency, and user perceptions; it should repeat the measures at regular intervals. The 147-person survey of the Blueprint's suggests hypotheses about access and tenant mix but a pilot would test which interventions actually change behavior (City of Westminster, 2021).

- Improve circulation before promising new transit service

Transportation recommendations should be based on existing travel data and coordination with the Orange County Transportation Authority. Low-cost actions include more transparent bus information, shaded and accessible stops, safer crossings, pedestrian connections between shopping centers, bicycle parking and event-day circulation management. If origin-destination and ridership data show a demand, a limited-stop bus might be studied. The 2016 Route 64X operated by OCTA that reduced stops from 48 to 16 and claimed savings of up to 19 minutes provides a historical example of how limited-stop service can improve travel time, but it is not evidence that the same service exists now or would succeed under current conditions (Orange County Transportation Authority, 2016).

- Build representative governance and phase spending

Implementation should be overseen by a Little Saigon partnership with seats for small-business tenants, property owners, residents, youth, workers, cultural organizations, and the city. The partnership should publish meeting summaries, annual budgets, project responsibilities, and scorecard results. Spending should be phased: first establish baselines and low-cost technical assistance; next test access, public-space, and legacy-business pilots; then scale capital projects that demonstrate demand and maintenance capacity. The cost categories of the Blueprint are a good starting point, but each project should also identify its funding source, operating cost, accountable entity, timeline, and measurable outcome (City of Westminster, 2021).

Conclusion

Little Saigon became a cultural hub thanks to Vietnamese refugees and immigrants building a small-scale economic system, going from households settling next to one another, entrepreneurs creating specialized firms, customers travelling to a recognizable marketplace, and institutions reinforcing trust and belonging. This further shows that cultural preservation cannot be achieved solely from symbols but also from economics. If firms cannot transfer to new owners, workers and renters cannot remain nearby, or visitors cannot move safely among

businesses, the district may retain its name while losing the relationships that give the name meaning.

Westminster's Blueprint and current streetscape projects are important as they acknowledge the need for a coordinated effort, public funding, and a stronger district identity (City of Westminster, 2021; City of Westminster, 2026a). However, the present economic data strongly suggests the need for a broader agenda. Little Saigon is a high-volume regional marketplace mostly created from small establishments located in the lower-payroll and highly rent-burdened area (Puri, Popp, & Fleissig, 2024). Physical improvement can increase demand, but without succession assistance, commercial monitoring, housing stability, and representative governance, reinvestment may not support the businesses and the people it is intended to help.

The most defensible path is therefore managed adaptation. Little Saigon should welcome new businesses, younger customers, safer streets, and contemporary forms of cultural expression while protecting access to the district's economic opportunities. A baseline scorecard, bilingual business support, targeted legacy-business assistance, housing partnerships, measured youth pilots, circulation improvements, and transparent governance would allow Westminster to test progress without relying on complex models. The district's future will be strongest when economic change becomes a means of cultural continuity rather than a force that separates culture from the community that created it.

Works Cited

Aguilar-San Juan, K. (2009). *Little Saigon: Staying Vietnamese in America*. University of Minnesota Press.

Alvarez, N., Andrews, B., & Lung-Amam, W. (2026). Keeping BIPOC- and immigrant-owned businesses in place: Policies to combat commercial gentrification and small business displacement. Journal of the American Planning Association. https://doi.org/10.1177/08854122251323897

Berg, T. (2015, April 30). Why Westminster? Eleven reasons the Vietnamese came to Little Saigon—and why they stayed. *Orange County Register*. https://www.ocregister.com/2015/04/30/why-westminster-eleven-reasons-the-vietnamese-came-to-little-saigon-and-why-they-stayed/

Choi, M., Van Zandt, S., & Matarrita-Cascante, D. (2018). Can community land trusts slow gentrification? Journal of Urban Affairs, 40(3), 394–411. https://doi.org/10.1080/07352166.2017.1362318

Choy, C. C. (2017). New Asian American communities: Building and dismantling. In F. Odo (Ed.), *Finding a path forward: Asian American and Pacific Islander National Historic Landmarks theme study* (pp. 307–324). National Park Service. https://www.nps.gov/articles/upload/16-Essay-16-New-Asian-American-Communities.pdf

City and County of San Francisco, Office of Small Business. (2023). *Annual report on the Legacy Business Historic Preservation Fund, fiscal year 2022–23*. https://www.sf.gov/sites/default/files/2023-06/LBP%20Grants%20Annual%20Report%202022-23.pdf

City of Los Angeles, Office of the City Clerk. (2015). *Council file 14-1686: Sawtelle Japantown/community naming or renaming*. https://cityclerk.lacity.org/lacityclerkconnect/index.cfm?cfnumber=14-1686&fa=ccfi.viewrecord

City of Westminster. (2026a, April 2). *City of Westminster selects Los Angeles-based HKLA to launch Little Saigon Streetscape Feasibility Study*. https://www.westminster-ca.gov/Home/Components/News/News/1047/

City of Westminster. (2026b). *Request for proposal: Little Saigon brand identity, environmental integration, and marketing strategy*. https://www.westminster-ca.gov/home/showpublisheddocument/10154/639075369853430000

Disability Rights California. (2020, January 30). *Residents of Green Lantern Mobile Home Parks and Village get to stay in their homes*. https://www.disabilityrightsca.org/press-release/residents-of-green-lantern-mobile-home-parks-and-village-get-to-stay-in-their-homes

Dang, V. T. (2013). *Little Saigon and Vietnamese American Communities*. In X. Zhao & E. J. W. Park (Eds.), *Asian Americans: An Encyclopedia of Social, Cultural, Economic, and Political History*. ABC-CLIO.

Do, A. (2020, February 1). Elderly residents fighting closure of Little Saigon mobile home park notch a victory. *Los Angeles Times*. https://www.latimes.com/california/story/2020-02-01/little-saigon-westminster-mobile-home-park-closure-reversed

Drueding, M. (2024, Summer). L.A.'s Little Tokyo faces threats from development and rising rents. *Preservation*. https://savingplaces.org/stories/las-little-tokyo-faces-threats-from-development-and-rising-rents

Gandhi, A., & Minner, J. (2017). Economic development challenges for immigrant retail corridors: Observations from Chicago's Devon Avenue. Journal of Planning Education and Research. https://doi.org/10.1177/0891242417730401

Ha, N. Q. (2002). *Business and politics in Little Saigon, California* [Undergraduate thesis, Rice University]. https://bos.ocgov.com/D1PDF/loa_090424/A21%20Promoting%20Little%20Saigon/OTH%20-%20Business%20and%20politics%20in%20Little%20Saigon%2C%20California.pdf

Hackworth, J., & Rekers, J. (2005). Ethnic packaging and gentrification: The case of four neighborhoods in Toronto. Urban Affairs Review, 41(2), 211–236.

Hirahara, N. (2015, April 15). Thinking L.A.: How West L.A. became a haven for Japanese-Americans. *UCLA Newsroom*. https://newsroom.ucla.edu/stories/thinking-l-a-how-west-l-a-became-a-haven-for-japanese-americans

Kloosterman, R., van der Leun, J., & Rath, J. (2002). Mixed embeddedness: (In)formal economic activities and immigrant businesses in the Netherlands. *International Journal of Urban and Regional Research, 23*(2), 252–266. https://doi.org/10.1111/1468-2427.00194

Lee, D. (1997, August 5). One man's vision for Little Saigon. *Los Angeles Times*. https://www.latimes.com/archives/la-xpm-1997-aug-05-mn-19616-story.html

Lieu, N. T. (2011). Vietnamese by other means: The overlapping diasporas of Little Saigon. In *The American dream in Vietnamese America*. University of Minnesota Press.

Little Tokyo Service Center. (n.d.). *First Street North Residences: Application resources*. Retrieved September 21, 2026, from https://www.ltsc.org/fsn_residences/application-resources/

Los Angeles City Planning. (2020). *Little Tokyo Community Design Overlay: Design standards and guidelines*. https://planning.lacity.gov/odocument/d8182b92-4603-42ce-a5d2-425dc7960fe6/LittleTokyoGuidlines_amendments_fal2020.pdf

Marcuse, P. (1985). Gentrification, abandonment, and displacement: Connections, causes, and policy responses in New York City. Washington University Journal of Urban and Contemporary Law, 28, 195–240.

Mirabal, N. R. (2009). Geographies of displacement: Latina/os, oral history, and the politics of gentrification in San Francisco's Mission District. The Public Historian, 31(2), 7–31. https://doi.org/10.1525/tph.2009.31.2.7

Mazumdar, S., Mazumdar, S., Docuyanan, F., & McLaughlin, C. M. (2000). *Creating a sense of place: The Vietnamese-Americans and Little Saigon*. Journal of Environmental Psychology, 20(4), 319–333.

Newman, K., & Wyly, E. K. (2006). The right to stay put, revisited: Gentrification and resistance to displacement in New York City. Urban Studies, 43(1), 23–57. https://doi.org/10.1080/00420980500388710

Orange County Transportation Authority. (2016, February 12). *Limited-stop bus service comes to popular route through Little Saigon with launch of 64Xpress*. https://www.octa.net/news/news-releases/limited-stop-bus-service-comes-to-popular-route-through-little-saigon-with-launch-of-64xpress

Puri, A. K., Popp, A., & Fleissig, A. (2024). *Little Saigon, Orange County: A demographic and economic profile*. Woods Center for Economic Analysis and Forecasting, California State University, Fullerton. https://business.fullerton.edu/engagement/economic-analysis-and-forecasting/assets/pdf/little-saigon-profile.pdf

Rico, B., Hahn, J. K., & Spence, C. (2023, September 21). Chinese, except Taiwanese, was the largest Asian alone or in any combination group; Nepalese population grew fastest. *U.S. Census Bureau*. https://www.census.gov/library/stories/2023/09/2020-census-dhc-a-asian-population.html

San Francisco Planning Department. (2020). *Calle 24 Special Area Design Guidelines*. https://sfplanning.org/sites/default/files/documents/citywide/calle-24-special-area-design-guidelines/calle_24_design_guidelines.pdf

Sawtelle Japantown Association. (n.d.). *Sawtelle Japantown history*. https://sawtelleja.org/sawtelle-japantown-history/

Strategic Economics. (2016). *Calle 24 retail study: Final report*. San Francisco Office of Economic and Workforce Development. https://oewd.org/sites/default/files/Invest%20In%20Neighborhoods/Calle24RetailStudyReport_Final_Jan2017.pdf

Streetsense. (2021). *Little Saigon: Blueprint for investment*. City of Westminster. https://www.westminster-ca.gov/home/showpublisheddocument/4248/637716165086830000

Throsby, D. (2014). The role of culture in sustainable urban development: Some economic issues. International Journal of Global Environmental Issues, 13(2–4), 89–99.

U.S. Environmental Protection Agency. (2012). *Smart growth and economic success: Benefits for real estate developers, investors, businesses, and local governments*. https://www.epa.gov/sites/production/files/documents/smart_growth_and_economic_success.pdf

Vo, L. T. (2008). Constructing a Vietnamese American Community: Economic and Political Transformation in Little Saigon, Orange County. Amerasia Journal, 34(3), 84–109. https://doi.org/10.17953/amer.34.3.82340l2rmn866108