AJUR

American Journal of *Undergraduate Research*

Volume 22 | Issue 2 | June 2025

www.ajuronline.org

Print Edition ISSN 1536-4585
Online Edition ISSN 2375-8732

AJUR
American Journal of
Undergraduate Research

Volume 22 | Issue 2 | June 2025 | *https://doi.org/10.33697/ajur.2025.136*

2 **AJUR History and Editorial Board**

3 **Understanding the Molecular Mechanisms that Govern Opioid Potency Through a Course-embedded Computational Research Project**
Keegan Gunderson, Faith Oldenburg, Sudeep Bhattacharyay, & Sanchita Hati

13 **Isolation and Characterization of Novel Marine Bacteriophages from Santa Rosa Island, Florida, Using Local and Nonlocal Bacterial Strains as Hosts**
Brittany C. Yencho, Charles J. West, Andrew J. Brown, & Conor R. Flannigan

27 **Uninformed Consent? The Impact of Reading Level, Format, and Interactivity of Consent Forms on Participant Comprehension**
Kaitlyn A. Carr & Clare Conry-Murray

35 **A Finite Difference Approach and Its Error Estimate to Two-Dimensional Poisson Equation for Dirichlet Boundary Conditions**
Matthew Whalen

49 **A Comparison of Zero-Inflated Models for Modern Biomedical Data**
Max Beveridge, Zach Goldstein, & Hee Cheol Chung

69 **Amplifying Disparities? The Inequitable Burden of Transportation Noise in Rural and Minority Communities**
Erin Koster & Michelle Stuhlmacher

81 **Utilizing a Large Language Model for Training Students in Personal Care Product Formulation**
McKinnley Bilbao, Caitlin West, Tomas Carmona, Morgan Covarrubia, Alex Goslin, Katherine Judge, Garland Munn, Hazel Ticas, Abe Tonioli, Collin Tuttle, & Daniel Scott

American Journal of Undergraduate Research (AJUR) *is a national, independent, peer-reviewed, open-source, quarterly, multidisciplinary student research journal. Each manuscript of AJUR receives a DOI number. AJUR is archived by the US Library of Congress. AJUR was established in 2002, incorporated as a cheritable not-for-profit organization in 2018. AJUR is indexed internationally by EBSCO and Crossref with ISSNs of 1536-4585 (print) and 2375-8732 (web).*

EDITORIAL TEAM

Dr. Peter Newell, Editor-in-Chief
Dr. Kestutis Bendinskas, Executive Editor
Dr. Anthony Contento, Copy Editor

EDITORIAL BOARD *by subject area*

ACCOUNTING
Dr. Dean Crawford,
dean.crawford@oswego.edu

ART HISTORY
Dr. Lisa Seppi,
lisa.seppi@oswego.edu

BEHAVIORAL NEUROSCIENCE
Dr. Aileen M. Bailey,
ambailey@smcm.edu

BIOCHEMISTRY
Dr. Kestutis Bendinskas,
kestutis.bendinskas@oswego.edu
Dr. Nin Dingra,
ndingra@alaska.edu

BIOENGINEERING
Dr. Jorge I. Rodriguez,
jorger@uga.edu
Dr. Jessica Amber Jennings,
jjnnings@memphis.edu

BIOINFORMATICS
Dr. John R. Jungck,
jungck@udel.edu
Dr. Isabelle Bichindaritz,
ibichind@oswego.edu

BIOLOGY, PHYSIOLOGY
Dr. David Dunn,
david.dunn@oswego.edu

BIOLOGY, DEVELOPMENTAL
Dr. Poongodi Geetha-Loganathan,
p.geethaloganathan@oswego.edu

BIOLOGY, MICROBIOLOGY
Dr. Peter Newell,
peter.newell@oswego.edu

BOTANY
Dr. Julien Bachelier,
julien.bachelier@fu-berlin.de

CHEMISTRY
Dr. Alfredo Castro,
castroa@felician.edu
Dr. Charles Kriley,
cekriley@gcc.edu
Dr. Vadoud Niri,
vadoud.niri@oswego.edu

COMPUTER SCIENCES
Dr. Dele Oluwade,
deleoluwade@yahoo.com
Dr. Mais W Nijim,
Mais.Nijim@tamuk.edu
Dr. Bastian Tenbergen,
bastian.tenbergen@oswego.edu

COMPUTATIONAL CHEMISTRY
Dr. Alexander Soudackov,
alexander.soudackov@yale.edu

ECOLOGY
Dr. Chloe Lash,
cllash@ollusa.edu

ECONOMICS
Dr. Elizabeth Schmitt,
elizabeth.schmitt@oswego.edu

EDUCATION
Dr. Charity Dacey,
cdacey@touro.edu
Dr. Marcia Burrell,
marcia.burrell@oswego.edu

EDUCATION, PHYSICS
Dr. Andrew D. Gavrin,
agavrin@iupui.edu

ENGINEERING, ELECTRICAL
Dr. Michael Omidiora,
momidior@bridgeport.edu

ENGINEERING, ENVIRONMENTAL
Dr. Félix L Santiago-Collazo,
fsantiago@uga.edu

FILM AND MEDIA STUDIES
Dr. Lauren Steimer,
lsteimer@mailbox.sc.edu
Dr. Ashely Young,
AY13@mailbox.sc.edu

GEOLOGY
Dr. Rachel Lee,
rachel.lee@oswego.edu

GEOGRAPHY &GIS
Dr. Jamie Spinney,
jamiespinney@suu.edu

HISTORY
Dr. Richard Weyhing,
richard.weyhing@oswego.edu
Dr. Murat Yasar,
murat.yasar@oswego.edu

HONORARY EDITORIAL BOARD MEMBER
Dr. Lorrie Clemo,
lorrie.a.clemo@gmail.com

JURISPRUDENCE
Bill Wickard, Esq,
William.Wickard@KLGates.com

KINESIOLOGY
Dr. David Senchina,
david.senchina@drake.edu

LINGUISTICS
Dr. Taylor Miller,
taylor.miller@oswego.edu

LITERARY STUDIES
Dr. Melissa Ames,
mames@eiu.edu
Dr. Douglas Guerra,
douglas.guerra@oswego.edu

MATHEMATICS
Dr. Dele Oluwade,
deleoluwade@yahoo.com
Dr. Christopher Baltus,
christopher.baltus@oswego.edu
Dr. Mark Baker,
mark.baker@oswego.edu
Dr. Monday Nnakwe,
mondaynnakwe@gmail.com (Auburn University)
Dr. Roza Aceska,
raceska@bsu.edu (Ball State University)

MEDICAL SCIENCES
Dr. Thomas Mahl,
Thomas.Mahl@va.gov
Dr. Jessica Amber Jennings,
jjnnings@memphis.edu

METEOROLOGY
Dr. Steven Skubis,
steven.skubis@oswego.edu

NANOSCIENCE AND CHEMISTRY
Dr. Gary Baker,
bakergar@missouri.edu

PHYSICS
Dr. Priyanka Rupasinghe,
priyanka.rupasinghe@oswego.edu

POLITICAL SCIENCE
Dr. Kaden Paulson-Smith,
Paulsonk@uwgb.edu

PSYCHOLOGY
Dr. Matthew Dykas,
matt.dykas@oswego.edu
Dr. Damian Kelty-Stephen,
keltystd@newpaltz.edu
Dr. Kenneth Barideaux Jr.,
kbaridea@uscupstate.edu

SOCIAL SCIENCES
Dr. Rena Zito,
rzito@elon.edu
Dr. Dana Atwood,
atwoodd@uwgb.edu

STATISTICS
Dr. Mark Baker,
mark.baker@oswego.edu

TECHNOLOGY, ENGINEERING
Dr. Reg Pecen,
regpecen@shsu.edu

ZOOLOGY
Dr. Chloe Lash,
CLash@stfrancis.edu

Understanding the Molecular Mechanisms that Govern Opioid Potency Through a Course-embedded Computational Research Project

Keegan Gunderson, Faith Oldenburg, Sudeep Bhattacharyay, & Sanchita Hati**

Department of Chemistry and Biochemistry, University of Wisconsin - Eau Claire, Eau Claire, WI

https://doi.org/10.33697/ajur.2025.137

Students: gunderkr8370@uwec.edu, oldenbfa6097@uwec.edu*
*Mentors: bhattas@uwec.edu, hatis@uwec.edu**

ABSTRACT
In recent decades awareness surrounding the class of drugs known as opioids has risen due to what many have termed the "opioid epidemic." Rates in opioid-related drug overdoses have spiked due to increased opioid addiction and the illegal lacing of illicit drugs, such as marijuana, with opioid compounds like fentanyl, increasing their potency. Coexisting with the devastating realities of overdose and addiction, however, is the clinical demand for these drugs and their potency to manage extreme pain. Herein, we describe the results of a course-embedded computational research project to understand the factors responsible for opioid potency. In the present study, the chemical properties of commonly used opioids and their interactions with receptors are investigated using computational techniques. In particular, the molecular basis for the high potency of fentanyl is investigated. The drug molecules were constructed using a web-based tool, WebMO. The electronegativity and chemical hardness of nine opioids and one pain reliever were determined using a quantum chemistry software package named Q-Chem. The binding interactions between the same set of opioid molecules and their receptors were studied using AutoDock FR, interactions within the receptor's binding pocket were analyzed using Visual Molecular Dynamics (VMD), and SwissADME was used to investigate the pharmacokinetic properties of each drug molecule.

KEYWORDS
Opioid Addiction, Opioid-Receptor Interactions, Opioid-Receptor Binding Affinity, Electronegativity and Chemical Hardness, Course-embedded Research

INTRODUCTION

The pharmaceuticals known as opioids are a class of analgesic, or pain-relieving drugs that function as agonists of human opioid receptor (OR) proteins. Opioids are powerful pain relievers that are commonly used in medical settings, especially in acute care contexts. Drugs that fall beneath the opioid umbrella include codeine and morphine, which are naturally derived from opium, as well as semi-synthetically and synthetically derived drugs such as fentanyl, oxycodone, methadone, and illegal heroin.

While these drugs are necessary for pain management for patients who have undergone surgery, physical trauma, or certain chemotherapies, they have health risks. The mechanism of action of opioids triggers the same chemical response as the human body's dopamine reward system, causing them to be highly addictive for some individuals, especially when used to treat chronic pain.[1] Opioids function by binding to and stimulating G-protein coupled receptors (GPCRs) called ORs, found on the plasma membranes of cells in the human central nervous system and gastrointestinal tract.[2] They are attracted to and bound to their protein receptors via noncovalent interactions. There are four primary ORs, which belong to the family of cell membrane proteins GPCRs - mu (μ OR), delta (δ OR), kappa (k OR), and nociceptin/orphanin FQ (N/OFQ) receptors (**Figure 1**).[3,4] Once the surface proteins of the GPCR complexes are activated, they trigger a cascade of interactions within the cell, releasing endorphins, neurotransmitters involved with the mesolimbic system. Endorphins reduce feelings of pain by slowing down the release of proteins involved in pain signaling. These endorphins also catalyze the release of dopamine, which plays a vital role in the perception of pain and provides natural analgesia to various portions of the brain and body.[5] Because opioids activate the mesolimbic system, prolonged usage commonly results in the development of opioid drug tolerance and leads to a high probability that those who develop this tolerance will display signs of addiction and subsequent withdrawal when coming off the medication. In the last decade, opioid addiction has surged at an exponential rate. Recent data shows that 81,806 Americans died from opioid overdose in 2021.[6] Thus, there is rising concern within the medical and broader community surrounding their uses, their mechanism of action, and what precautions or treatment alternatives can be implemented to fight back against the destruction of opioid addiction.

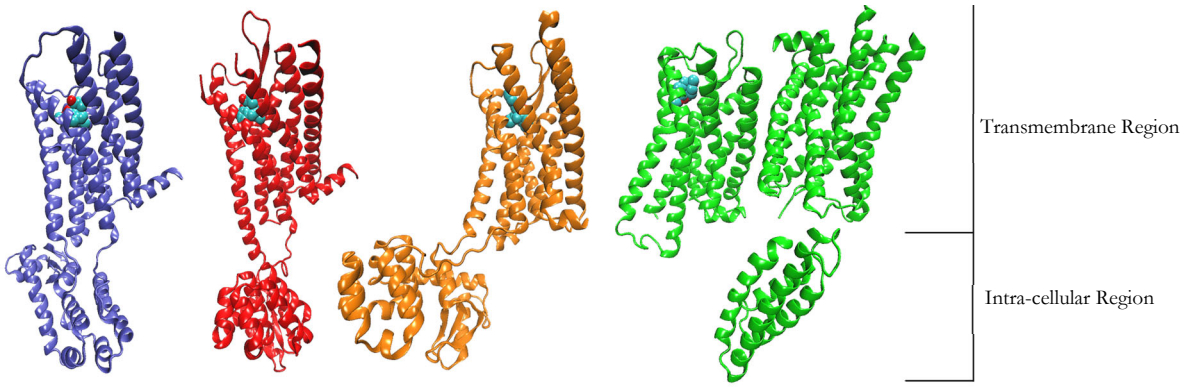


Figure 1. A visualization of the protein structure and binding pocket for each of the four opioid receptors.

The present study was conducted in a General Chemistry I Lab to provide students with hands-on experience with computation early in their college careers. It aimed to examine the molecular characteristics of nine opioids and the pain reliever imitrex and assess their potency based on their chemical and pharmacological properties. This involved computationally constructing the 3D molecular structures of the ten molecules and using these models to determine their electronegativity and chemical hardness values, which are key properties affecting chemical reactivity. It also aimed to identify basic chemical characteristics such as lipophilicity, gastrointestinal (GI) absorption, and blood-brain barrier permeability. Additionally, it determined the binding affinity values and analyzed the 3-dimensional structures of opioid-receptor complexes for all ten compounds (nine opioids and imitrex) and the four opioid receptors. Overall, the goal of this study was to identify molecular determinants of opioid potency, with a specific focus on fentanyl. This was achieved by analyzing the chemical properties of various opioids and their binding interactions with opioid receptors. These findings could be used in designing effective analgesics that pose fewer health risks, such as reduced potential for addiction or fatal overdose.

METHODS

The chemical and structural formulas of each opioid were obtained from PubChem.[7] Molecular structures were built using WebMO[8] and molecular geometry optimization was performed using QChem.[9] For geometry optimization, the Hartree-Fock method was used with the correlation-consistent polarized basis set with double zeta basis function.[10] This is abbreviated as cc-pVDZ. The convergence criteria for this optimization were 10^{-8} Hartree. Once the molecular optimization was completed, the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies for each opioid were collected. The HOMO and LUMO energies were used to determine the electronegativity (χ) and chemical hardness (η) of each molecule using equations 1 and 2 below:[11]

$$\chi = \frac{(IE + EA)}{2} = -\frac{1}{2}(\epsilon_{HOMO} + \epsilon_{LUMO}) \quad (1)$$

$$\eta = \frac{(IE - EA)}{2} = -\frac{1}{2}(\epsilon_{HOMO} - \epsilon_{LUMO}) \quad (2)$$

Electronegativity and chemical hardness are important chemical properties that characterize a molecule. Electronegativity (χ) is the ability of an atom or a group of atoms to attract the bonding pair of electrons towards itself.[12] The higher the electronegativity of an atom, the stronger its attraction for the bonding pair of electrons. On the other hand, chemical hardness (η) is defined as a measure of the tendency of a chemical species to localize charge density.[13] It is a qualitative indication of the polarizability of chemical species (atom, ion, or molecule). Chemically hard molecules are less polarizable, and chemically soft ones are more polarizable. The hardness η is an important chemical property that provides information about molecular reactivity and selectivity.[14-17]

To quantify the interaction between opioids and opioid receptors, the binding affinity was assessed for each opioid-receptor combination. To do this, the structural information of the opioid receptors was obtained from the Protein Data Bank.[18] The spatial data file (sdf) of the opioids was also obtained from PubChem.[7] These files contain information necessary to determine the size, 3-dimensional shape, and molecular characteristics of drug molecules and bound proteins. The molecular docking was performed using AutoDock for flexible receptors (ADFR) software.[19] ADFR uses a genetic algorithm-based approach and is validated using a protein ligand docking performance set called Astex Diverse Set.[20] This algorithm predicts the favored orientation of the ligand (an opioid) in the opioid receptor's binding pocket using the AutoDock 4 scoring function to estimate binding affinity.[21] The AutoDock 4 scoring function uses a semi-empirical force field to model non-bonding interactions between the opioid and receptor in various conformations. These models are quantified by finding the difference in energy before and

after ligand docking for Van der Waals forces, hydrogen bonding, electrostatic forces, and desolvation interactions. The binding affinity is calculated from the thermodynamic quantities of these interactions.[22] For each docking simulation, the genetic algorithm was run 50 times. In these simulations, the grid box size was calculated from the dimensions of the opioid receptor with a padding of 4 angstroms.

Additionally, interactions between opioids and receptors were investigated for fentanyl and meperidine bound to the μ and κ ORs. These two opioids were used because they displayed the highest and lowest binding affinities for the two opioid receptors. The μ OR was analyzed due to its contribution to opioid overdose deaths and the κ OR was used because it displayed the highest overall binding affinity when bound to fentanyl. This was done by visually analyzing and measuring the distance of significant polar and hydrophobic interactions between the opioid molecule and the receptor backbone using Visual Molecular Dynamics (VMD) software.[23] The ligand-bound binding pockets were manipulated to show their atom types using the "CPK," or ball-and-stick representation for the protein backbone, and the "licorice" representation for the opioid molecule. The alpha-carbon backbone surrounding the binding pocket was shown using the "NewCartoon" representation to help better visualize the location of the interactions relative to the protein. Layering several unique representations of the protein-ligand system allowed us to visualize noncovalent interactions between receptors and opioids within the binding pocket. These observations provided insight into the cause of variable binding affinity in each docking simulation.

Finally, the chemical characteristics, namely absorption, distribution, metabolism, elimination, and toxicity, or ADMET, of each opioid were obtained using SwissADME,[24] where the input files were in the simplified molecular-input line-entry system (SMILES) format.

RESULTS

Electronegativity and Chemical Hardness

First, the chemical properties of the nine opioids and Imitrex were compared. The χ and η values of these ten molecules range between 2.10 – 3.06 eV and 5.12 – 6.61 eV, respectively (**Figure 2, Table S1**). There is a notable spike in electronegativity and chemical hardness values for fentanyl, which is the most potent synthetic opioid, 50 to 100 times more potent than morphine.[25]

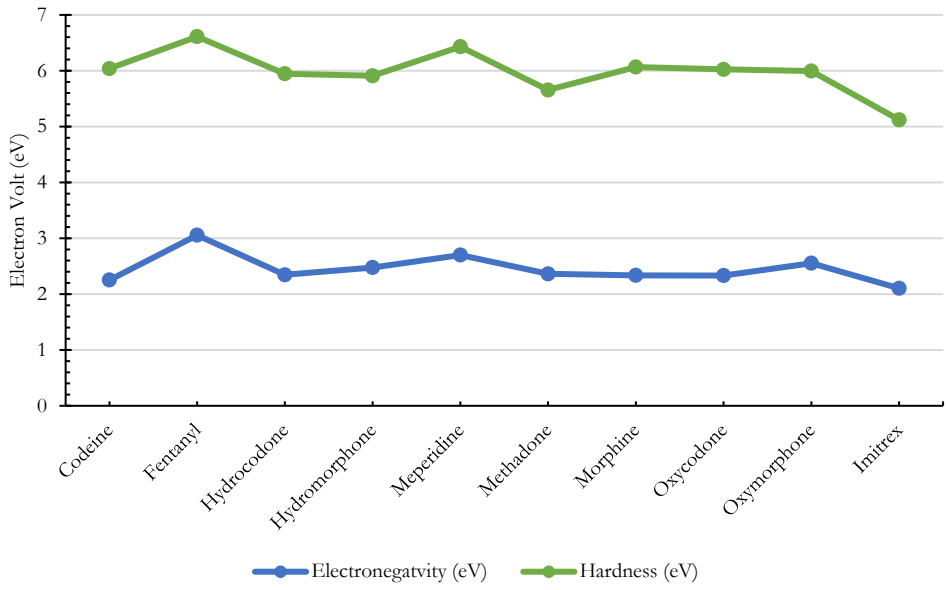


Figure 2. Electronegativity and chemical hardness, in electron volt (eV), of all nine opioids and Imitrex.

Docking and Binding Energy

The binding affinity for nine opioids and Imitrex, docked in the receptors binding pocket, were determined and shown in **Figure 3**. It is important to note that all binding affinity values are negative. Fentanyl showed the highest binding affinity for μ, κ, and δ ORs with values of -8.7 kcal/mol, -9.7 kcal/mol, and -9.2 kcal/mol, respectively. The N/OFQ OR had the best binding energy of -8.2 kcal/mol for methadone. The binding affinity was weakest (-6.4 kcal/mol, less tightly bound) for meperidine when bound to the μ OR. When comparing the binding affinities for different receptor proteins, the κ receptor exhibited tighter binding with a -

9.7 kcal/mol binding affinity for fentanyl. The κ OR exhibited an average binding affinity value of -8.12 kcal/mol while considering all 10 molecules. Also, fentanyl exhibited the highest binding affinity for three of the four receptors. However, the range for all the nine opioids for all the receptors was within 2.2 kcal/mol, suggesting all opioids are efficacious in stimulating the four primary opioid receptors in the body.

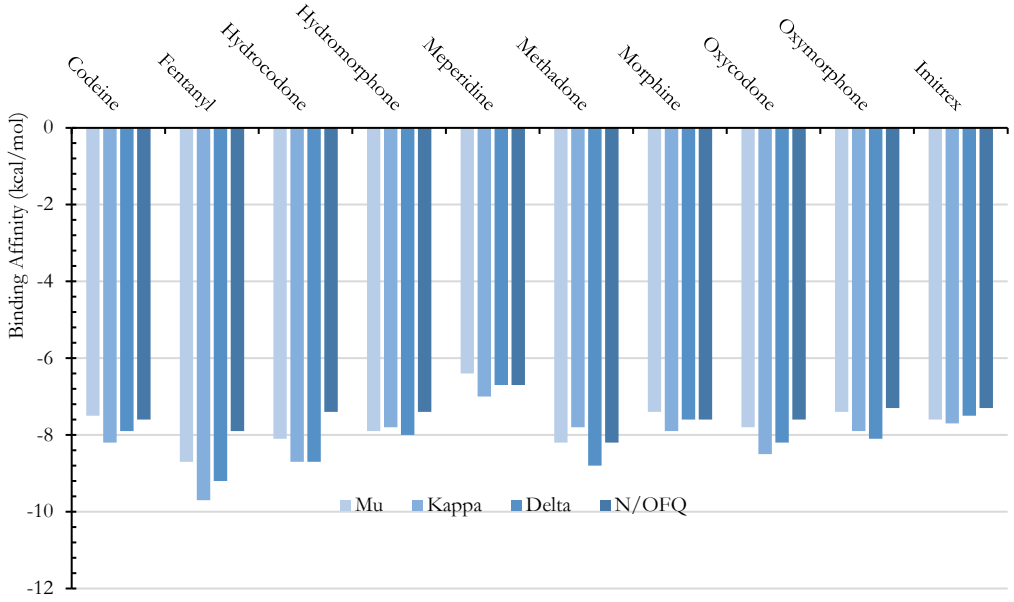


Figure 3. Computationally determined binding affinities for each of the four receptors, individually docked with all nine opioids and Imitrex.

Binding Pocket Analysis

To determine what kind of interactions were responsible for varying binding affinities among opioid-receptor pairs, the two extremes in binding affinity for κ and μ receptors, fentanyl and meperidine, were visualized using VMD. When these two molecules were bound to the κ OR (**Figure 4**), it was observed that their orientation in the pocket varied significantly, with fentanyl interacting more with the protein backbone (**Figure 4B, 4D**). Additionally, the two nonpolar phenyl groups of fentanyl (**Table S2**) appeared to have Van der Waals interactions with receptors that resulted in high binding affinity. Most notably, TRP287, ILE290, ILE294, and ILE316 in the binding pocket interact with one of fentanyl's two phenyl groups. Hydrogen bonding was also observed with residue ASP138 and fentanyl's carbonyl oxygen (**Figure 5A**). Hydrophobic interactions also could be observed for meperidine. The amino acid residues ILE294 and VAL230 interacted with the opioid's lone phenyl group. Weak interaction was also observed for residue HIS291 (**Figure 5B**). Hydrogen bonding was also observed between TYR139 and the oxygen group on meperidine. In the binding pocket of κ OR, hydrophobic interactions appear to be significant for both binding studies. However, polar interactions could also be observed.

When fentanyl and meperidine were bound to the μ OR (**Figure 6**), their orientation in the binding pocket appeared relatively similar (**Figure 6B, 6D**). However, fentanyl appeared to have more interactions. Again, fentanyl's two phenyl groups exhibited hydrophobic interactions with the protein backbone. This primarily occurred at amino acid residues VAL236, VAL300, TRP293, and TYR326. Also, a hydrogen bond was formed with fentanyl's carbonyl oxygen (**Table S2**) and ASP147 (**Figure 7A**). Similar interactions were observed to a lesser extent for meperidine. Namely, hydrophobic interactions occurred with amino acid residues VAL300 and TRP293. Also, a hydrogen bond interaction occurred between the amino acid residue ASP147 and the oxygen group in meperidine. A hydrogen bond also occurred between TYR326 and the nitrogen in meperidine (**Table S2**, **Figure 7B**).
For all four combinations, i.e. interactions of fentanyl and meperidine with κ and μ ORs, it is important to note that the noncovalent interactions between the receptors and opioids described here are those that were more significant to the opioid's binding; other amino acid residues also appeared to interact with the opioids (**Figures 6 and 7)**.

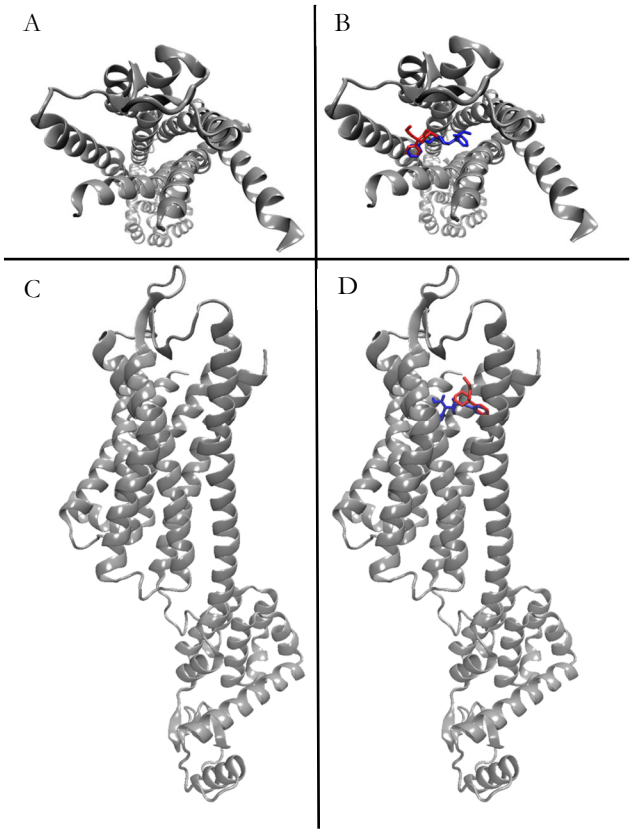


Figure 4. A) The top view of k-OR. B) k-OR displaying a bound meperidine (red) and fentanyl (blue). C) The front view of k-OR. D) k-OR front view displaying a bound meperidine (red) and fentanyl (blue).

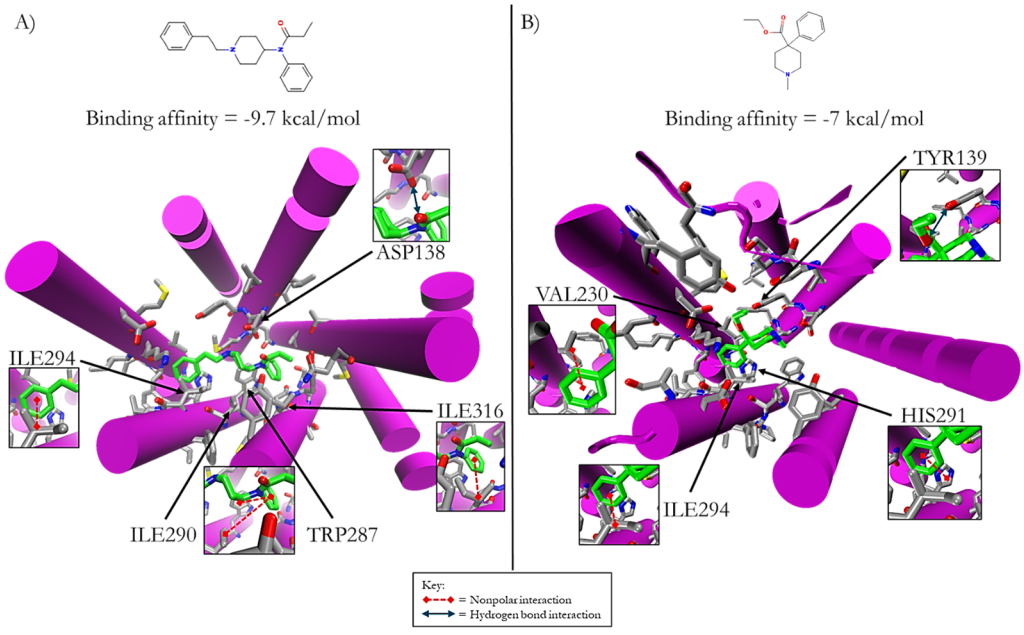


Figure 5. A) κ-OR bound to fentanyl. B) κ-OR bound to meperidine. Non-bonding interactions are depicted by the key and shown in the zoomed subpanels.

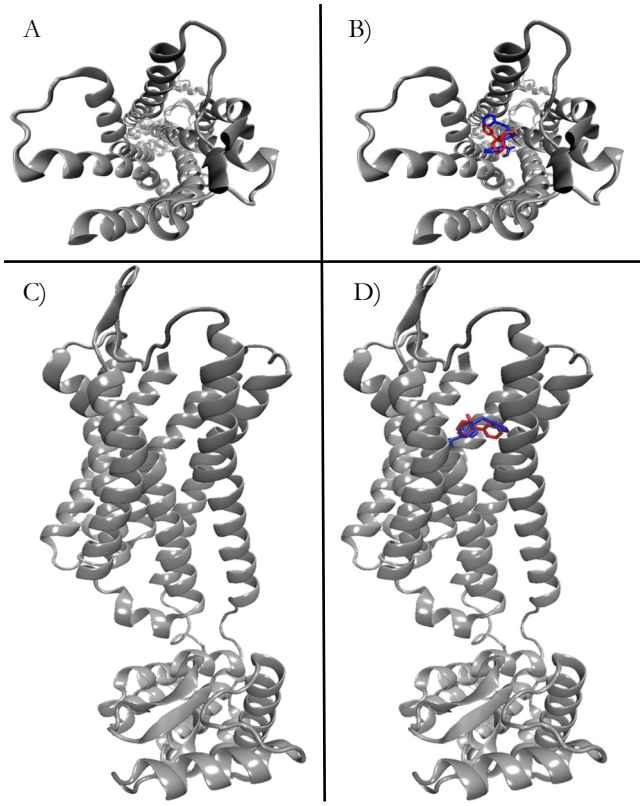


Figure 6. A) Top view of μ-OR. B) μ-OR displaying a bound meperidine (red) and fentanyl (blue). C) The front view of μ-OR. D) μ-OR front view displaying a bound meperidine (red) and fentanyl (blue) molecules.

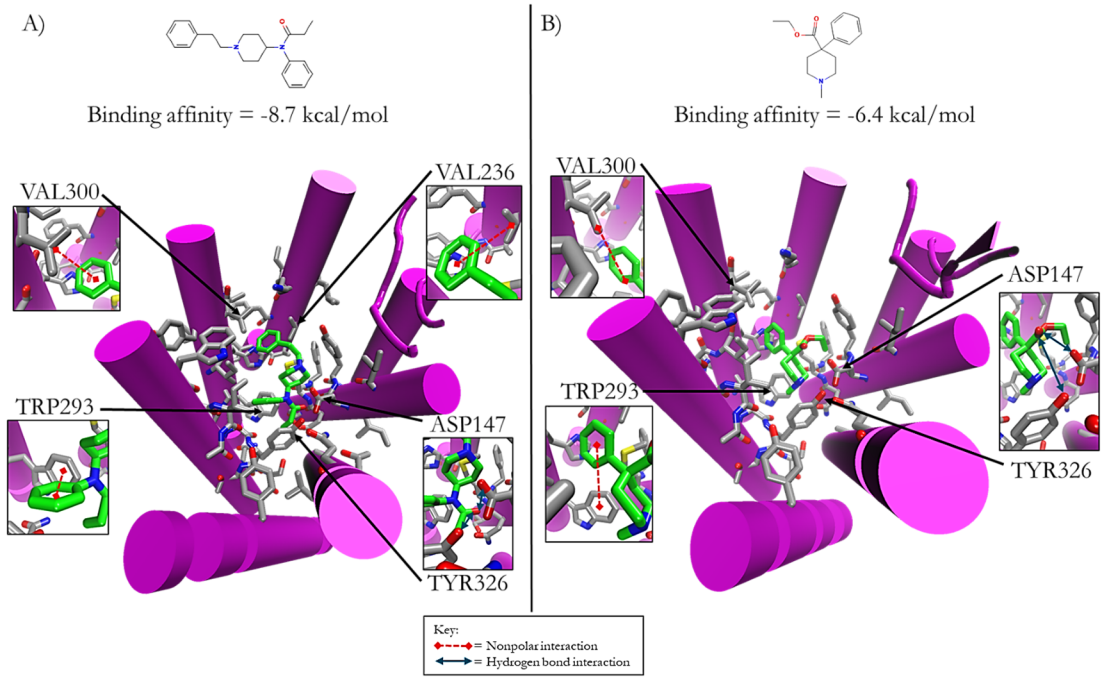


Figure 7. A) μ-OR bound to fentanyl. B) μ-OR bound to meperidine. Non-bonding interactions are depicted by the key and shown in the zoomed subpanels.

Absorption, Distribution, Metabolism, Elimination, and Toxicity (ADMET)

The ADMET values indicated that all tested drug molecules, apart from Imitrex, exhibited prominent levels of gastrointestinal absorption. It also demonstrated that all except three of the opioid molecules, Oxycodone, Oxymorphone, and the painkiller Imitrex, could permeate the body's blood-brain barrier complex, which is a vital function required for any pharmaceutical to be able to produce effects within the central nervous system tissues. These results are all shown in **Table 1.** One property of particular interest is lipophilicity which describes a chemical's ability to dissolve when placed into a hydrophobic solution. Highly lipophilic substances take longer to exit the body as they take longer to metabolize. They also more easily permeate membranes found in the body. Lipophilic opioids are effective because they can penetrate the blood-brain barrier faster and more effectively, leading to faster analgesic onset. They are also more readily absorbed into fatty tissues and are metabolized by the body at a lesser rate. This can lead to them having a more prolonged effect.[26] Notably, fentanyl and methadone exhibit heightened lipophilicity values relative to the other opioids tested. This may provide chemical insight into why fentanyl has such a dramatic effect on those who take it and why the level required for a dose of fentanyl to be lethal is so low, generally being estimated at 2 mg of the drug.[27] Methadone in contrast has an even higher lipophilicity, however, the other characteristics, such as electronegativity, chemical hardness, and 3-dimensional structure, are out of the ideal ranges, which may explain why it is not as potent.

Name	Lipophilicity	Water Solubility	GI absorption	BBB permeant	Drug-likeness	Lead-likeness	Synthetic accessibility
Codeine	2.88	Soluble	High	Yes	0.55	Yes	4.89
Fentanyl	**3.78**	Moderately Soluble	High	Yes	0.55	No; 1 violation	2.22
Hydrocodone	1.79	Very soluble	High	Yes	0.55	Yes	4.43
Hydromorphone	1.88	Soluble	High	Yes	0.55	Yes	4.33
Meperidine	2.53	Solubility	High	Yes	0.55	No; 1 violation	2.07
Methadone	**4.06**	Moderately Soluble	High	Yes	0.55	No; 1 violation	2.79
Morphine	1.47	Soluble	High	Yes	0.55	Yes	4.78
Oxycodone	1.12	Very soluble	High	No	0.55	No;1 violation	4.57
Oxymorphone	1.21	Very soluble	High	No	0.55	Yes	4.46
Imitrex	0.72	Very soluble	Low	No	0.55	No; 2 violations	3.07

Table 1. Absorption, distribution, metabolism, elimination, and toxicity values for each of the 10 opioids obtained using the SwissADME. GI and BBB stand for gastrointestinal and blood-brain barrier.

DISCUSSION

This study investigated the molecular characteristics of opioids and how they interact with opioid receptors. The goal was to understand opioid potency trends while also trying to find a chemical explanation for fentanyl's high potency.

The binding affinities determined computationally in this study aligned with published data for fentanyl and meperidine. Ellis et al. reported average docking scores for fentanyl and meperidine bound to the μ opioid receptor (μ OR) to be -9.43 and -7.77 kcal/mol, respectively.[28] These scores are comparable to the binding affinities of fentanyl (-8.7 kcal/mol) and meperidine (-6.4 kcal/mol) for μ OR found in this study. The small differences could be due to the use of different computational docking protocols. Additionally, the computed lipophilicity values obtained in this study were like other experimentally determined values, reinforcing conclusions about fentanyl and lipophilicity.[29,30] The computed electronegativity and chemical hardness values for fentanyl were also comparable to those reported by Sümeyya et al.[31] Overall, key findings for binding affinity, electronegativity, chemical hardness, lipophilicity, and binding pocket analysis were consistent with other studies. Future studies could validate these findings by performing multiple binding simulations to ensure consistency and by conducting bench-top experiments to determine if the results hold in a physical system. These steps would add more certainty to this study's conclusions.

Our findings show that fentanyl is an outlier in many ways. It exhibited the highest binding affinity for three of four opioid receptors, the highest electronegativity and chemical hardness, and the second-highest lipophilicity among opioids studied. When comparing binding pocket interactions between fentanyl, a highly potent opioid receptor agonist, to meperidine, a less potent opioid receptor agonist, it was observed that hydrophobic interactions may be important to fentanyl's high binding affinity. Specifically, fentanyl's long structure consisting of two phenyl groups on each end may allow it to interact more strongly with the μ OR. Specifically, these functional groups appeared to interact with the non-polar residues in the opioid receptor's binding pocket. This could help stabilize the compound and enhance opioid-receptor interactions. Similar studies have also highlighted

the importance of hydrophobic interactions in the μ OR-fentanyl complex. [32,33] This is also supported by fentanyl's high lipophilicity, which may promote the drug's interaction with the μ OR binding pocket. Additionally, the high fentanyl lipophilicity helps it to be absorbed more readily in the body and more easily cross the blood-brain barrier, which is key in opioids as they act on the central nervous system. Overall, these findings suggest that fentanyl's chemical properties are responsible for its high potency. This insight could aid in the development of effective and safer non-μ opioid receptor agonists.

The results of this study also explain how opioids, specifically fentanyl, can be linked to poor clinical outcomes. For example, it is known that stimulation of the μ OR triggers the body's reward system and, in overdose cases, the respiratory depression that leads to opioid overdose death. Fentanyl's high affinity for this receptor indicates that the risk of addiction and overdose is higher for fentanyl than that of other opioids. Furthermore, fentanyl's lipophilic nature allows it to cross the blood-brain barrier more easily, reaching opioid receptors efficiently. This characteristic means that fentanyl not only reaches opioid receptors effectively but also binds them strongly. Consequently, even trace amounts of fentanyl can provide analgesic effects, while a small overdose can be fatal. Ultimately, fentanyl's potency is highly beneficial in a clinical setting, but it also makes the drug extremely dangerous.

CONCLUSIONS

The present computational study demonstrated fentanyl to be the 'best fit' for three of the four opioid receptors in the body. Fentanyl's high affinity for the μ OR is significant because the binding of this receptor stimulates the release of endorphins and dopamine in the brain's limbic system and thus increases its addictiveness and potency relative to the other receptors studied. The μ OR is also associated with triggering respiratory depression, which causes death from an opioid overdose. In addition to its interactions with the μ OR, fentanyl was among the most lipophilic opioids, which means that the drug molecule takes a longer time to exit the body, granting it a longer period of action. The higher binding affinity for the μ OR combined with this elevated lipophilicity may provide chemical insight into the addictiveness and lethality of the drug.

As we currently stand amid a systemic opioid crisis, knowing why fentanyl is so lethal is essential information. This research highlights the lethal nature of fentanyl by identifying key chemical properties, such as lipophilicity and electronegativity, that contribute to its potency as a μ OR agonist. These insights could inspire the design of new opioids that avoid these properties, reducing the risks of addiction and overdose. The study also explores the chemical characteristics of strong agonists for other opioid receptors, potentially guiding targeted drug design for non-μ receptors. Minor variations in electronegativity, chemical hardness, and lipophilicity significantly impact opioid potency, suggesting that slight modifications in chemical structure could alter receptor selectivity and potency. Overall, this study provides valuable information on the structural and chemical factors influencing opioid-receptor binding, which could lead to the development of safer and more effective analgesics.

APPENDIX

The appendix contains two tables and is located on *https://ajuronline.org/*

ACKNOWLEDGMENTS

The authors thank all students enrolled in General Chemistry I (Sections 501 and 502) in Fall 2022 for performing the initial calculations, and the Blugold Center for High-Performance Computing at the University of Wisconsin-Eau Claire (NSF grant CNS 1920220) for the computational facilities.

REFERENCES

1. Chaves, C., Remiao, F., Cisternino, S., Decleves, X. (2017) Opioids and the Blood-Brain Barrier: A Dynamic Interaction with Consequences on Drug Disposition in Brain. *Curr Neuropharmacol* 15, 1156–1173. *https://doi.org/10.2174/1570159x15666170504095823*
2. Pathan, H., Williams, J. (2012) Basic opioid pharmacology: an update. *British Journal of Pain* 6, 11–16. *https://doi.org/10.1177/2049463712438493*
3. Faouzi, A., Varga, B. R., Majumdar, S. (2020) Biased Opioid Ligands. *Molecules* 25, 4257. *https://doi.org/10.3390/molecules25184257*
4. Dhaliwal A., Gupta, M. (2022) Physiology, Opioid Receptor, in *StatPearls [Internet]*. *https://www.ncbi.nlm.nih.gov/books/NBK546642/*
5. Ali, A. H., Ahmed, H. S., Jawad, A. S., Mustafa, M. A. (2021) Endorphin: function and mechanism of action. *Science Archives* 2, 09–13. *https://doi.org/10.47587/sa.2021.2102*
6. Drug Overdose Death Rates, National Institute on Drug Abuse (NIDA) *https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates* (accessed Mar 2023).
7. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B., Thiessen, P., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E. E. (2023) PubChem 2023 update. *Nucleic Acids Research* 51, D1373–D1380. *https://doi.org/10.1093/nar/gkac956*

8. Polik, W. F., Schmidt, J. R. (2021) WebMO: Web-based Computational Chemistry Calculations in Education and Research. *WIREs Computational Molecular Science* 12. *https://doi.org/10.1002/wcms.1554*

9. Shao, Y., Gan, Z., Epifanovsky, E., Gilbert, A. T. B., Wormit, M., Kussmann, J., Lange, A. W., Behn, A., Deng, J., Feng, X., Ghosh, D., Goldey, M., Horn, P. R., Jacobson, L. D., Kaliman, I., Khaliullin, R. Z., Kuś, T., Landau, A., Liu, J., Proynov, E. I., Rhee, Y. M., Richard, R. M., Rohrdanz, M. A., Steele, R. P., Sundstrom, E. J., Woodcock, H. L., Zimmerman, P. M., Zuev, D., Albrecht, B., Alguire, E., Austin, B., Beran, G. J. O., Bernard, Y. A., Berquist, E., Brandhorst, K., Bravaya, K. B., Brown, S. T., Casanova, D., Chang, C.-M., Chen, Y., Chien, S. H., Closser, K. D., Crittenden, D. L., Diedenhofen, M., DiStasio, R. A., Do, H., Dutoi, A. D., Edgar, R. G., Fatehi, S., Fusti-Molnar, L., Ghysels, A., Golubeva-Zadorozhnaya, A., Gomes, J., Hanson-Heine, M. W. D., Harbach, P. H. P., Hauser, A. W., Hohenstein, E. G., Holden, Z. C., Jagau, T.-C., Ji, H., Kaduk, B., Khistyaev, K., Kim, J., Kim, J., King, R. A., Klunzinger, P., Kosenkov, D., Kowalczyk, T., Krauter, C. M., Lao, K. U., Laurent, A. D., Lawler, K. V., Levchenko, S. V., Lin, C. Y., Liu, F., Livshits, E., Lochan, R. C., Luenser, A., Manohar, P., Manzer, S. F., Mao, S.-P., Mardirossian, N., Marenich, A. V., Maurer, S. A., Mayhall, N. J., Neuscamman, E., Oana, C. M., Olivares-Amaya, R., O'Neill, D. P., Parkhill, J. A., Perrine, T. M., Peverati, R., Prociuk, A., Rehn, D. R., Rosta, E., Russ, N. J., Sharada, S. M., Sharma, S., Small, D. W., Sodt, A., Stein, T., Stück, D., Su, Y.-C., Thom, A. J. W., Tsuchimochi, T., Vanovschi, V., Vogt, L., Vydrov, O., Wang, T., Watson, M. A., Wenzel, J., White, A., Williams, C. F., Yang, J., Yeganeh, S., Yost, S. R., You, Z.-Q., Zhang, I. Y., Zhang, X., Zhao, Y., Brooks, B. R., Chan, G. K. L., Chipman, D. M., Cramer, C. J., Goddard, W. A., Gordon, M. S., Hehre, W. J., Klamt, A., Schaefer, H. F., Schmidt, M. W., Sherrill, C. D., Truhlar, D. G., Warshel, A., Xu, X., Aspuru-Guzik, A., Baer, R., Bell, A. T., Besley, N. A., Chai, J.-D., Dreuw, A., Dunietz, B. D., Furlani, T. R., Gwaltney, S. R., Hsu, C.-P., Jung, Y., Kong, J., Lambrecht, D. S., Liang, W., Ochsenfeld, C., Rassolov, V. A., Slipchenko, L. V., Subotnik, J. E., Van Voorhis, T., Herbert, J. M., Krylov, A. I., Gill, P. M. W., Head-Gordon, M. (2014) Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Molecular Physics* 113, 184–215. *https://doi.org/10.1080/00268976.2014.952696*

10. Hill, A. N., Meijer, A. J. H. M., Hill, J. G. (2022) Correlation Consistent Basis Sets and Core Polarization Potentials for Al–Ar with ccECP Pseudopotentials. *The Journal of Physical Chemistry A* 126, 5853–5863. *https://doi.org/10.1021/acs.jpca.2c04446*

11. Putz, M. V. (2021) Chemical Bonding by the Chemical Orthogonal Space of Reactivity. *Int J Mol Sci* 22, 223. *https://doi.org/10.3390/ijms22010223*

12. Parr, R. G., Donnelly, R. A., Levy, M., Palke, W. E. (1977) Electronegativity: The Density Functional Viewpoint. *J Chem Phys* 68, 3801–3807. *https://doi.org/10.1063/1.436185*

13. Pearson, R. G. (2005) Chemical Hardness and Density Functional Theory. *J Chem Sci* 117, 369–377. *https://doi.org/10.1007/BF02708340*

14. Parr, R. G., Pearson, R. G. (1983) Absolute Hardness: Companion Parameter to Absolute Electronegativity. *J Am Chem Soc* 105, 7512–7516. *https://doi.org/10.1021/ja00364a005*

15. Hati, S., Datta, D. (1996) Hardness: A Concept in Inorganic Chemistry. Some Aspects. *Proc Ind Acad Sci: Chem Sci* 108, 143–158. *https://doi.org/10.1007/BF02870020*

16. Hati, S., Datta, D. (1994) Hardness and Electric Dipole Polarizability. Atoms and Clusters. *J Phys Chem* 98, 10451–10454. *https://doi.org/10.1021/j100092a012*

17. Hati, S., Datta, D. (1994) Chemical Hardness of Metal Ions in the Gas Phase: A Thermochemical Approach. *J Chem Soc, Dalton Trans*, 2177–2180. *https://doi.org/10.1039/DT9940002177*

18. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000) The Protein Data Bank. *Nucl Acids Res* 28, 235–242. *https://doi.org/10.1093/nar/28.1.235*

19. Ravindranath, P. A., Forli, S., Goodsell, D. S., Olson, A. J., Sanner, M. F. (2015) AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *PLoS Computational Biology* 11. *https://doi.org/10.1371/journal.pcbi.1004586*

20. Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T. M., Mortenson, P. N., Murray, C. W. (2007) Diverse, High-Quality Test set for the validation of Protein−Ligand docking performance. *Journal of Medicinal Chemistry* 50, 726–741. *https://doi.org/10.1021/jm061277y*

21. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. and Olson, A. J. (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Computational Chemistry* 16, 2785-2791. *https://doi.org/10.1002/jcc.21256*

22. Morris, G. M., Goodsell, D. S., Pique, M. E., Lindstrom, W., Huey, R., Forli, S., Hart, W. E., Halliday, S., Belew, R., Olson, A. J. (2014*)* AutoDock Version 4.2: Automated Docking of Flexible Ligands to Flexible Receptors, *AutoDock Tools.* https://autodock.scripps.edu/wp-content/uploads/sites/56/2021/10/AutoDock4.2.6_UserGuide.pdf (accessed Mar 2025).

23. Humphrey, W., Dalke, A., Schulten, K. (1996) VMD: Visual Molecular Dynamics. *J Mol Graph* 14, 33–38. *https://doi.org/10.1016/0263-7855(96)00018-5*

24. Daina, A., Michielin, O., Zoete, V. (2017) SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Scientific Reports* 7, 42717. *https://doi.org/10.1038/srep42717*

25. Fentanyl Drug Facts, National Institute on Drug Abuse, *https://nida.nih.gov/publications/drugfacts/fentanyl* (accessed Mar 2025).

26. Morak-Młodawska, B., Jeleń, M., Martula, E., Korlacki R. (2023) Study of Lipophilicity and ADME Properties of 1,9-Diazaphenothiazines with Anticancer Action. *Int J Mol Sci* 24, 6970. *https://doi.org/10.3390/ijms24086970*

27. Facts about Fentanyl, United States Drug Enforcement Administration, *www.dea.gov/resources/facts-about-fentanyl* (accessed Mar 2025).

28. Ellis, C. R., Kruhlak, N. L., Kim, M. T., Hawkins, E. G., Stavitskaya, L. (2018) Predicting opioid receptor binding affinity of pharmacologically unclassified designer substances using molecular docking. *PLOS One* 13. *https://doi.org/10.1371/journal.pone.0197734*

29. Bird, H. E., Huhn, A. S., Dunn, K. E. (2023) Fentanyl absorption, distribution, metabolism, and excretion: narrative review and clinical significance related to illicitly manufactured fentanyl. *Journal of Addiction Medicine* 17, 503–508. *https://doi.org/10.1097/adm.0000000000001185*

30. Sutcliffe, K. J., Corey, R. A., Alhosan, N., Cavallo, D., Groom, S., Santiago, M., Bailey, C., Charlton, S. J., Sessions, R. B., Henderson, G., Kelly, E. (2022) Interaction with the lipid membrane influences fentanyl pharmacology. *Advances in Drug and Alcohol Research* 2. *https://doi.org/10.3389/adar.2022.10280*

31. Serîn, S., Utku, T., Kaya, G. (2021) Quantum chemical calculations on fentanyl used as potent analgesic. *MTU Journal of Engineering and Natural Sciences* 2. *https://doi.org/10.46572/naturengs.1022735*

32. Zhuang, Y., Wang, Y., He, B., He, X., Zhou, X. E., Guo, S., Rao, Q., Yang, J., Liu, J., Zhou, Q., Wang, X., Liu, M., Liu, W., Jiang, X., Yang, D., Jiang, H., Shen, J., Melcher, K., Chen, H., Jiang, Y., Cheng, X., Wang, M.-W., Xie, X., Xu, H. E. (2022) Molecular recognition of morphine and fentanyl by the human μ-opioid receptor. *Cell* 185, 4251–4253. *https://doi.org/10.1016/j.cell.2022.09.041*

33. Vo, Q., Mahinthichaichan, P., Shen, J., Ellis, C. (2021) How mu-Opioid Receptor Recognizes Fentanyl. *Nat Commun* 12, 984. *https://doi.org/10.1038/s41467-021-21262-9*

ABOUT THE STUDENT AUTHORS

Keegan Gunderson is a Biomedical Engineering student at the University of Wisconsin–Eau Claire. After graduation, he plans to go to medical school. He is passionate about using current clinical knowledge and research to provide exceptional patient care to individuals in rural communities.

Faith Oldenburg is a third-year student of Nursing and Spanish at the University of Wisconsin Eau Claire. After graduation, she hopes to pursue a doctoral degree to become a certified nurse midwife. She is passionate about advancing and promoting research-based practice in order to provide excellent healthcare to underserved communities, especially those experiencing elevated rates of maternal and infant mortality.

PRESS SUMMARY

Deaths from opioid overdoses are at an all-time high due to their use clinically and increasing prevalence in street drugs. This study intends to identify the mechanisms that make them such potent stimulants for opioid receptors, specifically the Mu opioid receptor, which causes respiratory depression and potentially death when overstimulated. Using computational chemistry techniques, it was found that multiple chemical properties, namely a high electronegativity and chemical hardness, high lipophilicity, and the presence of non-polar groups for hydrophobic interactions may be responsible for making some opioids potent ligands for opioid receptors. This knowledge is significant to the design of future analgesics that minimize health risks, including addiction and fatal overdose.

Isolation and Characterization of Novel Marine Bacteriophages from Santa Rosa Island, Florida, Using Local and Nonlocal Bacterial Strains as Hosts

Brittany C. Yencho[*a,b], Charles J. West*[a,c], Andrew J. Brown*[a,d], & Conor R. Flannigan*[a,e]

[a]*Department of Biology, University of West Florida, Pensacola, FL*
[b]*School of Veterinary Medicine, Louisiana State University, Baton Rouge, LA*
[c]*Department of Microbiology, University of Alabama at Birmingham, Birmingham, AL*
[d]*United States Environmental Protection Agency, Gulf Breeze, FL*
[e]*Department of Surgery, Duke University, Durham, NC*

https://doi.org/10.33697/ajur.2025.138

Students: yenchobrittany@gmail.com, westc@uab.edu, brown.andrew.fl@gmail.com, conorflannigan@gmail.com*
*Mentor: hchung@uwf.edu**

ABSTRACT
Bacteriophages, or phages, are viruses that infect bacteria and are the most prevalent biological entities in the world. While phage research has continued to develop, the investigation of phages within ecological biology remains in the early stages and requires an extensive database of environmentally derived phages, such as from marine ecosystems. The objectives of this research were to a) determine the local beach bacterial composition and select locally isolated strains as potential bacteriophage hosts, b) determine the efficacy of local bacteria as marine bacteriophage hosts (*Erythrobacter citreus* Pensacola AB and *Microbacterium oleivorans* Pensacola AB), and c) compare the host efficacy between locally isolated bacterial strains and non-local *Microbacterium* isolates (*Microbacterium* sp. Casco Bay and *Microbacterium foliorum* NRRL B-24224). The results suggest that of the two locally isolated bacterial strains tested, *Microbacterium oleivorans* was not a successful host for the scope of this research. In contrast, *Erythrobacter citreus* Pensacola AB was an efficient host across several trials. Of the non-local bacteria, both *Microbacterium* sp. Casco Bay and *Microbacterium foliorum* NRRL B-24224 were successful hosts at isolating phage, although not at the same consistency as *Erythrobacter citreus*. This research concluded that local bacterial strains could be effective for phage hunting and that the overall success of finding local marine phage was not dependent on using local bacterial strains. Four phages, two *M.* sp. Casco Bay phages, one *M. foliorum* phage, and one *E. citreus* phage, were selected for further purification and characterization. The four phage genomes were sequenced to characterize the molecular nature of these marine phages.

KEYWORDS
Microbiome; Bacteriophage; Marine phage; Plaque Assay; Genomic sequencing; Marine ecology; Zinc chloride ($ZnCl_2$) precipitation; PEG (polyethylene glycol) precipitation

INTRODUCTION

Bacteriophage (phage) biology has long been an important research topic in medical microbiology. Over 100 years ago, phage research was popularized by investigating the mechanisms by which phage particles destroy bacterial cells, hence the initial phase of phage therapy. This trend was halted as antibiotics became prevalent and replaced clinical phage research.[1] At the time, the bulk of phage biology was established by studying a small subset of *E. coli* phages such as lambda phage, however, there remained many gaps in the knowledge of phage biology. With the increase in bacterial antibiotic resistance,[2] interest in phage research has resurged in the 21st century.[1] It has been revealed that phages play essential roles in influencing their bacterial hosts' metabolic behaviors by utilizing phage-encoded auxiliary metabolic genes.[3] Such influence in the microbial ecosystem was especially noted in studying marine phages.[3]

The prevalence of phage in the marine environment was highly underestimated, with the original estimated virus-to-bacteria ratio (VBR) being around ten virus-like particles per bacterial cell in surface seawater. However, as technology has advanced, the VBR in marine environments has been reevaluated to be as high as 100:1.[3] As a result, a large pool of unknown marine phage genes may exist to regulate phage-bacteria interaction, thus creating an untapped reservoir for societal benefit. Societal niches potentially impacted by phage functions include, but are not limited to, applications in molecular editing,[4] drug development, [5,6] and environmental mitigation.[7,8] By characterizing new phages from the marine environment, one can better understand overall phage biology and their molecular mechanisms.

While the exact composition of bacterial communities varies throughout the oceans, the literature suggests that *Pseudomonas* sp., *Vibrio* sp., *Achromobacter* sp., *Flavobacterium* sp., and *Micrococcus* sp. are the most prevalent marine bacterial residents.[9,10] Currently, marine bacterial lineages commonly used for phage hunting efforts reside within the phyla Cyanobacteria, Proteobacteria, and Bacteroidetes.[11] To isolate phages from the local marine environment, it is crucial to understand which bacterial species represent the local microbiome within environmental samples. Metagenomic and lab culturing approaches can be used to determine the local bacterial composition, and to culture and identify bacterial strains as possible hosts for phage hunting. However, several unknowns challenge this approach, including the ability to culture marine bacteria in the local environmental sample under laboratory conditions [12, 13] and the abundance and geographical distribution of phages infecting a marine bacterial strain of interest. It is also valuable to explore the use of nonlocal hosts with local environmental samples.

This research tested the hypotheses that the local beach bacterial composition could be determined through environmental DNA (eDNA) analysis, that locally isolated marine bacterial strains would have associated phage in the environmental samples, and that the locally isolated bacteria are more efficient in isolating marine phages than non-local bacterial strains. Sand samples from various locations along Santa Rosa Island, Florida, were used to culture eight bacterial strains. Two non-pathogenic bacterial strains of the eight strains isolated were chosen as local phage hunting hosts, *Microbacterium oleivorans* Pensacola AB (*M. oleivorans*) and *Erythrobacter citreus* Pensacola AB (*E. citreus*). Additionally, two non-local bacterial strains were also chosen as phage hunting hosts, *Microbacterium foliorum* NRRL B-24224 (*M. foliorum*) and *Microbacterium* sp. Casco Bay [14] (*M.* sp. Casco Bay). The efficiency of phage hunting using these hosts was compared by infecting these bacterial strains with various environmental samples. Plaques are observed within a bacterial lawn of the host bacteria to confirm the presence of phage. A plaque is a clear zone on a bacterial lawn that is formed when a single phage particle infects, replicates, and lyses bacteria during the incubation time. The morphology and the number of plaques formed on the plate are good indicators to determine the efficacy of bacterial host and to characterize phage morphology during the initial stage of the phage hunting process.[15] At the end of this research in the summer of 2023, four phages were further isolated and characterized by their morphological, physiological, and genomic features.

METHODS AND PROCEDURES

Profiling the local bacterial composition

Wet intertidal sand was collected from Pensacola Beach near the fishing pier at 30.3304928, -87.1409214 on January 23, 2022. Environmental DNA was extracted from a 0.5 g sand sample using a Qiagen DNEasy® PowerSoil® isolation kit. A metagenomic library was prepared using Oxford Nanopore Technologies' 16S Barcoding Kit, which utilizes barcoded 27F and 1492R primers [16] and quantitative PCR amplification. The amplified 16S rRNA segments were then sequenced in situ using MinION MK1C flow-through sequencer with an R9 flow cell (Oxford Nanopore Tech Inc.), and the data returned was further analyzed using EPI2ME 16S workflow software (Oxford Nanopore Tech Inc.) to identify the bacterial taxa in the metagenomic sample.

Bacteria isolation and identification

A 20 g sample of the sand taken from Pensacola Beach was combined with 50 mL sterile water and shaken at 33 RPM in an incubator at 37 °C for 24 hours to isolate bacterial cultures for use as phage hosts. The resulting supernatant was serially diluted from 10^0 to 10^{-5}. Using sterile microbiology techniques, duplicate 5 mL aliquots from each dilution factor were spread onto Luria-Bertani (LB) and Marine 2216 nutrient agar plates, incubated at 25 °C, and monitored daily for growth. Colonies were viewed under a microscope to determine isolation (i.e., not touching other colonies) and uniqueness based on morphological factors such as color, form, elevation, and margin. Single colonies were then inoculated in 500 µl of appropriate liquid nutrient media (LB or Marine 2216) and purified twice by streaking plates, picking single colonies, and inoculating. Single colonies from the second round of purification were used to produce a final uncontaminated streak plate. A Qiagen DNEasy® Blood & Tissue spin-column isolation kit was used to extract DNA from a single pure colony of each cultured bacterium. Subsequently, using non-barcoded 27F and 1492R primers [16], the 16S rRNA genes from each bacterial DNA were amplified; the resulting amplicons were sent to North Carolina State University's Genomic Sciences Laboratory for Sanger DNA sequencing. The returned sequence data from each sample were then compared to the NCBI database using the nucleotide BLAST tool to identify the species name of the isolated bacterial strain.[17]

Culturing hosts on multiple growth media

The two non local hosts *Microbacterium* sp. Casco Bay and *Microbacterium foliorum* were acquired through other labs. *M.* sp. Casco Bay[14] was gifted to us by the Tarbox Lab at Maine Community College and originally discovered from marine mud samples in South Portland, Maine. This strain is obtainable from Bigelow Laboratory National Center for Marine Algae and Microbiota as *Microbacterium* sp. NCMA B81. *Microbacterium foliorum* NRRL B-24224 was a stock strain obtained from the SEA-PHAGES stock center. Each of the chosen bacterial hosts (*M. oleivorans, M. foliorum, M.* sp. Casco Bay*, and E. citreus*) was tested using streak plates and liquid cultures on four different medium types. The bacterial growth was analyzed after 48 hours at 28 °C.[15] The media types used were peptone yeast calcium agar media (PYCa) with 0.01 g/L cycloheximide (CHX) and a 0.1% dextrose concentration,[15]

Marine 2216 agar (BD Difco), and mock seawater made with 25 g/L Instant Ocean salt, 0.5 g/L yeast, 0.5 g/L peptone, 0.5 g/L Casamino acids (Bacto™ Casamino Acids), and 1.2 g/L glycerol in water.#

Collecting environmental samples
Environmental samples were collected in two different locations on Santa Rosa Island: Pensacola Beach, Florida, near the fishing pier (30.3304928, -87.1409214), and Navarre Beach, Florida (30.38034, -86.85560). Sample types included dry beach sand, wet intertidal sand, underwater sand, Gulf of Mexico seawater, seagrass of unknown species, coquina clams (*Donax variabilis*), and Atlantic sand crabs (genus *Emerita*).

Phage Isolation and Purification
Phage isolation was conducted using *Phage Discovery Guide* methods.[15] The environmental samples tested included dry sand, underwater sand, dissected marine animals, seagrass, and mixed samples. 30 ml Marine 2216 or PYCa broth was added to each sample and shaken overnight. The supernatant was centrifuged, filtered through a 0.22 μm vacuum filter, and aliquoted for testing with direct and enriched isolation methods. Filtrates for the direct isolation method were followed by plaque assay. Filtrates for the enriched isolation method were mixed with 0.5 ml of a bacteria host culture and incubated for two days at 220 rpm, followed by filtration, before proceeding with plaque assay. Alternatively, the enriched filtrate was precipitated with PEG solution before plaque assay.

Concentrating phage particles using polyethylene glycol (PEG) and zinc chloride ($ZnCl_2$)
Each one mL of enriched sample filtrate or phage lysate of known titer was mixed with 600 μl 20% PEG/2.5M NaCl through inversion and incubated at 4 °C overnight.[15] The lysate was centrifuged, resuspended in 1/100th of the original volume in phage buffer, and plated for a plaque assay. For testing the efficiency of using $ZnCl_2$ to concentrate phage lysate, one mL of lysate was mixed with 25 μL of 2M $ZnCl_2$ through inversion and incubated at 37 °C for 10 mins.[15] The lysate was centrifuged, resuspended in 1/100th of the original volume in phage buffer, and plated for a plaque assay.

To determine the quality of phage particles post-concentration, the concentrated phage particles were examined by a plaque assay. Gel electrophoresis was used to examine the phage DNA: 2 μL of 0.8% SDS/6X loading dye was added to 10 μL of concentrated lysate and incubated at 65 °C for 5 mins; this treatment broke the phage particles to release phage DNA.

Plaque assay
A plaque assay was carried out by first incubating 250 μl of the bacterial host in broth with 500 μl of filtrate derived from the direct method or 10 μl from enriched isolation for 15 minutes. Subsequently, the sample was mixed with 3 ml of the desired top agar (1X PYCa top agar or 1.25X Marine 2216 top agar) and plated onto an appropriate agar plate. The plates were incubated at 25 °C for 24-48 hours before observing plaques.

After a plaque of interest was identified, it was picked and dipped into a tube containing 100 μl of phage buffer, generating a phage lysate containing phage particles. The phage lysate was then serially diluted and examined by plaque assay for consistency of plaque morphology to ensure the purity of the phage of interest (purification round 1). A plaque derived from this process was to go through a similar process for another round of purification. After three rounds of purification, the phage of interest was amplified using the amplification process described in the Phage Discovery Guide to produce a high-titer lysate for characterization and archiving.[15]

Phage DNA isolation, sequencing, and genome assembly
DNA was extracted from the high-titer lysates using the Norgen Phage DNA isolation kit. The quantity and quality of each sample of extracted DNA were determined using a Qubit fluorometer and gel electrophoresis, respectively. A total of 2 μg of DNA from each phage sample was sent off for sequencing and genome assembly at the North Carolina State Genomic Sciences Laboratory.

Transmission electron microscopy (TEM)
Phage lysates were shipped to Taiwan for transmission electron microscopy (TEM) analysis by the Lin lab at National Taiwan University. A high-titer lysate of interest placed on Formvar-coated copper grids was negatively stained with 1% uranyl acetate.[18] Imaging was carried out with a Hitachi H-7650 microscope operated at 75 kV and equipped with a 1kx1k CCD detection camera (Gatan 782). For each novel phage, two photos capturing two different phage particles (two individual phage particles per isolate) were analyzed using ImageJ measuring software to measure the length and width of phage capsids and tails.

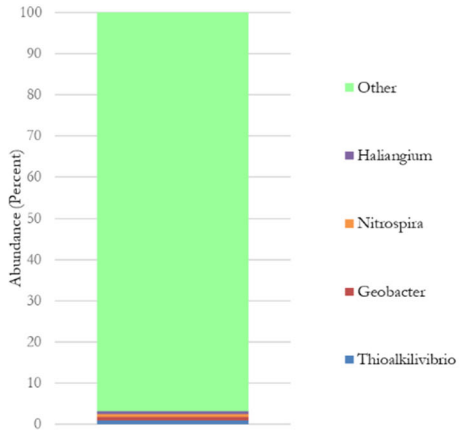
RESULTS

To test if the locally isolated bacterial strains are more efficient in isolating marine phages than non-local bacterial strains, this project began by isolating and characterizing local bacterial strains from Pensacola Beach. This project used common methodologies of phage hunting to isolate, purify, and characterize novel phages.

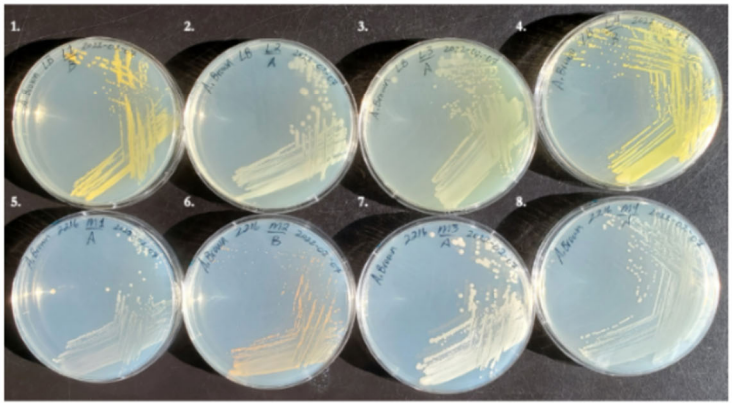
Isolating local marine bacterial strains and choosing hosts for phage hunting

Wet intertidal sand was collected from Pensacola Beach to determine the local bacterial composition and isolate bacterial strains (see Methods and Procedures). The former relied on metagenomic information based on profiling 16S rRNA data derived from the sand sample, and the latter utilized microbiological culturing techniques and subsequent 16S rRNA sequences of isolated bacterial strains.

1A.



1B.



1C.

No.	Bacterial Species	Percent Abundance
1	*Sphingomonas dokdonensis*	0.000
2	*Pseudomonas guineae*	0.001
3	*Pseudomonas alcaligenes*	0.000
4	*Microbacterium oleivorans*	0.001
5	*Pseudomonas benzenivorans*	0.008
6	*Erythrobacter citreus*	0.001
7	*Rheinheimera aquimaris*	0.005
8	*Pseudomonas benzenivorans*	0.008

Figure 1. Determining the bacterial composition of bacterial strains isolated at Pensacola Beach. (1A) This graph shows the bacterial genera that made up the majority of the sand sample analysis as calculated using total metagenomic reads, which were generated via Nanopore sequencing and EPI2ME analysis. **(1B)** The eight bacterial isolates streaked on media; Luria-Bertani agar in the top row of plates (1-4), and Marine 2216 agar in the bottom row (5-8). **(1C)** The eight bacterial isolates, which correspond to **1C** by number, were identified employing 16S rRNA Sanger sequencing and BLAST analysis and the bacterial abundance was determined referencing the Epi2Me report.

Analysis of barcoded 16S metagenomic samples extracted from Pensacola Beach sand revealed that Thioalkalivibrio was the most abundant genus relative to the sample (**Figure 1C**). Geobacter, Nitrospira, and Halangium followed close behind. Most genera described in the EPI2ME report, including the four mentioned previously, were gram-negative. While the metagenomic data provided valuable information about the types of bacteria that thrive in our site of interest, culturing bias prevents us from growing the most present bacteria due to non-reproducible environmental conditions. Despite this, serially diluted spread plates were made using a portion of the same Pensacola Beach sand sample to isolate bacterial strains for phage hunting. Single colonies appeared on LB and Marine 2216 plates with a dilution factor of 10^{-2} to 10^{-5}, and morphology was analyzed under a light microscope. Eight unique single colonies were picked, four from LB plates and four from Marine 2216 plates. The strains were

purified twice and streaked on LB and Marine 2216 nutrient agar plates (**Figure 1B**). Subsequently, 16S rRNA genes from each purified bacterial sample were amplified and sequenced by North Carolina State University's Genomic Sciences Laboratory. The resulting sequence data were then compared to the NCBI database using the nucleotide BLAST tool [19]; the identified bacterial species are summarized in the table of (**Figure 1A)**. All isolated strains were found to be present in the metagenomic data. The eight sequenced isolates matched seven different species from the NCBI database. Two isolates were named *M. oleivorans* Pensacola AB and *E. citreus* Pensacola AB based on their matches with *M. oleivorans* and *E. citreus*, respectively; the former is gram-positive, and the latter is gram-negative. These strains were selected as representative local hosts for marine phage hunting due to their differing gram stains to study the potential phage diversity.

Choosing growth media for phage hunting

The medium preferences of *M. oleivorans* and *E. citreus*, *M. foliorum,* and *M.* sp. Casco Bay) were evaluated to optimize growth conditions for phage hunting. All four bacterial hosts were tested for growth on agar plates of the three medium types: PYCa, Marine 2216, and mock seawater medium. All plates were incubated at 28 °C for 48 hours (see **Table 1** and **Figure 2**, and data not shown).

M. foliorum, *M.* sp. Casco Bay, and *M. oleivorans* grew best on PYCa media and grew moderately on both Marine 2216 and mock seawater media despite a slower generation time. For *M. foliorum* and *M.* sp. Casco Bay, there was rapid growth in the PYCa medium. *M. oleivorans* grows well in the PYCa medium but at a slower pace than *M. foliorum* and *M.* sp. Casco Bay. All three *Microbacterium* strains grew consistently but slowly in Marine 2216 and had the most minor growth on mock seawater media. *E. citreus* grew best on the Marine 2216 medium and had significantly reduced colony growth on the mock seawater medium. *E. citreus* had no growth onPYCa media. Ultimately, PYCa was chosen to grow, assay, and amplify *M. foliorum*, *M. oleivorans*, *M.* sp. Casco Bay, and respective phages. Likewise, Marine 2216 was chosen for *E. citreus* and its isolated phages.

Bacteria Growth			
Bacterial Strain	**PYCa**	**Mock Seawater**	**Marine 2216**
M. foliorum	++	+	+
***M.* sp. Casco Bay**	++	+	+
M. oleivorans	++	+	+
E. citreus	–	+	+

Table 1. Testing three different growth media on the four bacterial hosts for optimal growth. Each bacterial host was tested on three different growth media. A single plus (+) sign indicates decent growth, a double plus (++) indicates rapid and significant growth, and a dash (-) symbol indicates no growth. The responses in red indicate the chosen media used in the subsequent phage research.

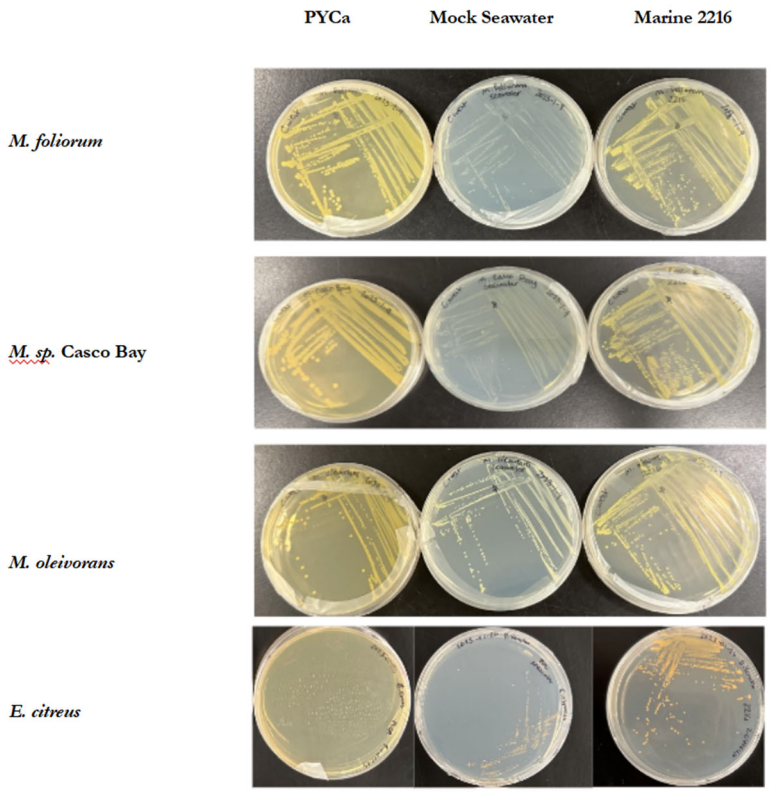


Figure 2. Bacterial growth on three different media. Hdfk#edfwhuldd#krw#z dv#whwhg#rq#wkh#wkuhh#gliihuhqw#jurz wk#p hgld#S\ F d/#P rfn#Vhdz dwu/#bqg# P dulqh#5549,#diwu#7;#krxu#ri#ltxlg#fxowuh#qfxedwrq#dw#5;°C.

Phage hunting trials

This project used plaque assay to identify phages' presence and measure the efficiency of phage hunting using the locally isolated marine bacterial strains and non-marine bacterial strains. **Figure 3** shows two examples of plaque morphology observed during phage hunting trials described below.

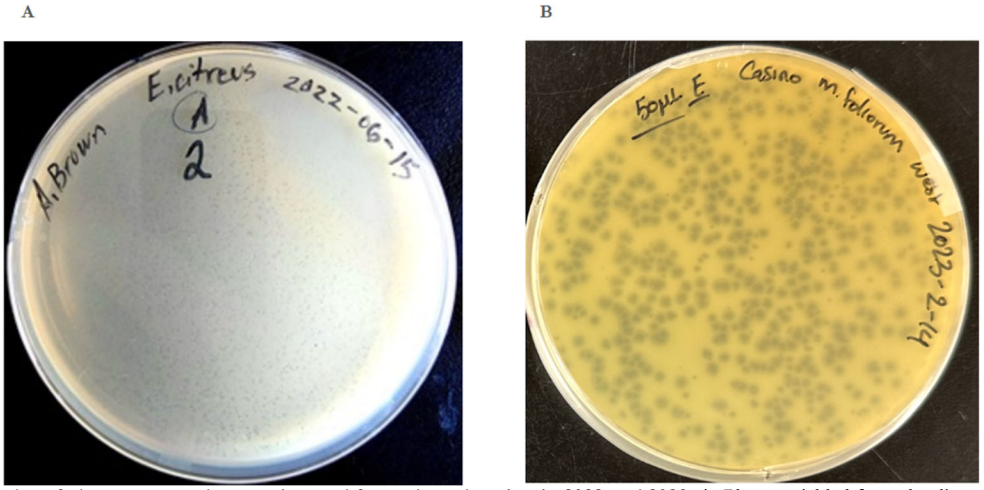


Figure 3. Two examples of plaque assay- plaques observed from phage hunting in 2022 and 2023. A. Plaques yielded from the direct isolation of a dry sand sample from Pensacola Beach, Florida, infecting *E. citreus*. Marine 2216 was used for the media preparations. **B.** Plaques yielded from enriched isolation of a mixed marine sample from Pensacola Beach, Florida, infecting *M. foliorum* in PYCa media preparations.

The first two marine phage hunting attempts were to isolate phages from dry, wet, and underwater sand samples collected from Pensacola Beach. These sample filtrates were tested using *M. oleivorans* and *E. citreus*. No phages from the three environmental

samples were found when using *M. oleivorans* as the host either through direct isolation or medium-enrichment methods. All three ecological samples showed the presence of phage for *E. citreus* for direct samples using the direct isolation method. Wet and underwater sand had few plaques, and dry sand had too many to count.

Determining the preferred method of phage lysate precipitation

Lysate precipitations with various agents, such as $ZnCl_2$ and PEG solutions, help concentrate the sample lysate before infection.[20,21] To determine which reagent was more effective, *M. foliorum* phage Zepp, previously isolated in the lab, was used to test for usage in environmental phage hunts. After the precipitation treatment of Zepp with either $ZnCl_2$ or PEG, plaque assay was used to determine the efficiency of its infection of *M. foliorum* PEG-precipitated lysates showed a 1.6-fold titer increase from $2.5*10^8$ PFU/mL to $4*10^8$ PFU/mL, while $ZnCl_2$ precipitated lysates saw a 0.23-fold decrease from $6*10^8$ PFU/mL to $1.4*10^8$ PFU/mL. The percentages of phage particles from the original lysate that remained in the precipitated lysates were calculated to be 80% and 0.2% for PEG and $ZnCl_2$ precipitation, respectively (**Figure 4A**).

DNAs from the PEG-precipitated and unprecipitated Zepp phage lysates were examined to verify that PEG-precipitation did not damage phage DNA. **Figure 4B** shows that the lane containing undiluted DNA extracts from PEG-precipitated samples had a band (12 kb and larger sized) with a higher concentration of phage DNA than unprecipitated, and the DNA was still intact for use in downstream sequencing applications. Note that the DNA extracts were not treated with nuclease before incubation with SDS (See Methods and Procedures), which indicates that the smear in Lane 3 might be due to either degraded phage or host DNA fragments that remained in the lysate buffer during the Zepp lysate preparation. Judged by the increased titer, high percentage of infectious particle yield, and maintained DNA quality of PEG-precipitated Zepp lysates, it was found that precipitation of phage lysates with a PEG/NaCl solution was highly effective and can be used in various lysate concentration methods. These results showed that for our purposes, PEG-precipitation, but not $ZnCl_2$, is an effective method to concentrate a sample lysate and can be used throughout the phage-hunting process to increase the isolation yield. The PEG-precipitation method was adopted in the phage hunting trial in 2023.

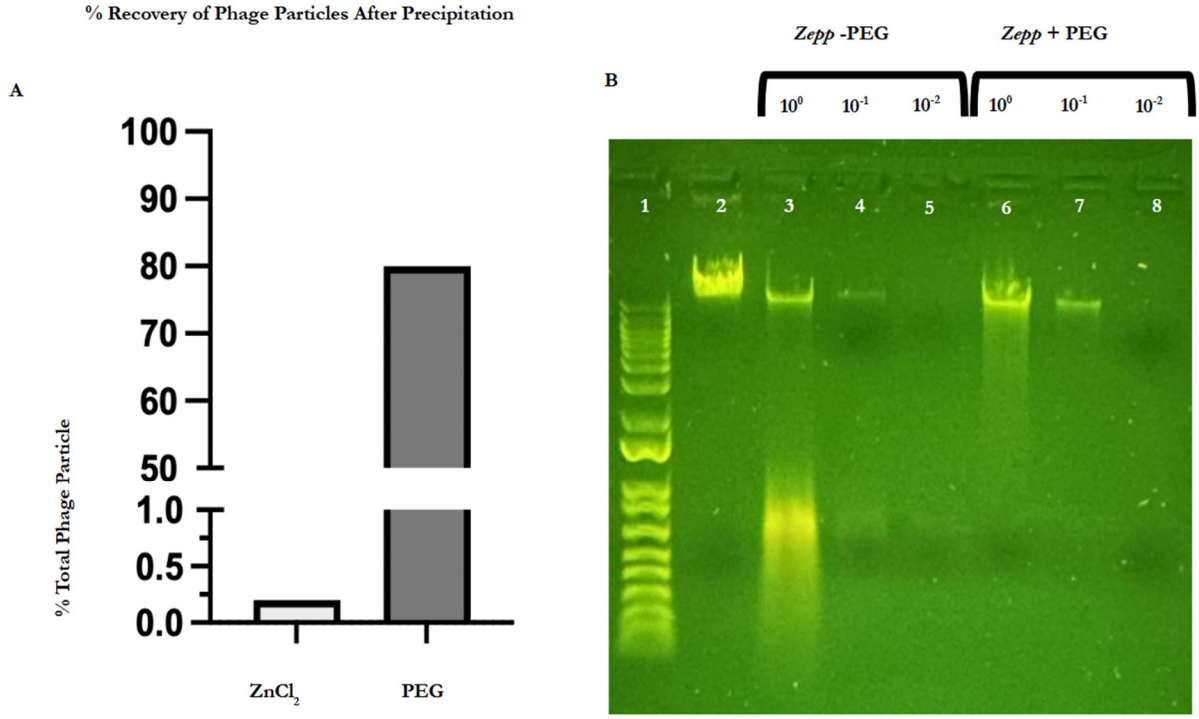


Figure 4. Comparison of Phage Lysate Precipitation Methods. The *M. foliorum* phage Zepp was precipitated with 2M $ZnCl_2$ or 20% PEG/2.5M NaCl. Titers were calculated through plaque assays with Zepp's host, *M. foliorum*. **Figure 4A** shows the percentage of phage particles remaining in lysate after precipitation. (**B**) Gel electrophoresis of isolated DNA from above PEG-precipitated samples. **Lane 2** shows DNA from the *E. coli* phage *Lambda* (λ). Lanes 3-5 are 10-fold dilutions of DNA from PEG-precipitated Zepp lysates and Lanes 6-8 are 10-fold dilutions of DNA from unprecipitated Zepp lysates.

Table 2 summarizes the phage-hunting results of all the trials in 2022-2023, with and without PEG precipitation treatments. Overall, six environmental samples collected from areas along Santa Rosa Island (Pensacola Beach and Navarre Beach) were used, including dry sand, underwater sand, ocean water, marine animals, seagrass, and mixed samples. Three findings were made: a) when using environmental sample lysates to infect *E. citreus*, the PEG-precipitation method was not helpful as plaque assays

identified no indication of phages. b) The PEG-precipitation method was successful when *M. foliorum* and *M.* sp. Casco Bay were used as bacterial hosts. c) Numerous *E. citreus* phages were isolated from sand samples. Phages isolated from the seagrass sample infected *M.* sp. Casco Bay, and the mixed sample infected *M. foliorum*. These results suggest that, without concentrating sample lysates, bacterial isolate *E. citreus* was more efficient in marine phage hunting than the non-local bacterial strains *M. foliorum* and *M.* sp. Casco Bay, which is consistent with our hypothesis.

	Phage Hunting Results							
	M. oleivorans		***M. foliorum***		***M. sp* Casco Bay**		***E. citreus***	
Samples	**-PEG**	**+PEG**	**-PEG**	**+PEG**	**-PEG**	**+PEG**	**-PEG**	**+PEG**
Dry /Wet Sands	–	N/A	–	–	–	–	+	–
Seagrass	N/A	N/A	–	–	–	+	–	–
Mixed sample**	N/A	N/A	–	+	–	N/A	N/A	N/A

Table 2. Summary of phage hunting results using either no PEG ("-PEG") or with PEG ("+PEG") treatment in 2022-2023. The results are denoted with (-) for no phage presence and (+) for the presence of phages. Phage hunting trials occurred in January, June, and August of 2022, February and March of 2023. Due to the culture maintenance issues with *M. oleivorans,* this host was not used in 2023 for phage hunting. ** Mixed sample included coquina, beach sand at the waterline and dry beach sand.

Morphological and Genomic characterization of four phages

Four phages were successfully isolated and subjected to further purification and amplification. Among the four, WestPM, derived from a mixed sample is an *M. foliorum* phage; PortlandC27 and PensacolaC28, both derived from the sea grass sample, are *M.* sp. Casco Bay phages; Horizon is an *E. citreus* phage. All four phages were characterized by observing plaque morphology and through transmission electron microscopy (TEM). Regarding the plaque morphology, Horizon is distinctly different from the other three. While the phages from Microbacterium species had plaques roughly 1.5-2 mm in diameter, Horizon's plaques are extremely tiny and barely visible to the naked eye (about 0.25 mm). **Table 3** and **Figure 5** provide a further description and images of the plaque characteristics of all four phages. The TEM analysis showed that Horizon has a prolate polyhedral capsid shape with a tail, and the other three phages have isometric polyhedral capsids with a tail. **Figure 5** shows the morphology of WestPM, PensacolaC28, PortlandC27, and Horizon (Detailed analysis of WestPM [22] and PensacolaC28 [23] has been published separately). Each phage's average capsid transverse diameter and tail length were calculated based on measurements taken from two TEM photographs using ImageJ software. WestPM had an average capsid transverse diameter of 53.5 nm and an average tail length of 114.0 nm.[22] PortlandC27 had an average capsid transverse diameter of 39.0 nm and an average tail length of 94.4 nm. Similarly, PensacolaC28 had an average capsid transverse diameter of 39.2 nm and tail length of 91.4 nm.[23] Lastly, Horizon was determined to have an average transverse diameter width of 52.8 nm and an average tail length of 222.9 nm.

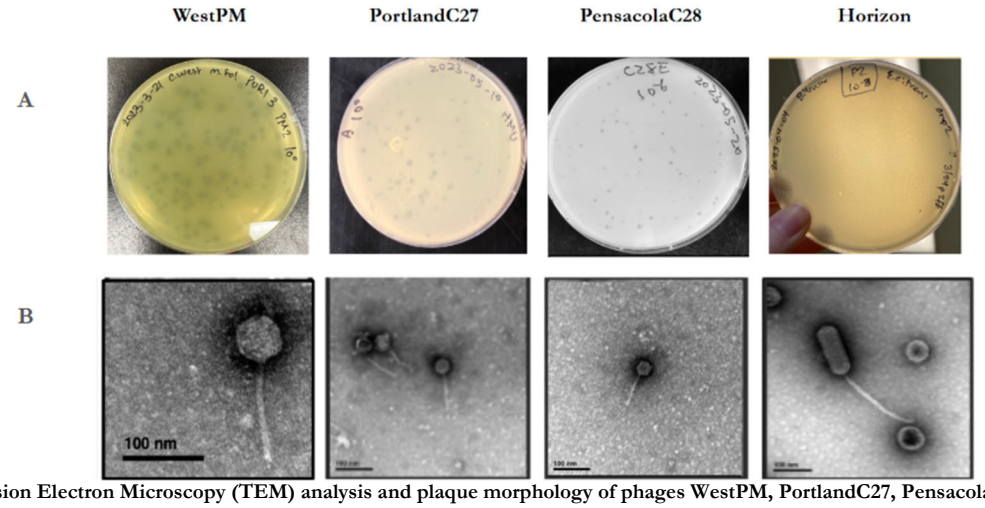


Figure 5. Transmission Electron Microscopy (TEM) analysis and plaque morphology of phages WestPM, PortlandC27, PensacolaC28, and Horizon. **A.** Plaque morphology of four phages (WestPM, PortlandC27, and PensacolaC28 on the PYCa medium, and Horizon on the Marine 2216 medium. **B.** Negative staining TEM of the four bacteriophages as seen in **A.** The average capsid width and tail length were measured using ImageJ software (n=2). Images for WestPM and PensacolaC28 were adapted from West *et al*, 2025 and Girard *et al*, 2025 respectively.

All four phages mentioned maintained a high enough titer throughout the amplification process for sequencing. The titer of WestPM was approximately 1 x 10^9 PFU/mL. PensacolaC28 and PortlandC27 had the highest titers at 4.0 x 10^{10} PFU/mL. The titer of Horizon was hard to determine because the plaque size was tiny, and the bacterial lawns were faintly colored (**Figure 5.**), however, an estimate was 2.3 x 10^8 PFU/mL. Gel electrophoresis was used to test the quality of the extracted phage DNA. **Figure 6** shows that each of the phage genomic DNA was larger than 12 kb. A total of 2 µg of each extracted DNA was then sequenced using Illumina sequencing by North Carolina State University's Genomic Sciences Laboratory. Over a million reads were obtained for each phage DNA to assemble a single contig, i.e., a phage genome. Sequencing of WestPM and PensacolaC28 resulted in 6541-fold and 19692-fold coverage respectively. It was determined that PortlandC27 and PensacolaC28 had the shortest genome lengths at around 17 kb among the four phage genomes. WestPM had a genome length of 39.7 kb, and Horizon had the largest genome length of 151.4 kb. Recent genomic analysis of WestPM [22] and PensacolaC28 [23] showed that WestPM is in the subcluster of EA11 and PensacolaC28 is a singleton of the Actinobacteriophage database [15] and Horizon is in progress. Additionally, for PortlandC27 and PensacolaC28, the genome of the two phages only differed by 39 base pairs, suggesting that PortlandC27 and Pensacola C28 are duplicates of the same phage isolated from independent plaques on the sample plate.

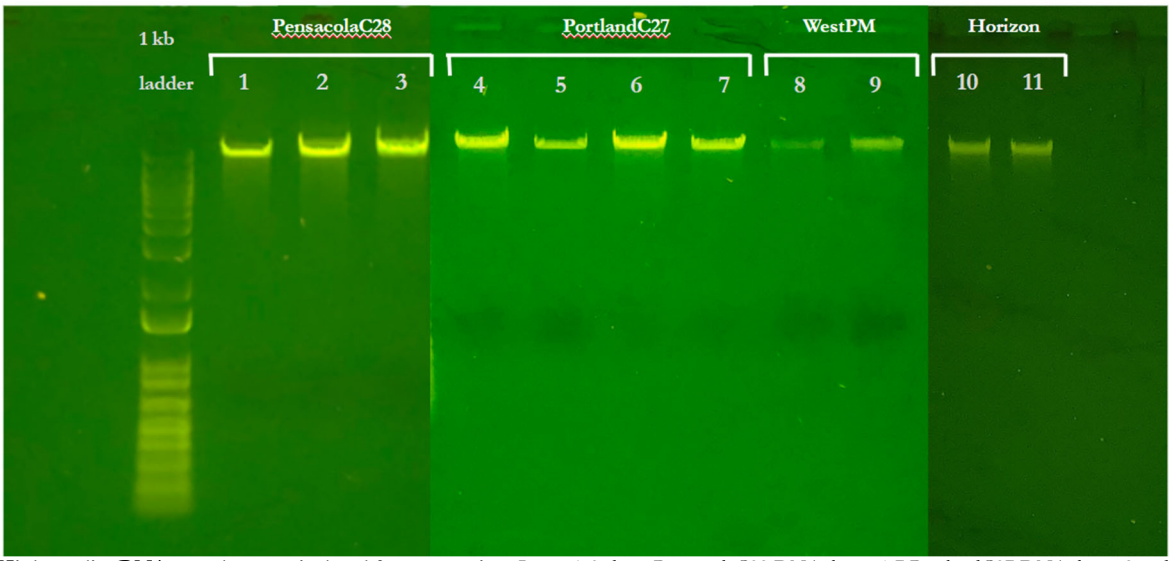


Figure 6. High-quality DNA samples were isolated for sequencing. Lanes 1-3 show PensacolaC28 DNA, lanes 4-7 PortlandC27 DNA, lanes 8 and 9 WestPM DNA, lanes 10 and 11 show Horizon DNA, 0.5 µg DNA per lane. DNA was extracted from the phage lysates using the Norgen Biotek Phage DNA Isolation Kit. A total of 2 µg of DNA per phage was sent off for sequencing.

Characterization of isolated phages					
Host	**Bacteriophage**	**Genome Length (kb)**	**GenBank Accession Number**	**Plaque Morphology**	**Phage Morphology**
Microbacterium foliorum	WestPM	39.693	PP978895	Turbid, with halo and irregular border ~1.5mm diameter	Siphoviral Capsid: 53.5nm Tail: 114.0nm
***Microbacterium* sp. Casco Bay**	PortlandC27	16.73	Duplicate of Pensacola C28	Translucent, with clear and consistent borders ~2mm diameter	Siphoviral Capsid: 39.0nm Tail: 94.4nm
***Microbacterium* sp. Casco Bay**	PensacolaC28	16.769	PP978844	Translucent, with clear and consistent borders ~2mm diameter	Siphoviral Capsid: 39.2nm Tail: 91.4nm
Erythrobacter citreus	Horizon	151.365	TBD	Tiny, pin prick-like plaques ~0.25mm diameter	Prolate Capsid: 52.8nm Tail: 222.9nm

Table 3. Summary of phage genome sizes. Four bacteriophage DNA samples were sent off for genome sequencing at the North Carolina State Genomic Sciences Laboratory. PortlandC27 and PensacolaC28 were determined to be identical. WestPM and PensacolaC28 genomes have been annotated and published in the GenBank database.

DISCUSSION/CONCLUSION

As the most abundant organisms in the ocean, phages play an essential ecological role in microbial balance.[25] The sheer abundance of phage particles in the ocean has been realized using traditional culture-based and metagenomic approaches.[3] Although there has been more research on marine microbial environments in recent years, it is essential to continue to better understand our local bacterial species and their interactions with associated phages. This undergraduate research project tested the hypotheses that the local beach bacterial composition could be determined through eDNA analysis, confirm that local bacterial strains could be used to isolate bacteriophage from local environmental samples, and that locally isolated bacterial strains are more efficient in isolating marine phages than non-local bacterial strains. By the end of this project in summer 2023, the authors were able to conclude that determining the bacterial composition is difficult due to low culturability of present bacteria in environmental samples, and that it is possible to use local marine bacterial isolates to perform phage hunting, however, which bacterial isolate is more suitable to use might require extensive characterization not within the limitations of this research.

Regarding choosing bacterial hosts for phage hunting

Not all bacteria in the environment can be readily cultured in a laboratory setting. The eight bacterial strains isolated from the Pensacola beaches have low abundance relative to the entire metagenomic sample, demonstrating a widely understood culturing bias. *M. oleivorans* and *E.citreus* were selected as representative local hosts for marine phage hunting due to the ubiquitous distribution of the genera and the absence of health risks. These two bacterial strains comprise a very small portion of the metagenomic bacteria composition profile, which might indicate a low number of marine phages capable of infecting these bacterial strains. However, this factor could not be solely responsible, as revealed by the drastic difference in phage hunting success between using *Microbacterium* and *Erythrobacter* species as hosts (**Table 2**). *Microbacterium* strains fall under the Actinobacteria phylum and are gram-positive, whereas *E. citreus* belongs to the phylum Pseudomonadota and is gram-negative; it is possible that these differences could have contributed to the varied results of our phage hunting. Some phage possess specialized endolysin production genes to combat the peptidoglycan layers in bacterial cell walls, such as Vibriophage VPp1, a double-stranded DNA phage specific to *Vibrio parahaemolyticus*.[25] Gram-negative bacteria are much more prevalent in the marine environment,[27] suggesting that marine phages are likely better suited to infect gram-negative hosts; this could cause the discrepancy in phage infection between *M. oleivorans* and *E. citreus*.

Another example of how the complex interactions between phages and bacteria in the marine world regulate the ocean's ecological balance is that some marine bacterial strains exist as lysogens, which contain prophage regions that inhibit infection by lytic phages.[28] While it is possible to identify prophage via sequencing the genome of newly isolated bacterial strains,[29] this poses an additional financial burden. The observed difficulties in using *M. oleivorans* as a host for phage hunting, might suggest the presence of a lysogen in this bacterial strain. Alternatively, it's possible that the inability of *M. oleivorans* to produce plaques from our environmental samples was due to the host's low abundance in the surrounding environment, as determined through bacterial profiling. Environments without a significant abundance of the chosen host bacteria present often yield lower phage hunting success due to decreased pressure for the development of associated phage. Because of this, increased volumes of environmental samples are necessary to increase phage hunting yield and it can be challenging to detect phages from a large volume of environmental samples without concentrating filtrates. While testing if the *M. oleivorans* isolate is lysogenic was beyond the scope of this research, concentrating the sample lysates to increase the success of phage hunting was worth trying.

When comparing another two *Microbacterium* strains, *M. foliorum* NRRL B-24224 [30] isolated from the phyllosphere of grasses and *M.* sp. Casco Bay, only five phages were isolated using *M.* sp. Casco Bay as the host.[24] In contrast, more than 5000 phages have been discovered using *M. foliorum* NRRL B-24224.[30] It is possible that originating from a close-to-marine environment, *M.* sp. Casco Bay might be a better *Microbacterium* strain host for marine phage hunting to be compared with *Erythrobacter* strains. Further research could shed light on this aspect.

Determining the preferred method of phage lysate precipitation

In nature, phages and their hosts compete with many other bacteria and viruses, typically leading to lower concentrations than in pure lab cultures.[31] Phage hunting in the harsh marine environment might require large volumes of environmental samples to be plated for adequate plaque formation. To avoid this, phage precipitation can be performed to concentrate the phage particles into a smaller volume for plaque assay and titering.[20] Practice of such has been used in phage isolation from water samples. [32,33] In this project, the first few phage hunting attempts used various environmental samples from Santa Rosa Island, including dry sand, wet intertidal sand, underwater sand, seagrass, marine animals, and mixed samples. Results indicated no phage presence for further purification when *M. oleivorans* was used as the bacterial host. However, using direct isolation, *E. citreus* exhibited numerous plaques from all sand samples.

Several of the standard phage precipitation methods work through similar mechanisms. A salt solution is added to the lysate, allowing for the separation of water and clustering of phage particles similar to the salting out of proteins.[21] $ZnCl_2$ and PEG solutions were tested to determine efficacy for the concentration of infectious particles for downstream application with environmentally isolated lysates (**Figure 4**). Although $ZnCl_2$ has been used for phage research,[21] it was observed that the PEG-precipitation method was well suited to our phage hunting project, which intended to isolate and preserve phages for long-term usage.

Choosing growth media for phage hunting

This project used the media PYCa and Marine 2216 to culture *M. oleivorans, M. foliorum*, *M.* sp. Casco Bay, and *E. citreus* for phage hunting. After several months of culturing *E. citreus*, it was observed that the bacteria cultures struggled, possibly due to *E. citreus* being an environmentally derived host. The mock seawater medium supports the growth of all four bacterial strains but at a much slower rate than PYCa (*M. oleivorans, M. foliorum,* and *M.* sp. Casco Bay*)* or Marine 2216 (*E. citreus*), and yields smaller colonies. Although mock seawater media was not advantageous for culturing these bacteria, the growth suppression may be a possible stressor for the bacteria cells, leading to a higher phage infection rate. Whether using a sub-optimal medium could increase the potential for phage infection could be further studied in experiments involving marine bacterial strains.

Morphological and Genomic characterization of four phages

The morphology of WestPM, PortlandC27, and PensacolaC28 showed an isometric polyhedral capsid shape with a tail. Horizon images showed a prolate polyhedral shape with a tail. Regarding capsid volume and tail length, Horizon has the most significant capsid volume and extended tail compared to the other three (**Figure 4**). The prolate capsid shape indicates increased capsid capacity or volume.[34] A study has suggested that capsid volume linearly correlates with genome length regarding DNA viruses, including phages,[35] which is consistent with what was observed in this research when comparing the genome sizes of PortlandC27, PensacolaC28, WestPM, and Horizon (**Table 3**). Interestingly, upon observing plaque sizes, Horizon gave rise to miniature pin-prick-like plaques that are difficult to see and photograph. The large genome size of Horizon is consistent with other phages isolated from the bacterial genus *Erythrobacter*[36]. *Erythrobacter* phage vB_EliS-L02 has a genome length similar to Horizon's — around 150,000 base pairs —and the shape of its capsid is similarly prolate.[36]

Summary

In conclusion, wrapped by the summer of 2023, this undergraduate research tested the hypothesis that marine bacterial strains are equally suitable or better for marine phage hunting. The authors explored the feasibility of isolating phages from marine environments using a "start from scratch" approach, comparing the efficiency of isolated *Erythrobacter* and *Microbacterium* species in marine phage hunting. The discovery of the four phages offered an excellent opportunity to examine the phage-host interaction and analyze the genomic information. Annotations of WestPM [22] and PensacolaC28 [23] genomes suggest that these newly isolated marine phages share similar genomic structures with non-marine phages through comparisons using NCBI Blast database [19] and HHPred [37]. The genome annotation of Horizon is in progress, which will open a window for molecular analysis of phage gene functions in the future. For instance, thus far, very few *Erythrobacter* phages have been identified; [36] Horizon may be a valuable model to better understand marine phage and *Erythrobacter* species as hosts. Considering the possibility of lysogeny leading to superinfection resistance, metagenomic analyses may uncover more information on phage diversity and abundance within the marine environment. The lessons learned from this project can help structure future undergraduate marine phage hunting projects.

ACKNOWLEDGMENTS

The authors thank Daniel Russell and Rebecca Garlena of the SEA-PHAGES program for genome assembly, Che-Yu Cheng and Dr. Nai-Chun Lin from the Department of Agricultural Chemistry at National Taiwan University for performing electron microscopy, Leila Harris, Grace Worley and Trevor McMullen for establishing a marine bacteria culturing guide, Brian Tarbox, Emily Savage, and Southern Maine Community College for their collaboration on phage hunting as well as providing the bacterial host *M.* sp. Casco Bay, SEA-PHAGES for supplying our bacterial host *M. foliorum* NRRL B-24224, and the Office of Undergraduate Research at UWF for awarding Charles West, Brittany Yencho, and Conor Flannigan undergraduate research awards and travel awards. We would especially like to thank our mentor, Dr. Hui-Min Chung, for her commitment, support, and patience as she guided us through our research. Hui-Min Chung was supported by the NSF grant 1711842.

REFERENCES

1.Clokie, M. R., Millard, A. D., Letarov, A. V., & Heaphy, S. (2011) Phages in nature. *Bacteriophage*, *1*(1), 31–45. *https://doi.org/10.4161/bact.1.1.14942*

2.Choudhury, R., Panda, S., & Singh, D. (2012) Emergence and dissemination of antibiotic resistance: A global problem. *Indian Journal of Medical Microbiology*, *30*(4), 384–390. *https://doi.org/10.4103/0255-0857.103756*

3. Breitbart, M., Bonnain, C., Malki, K., & Sawaya, N. A. (2018) Phage puppet masters of the marine microbial realm. *Nature Microbiology*, *3*(7), 754–766. *https://doi.org/10.1038/s41564-018-0166-y*

4. Drulis-Kawa, Z., Majkowska-Skrobek, G., & Maciejewska, B. (2015) Bacteriophages and Phage-Derived Proteins – Application Approaches. *Current Medicinal Chemistry*, *22*(14), 1757–1773. *https://doi.org/10.2174/0929867322666150209152851*

5. Ali, Y., Inusa, I., Sanghvi, G., Mandaliya, V. B., & Bishoyi, A. K. (2023) The current status of phage therapy and its advancement towards establishing standard antimicrobials for combating multi drug-resistant bacterial pathogens. *Microbial Pathogenesis*, *181*, 106199. *https://doi.org/10.1016/j.micpath.2023.106199*

6. Bolsan, A. C., Sampaio, G. V., Rodrigues, H. C., Silva De Souza, S., Edwiges, T., Celant De Prá, M., & Gabiatti, N. C. (2024) Phage formulations and delivery strategies: Unleashing the potential against antibiotic-resistant bacteria. *Microbiological Research*, *282*, 127662. *https://doi.org/10.1016/j.micres.2024.127662*

7. Bhatt, P., Engel, B. A., Reuhs, M., & Simsek, H. (2023) Cyanophage technology in removal of cyanobacteria mediated harmful algal blooms: A novel and eco-friendly method. *Chemosphere*, *315*, 137769. *https://doi.org/10.1016/j.chemosphere.2023.137769*

8. Zhu, X., Li, Z., Tong, Y., Chen, L., Sun, T., & Zhang, W. (2023) From natural to artificial cyanophages: Current progress and application prospects. *Environmental Research*, *223*, 115428. *https://doi.org/10.1016/j.envres.2023.115428*

9.Baharum, S. N., Beng, E. K., & Mokhtar, M. A. A. (2010) Marine Microorganisms: Potential Application and Challenges. *Journal of Biological Sciences*, *10*(6), 555–564. *https://doi.org/10.3923/jbs.2010.555.564*

10. Ameen, F., AlNadhari, S., & Al-Homaidan AA. Marine microorganisms as an untapped source of bioactive compounds. *Saudi Journal of Biological Sciences,* 28(1), 224-23. https://doi.org/10.1016/j.sjbs.2020.09.052

11. Tuttle, M. J., & Buchan, A. (2020) Lysogeny in the oceans: Lessons from cultivated model systems and a reanalysis of its prevalence. *Environmental Microbiology*, *22*(12), 4919–4933. *https://doi.org/10.1111/1462-2920.15233*

12. Rodrigues, C. J. C., & de Carvalho, C. C. R. (2022) Cultivating marine bacteria under laboratory conditions: Overcoming the "unculturable" dogma. *Frontiers in Bioengineering and Biotechnology, 10,* 3389. *https://doi.org/10.3389/fbioe.2022.964589*

13. Rey-Velasco, X., Deulofeu-Capo, O., Sanz-Sáez, I., Cardelús, C., Ferrera, I., Gasol, J. M., & Sánchez, O. (2023) Expanding success in the isolation of abundant marine bacteria after reduction in grazing and viral pressure and increase in nutrient availability. *Microbiology Spectrum*, *11*(5), e00890-23. *https://doi.org/10.1128/spectrum.00890-23*

14. Watkins, K. 2022. The Effect of Prophage Integration in a New Bacterial Host. ResearchGate *10.13140/RG.2.2.28243.39203*

15. Poxleitner, M., Pope, W., Jacobs-Sera, D., Sivanathan, V., & Hatfull, G. (2018) Phage discovery guide. *Howard Hughes Medical Institute, Chevy Chase, MD. https://seaphagesphagediscoveryguide.helpdocsonline.com/copyright* (accessed 28 Feb 2024)

16. Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 5029. *https://doi.org/10.1038/s41467-019-13036-1*
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990) Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410. *https://doi.org/10.1016/S0022-2836(05)80360-2*
18. Bradley, D. E. (1962) A Study of the Negative Staining Process. *Journal of General Microbiology*, *29*(3), 503–516. *https://doi.org/10.1099/00221287-29-3-503*
19. Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Sherry, S. T. (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. *https://doi.org/10.1093/nar/gkab1112*
20. Fouladvand, F., Bemani, P., Mohammadi, M., Amini, R., & Azizi Jalilian, F. (2020) A Review of the Methods for Concentrating M13 Phage. *Journal of Applied Biotechnology Reports*, *7*(1) *https://doi.org/10.30491/jabr.2020.105916*
21. Czajkowski, R., Ozymko, Z., & Lojkowska, E. (2016) Application of zinc chloride precipitation method for rapid isolation and concentration of infectious Pectobacterium spp. And Dickeya spp. Lytic bacteriophages from surface water and plant and soil extracts. *Folia Microbiologica*, *61*(1), 29–33. *https://doi.org/10.1007/s12223-015-0411-1*
22. West, C.J., Yencho, B.C., Brown, A.J., Flannigan, C.R., Chung, H.M. (2025) Genome sequence of WestPM, a phage infecting *Microbacterium foliorum* isolated from beach environmental samples. microPublication Biology. *10.17912/micropub.biology.001395*
23. Girard, L. R., Cousins, S. R., Heald, A. C., Tanzey, D. A., Omo, E. M., Flannigan, C. R., Chung, H.-M., Tarbox, B. P., & Savage, E. F. (2025) Genome sequence of bacteriophage PensacolaC28 isolated using *Microbacterium sp.* Casco Bay. *Microbiology Resource Announcements*, e01146-24. *https://doi.org/10.1128/mra.01146-24*
24. Russell, D. A., & Hatfull, G. F. (2017) PhagesDB: The actinobacteriophage database. *Bioinformatics*, *33*(5), 784–786. *https://doi.org/10.1093/bioinformatics/btw711*
25. Sun, M., Lin, H., Wang, M., Liu, Y., Liu, Z., Meng, X., Jiang, Y., Wang, D., Li, Y., & Xia, J. (2018) Characterization and complete genome of the marine Pseudoalteromonas phage PH103, isolated from the Yellow Sea, China. Marine Genomics, 42, 67–70. *https://doi.org/10.1016/j.margen.2018.04.003*
26. Li, M., Jin, Y., Lin, H., Wang, J., & Jiang, X. (2018) Complete Genome of a Novel Lytic *Vibrio parahaemolyticus* Phage VPp1 and Characterization of Its Endolysin for Antibacterial Activities. *Journal of Food Protection*, *81*(7), 1117-1125. *https://doi.org/10.4315/0362-028X.JFP-17-278*
27. Gnanagobal, H. & Santander, J. (2022) Host–Pathogen Interactions of Marine Gram-Positive Bacteria. *Biology 11*(9), 1316. *https://doi.org/10.3390/biology11091316*
28. Paul, J. H. (2008) Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *The ISME Journal, 2,* 579-589
29. Gan, R., Zhou, F., Si, Y., Yang, H., Chen, C., Ren, C., Wu, J., & Zhang, F. (2022) DBSCAN-SWA: An Integrated Tool for Rapid Prophage Detection and Annotation. *Frontiers in Genetics*, *13. https://doi.org/10.3389/fgene.2022.885048*
30. Russell, D. A., Garlena, R. A., & Hatfull, G. F. (2019) Complete Genome Sequence of Microbacterium foliorum NRRL B-24224, a Host for Bacteriophage Discovery. *Microbiology Resource Announcements*, *8*(5), e01467-18. *https://doi.org/10.1128/MRA.01467-18*
31. Salazar, G., & Sunagawa, S. (2017) Marine microbial diversity. *Current Biology*, *27*(11), R489–R494. *https://doi.org/10.1016/j.cub.2017.01.017*
32. Wommack, K.E., Hill, R. T., Colwell, R.R. (1995) A simple method for the concentration of viruses from natural water samples. *Journal of Microbiological Methods, 22(1)*, 57-67. https://doi.org/10.1016/0167-7012(94)00064-E
33. Bonilla, N., Rojas, M. I., Flores Cruz, G. N., Hung, S., Rohwer, F., Barr, J. J. (2016) Phage on tap- a quick and efficient protocol for the preparation of bacteriophage laboratory stocks. *PeerJ* 4:e2261. *https://doi.org/10.7717/peerj.2261*
34. Stone, N. P., Demo, G., Agnello, E., & Kelch, B. A. (2019) Principles for enhancing virus capsid capacity and stability from a thermophilic virus capsid structure. *Nature Communications*, *10*(1), 4471. *https://doi.org/10.1038/s41467-019-12341-z*
35. Chaudhari, H. V., Inamdar, M. M., & Kondabagil, K. (2021) Scaling relation between genome length and particle size of viruses provides insights into viral life history. *iScience*, *24*(5), 102452. *https://doi.org/10.1016/j.isci.2021.102452*
36. Li, X., Guo, R., Zou, X., Yao, Y., & Lu, L. (2022) The First Cbk-Like Phage Infecting Erythrobacter, Representing a Novel Siphoviral Genus. *Frontiers in Microbiology*, *13*, 861793. *https://doi.org/10.3389/fmicb.2022.861793*
37. Soding, J., Biegert, A., & Lupas, A. N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(Web Server), W244–W248. *https://doi.org/10.1093/nar/gki408*

ABOUT STUDENT AUTHORS

Brittany Yencho graduated from the University of West Florida in 2022 with a Bachelor's degree in Marine Biology. She is now attending Louisiana State University's School of Veterinary Medicine and hopes to use her veterinary degree to provide much-needed assistance and education to rural communities.

Charles West graduated from the University of West Florida in May 2023 with a Bachelor of Science in Biomedical Sciences. He is currently a Graduate Student Trainee at the University of Alabama at Birmingham. He plans to continue his education and research in the microbiology field.

Andrew Brown graduated from the University of West Florida in 2022 with a Bachelor's degree in Marine Biology. He is an ecotoxicologist for the United States Environmental Protection Agency in Gulf Breeze, Florida.

Conor Flannigan graduated from the University of West Florida in 2023 with a Bachelor's degree in Biology. He works in North Carolina at the Duke University Medical Center in the Department of Surgery. He plans to continue his education at Duke University as a Graduate Student Trainee.

PRESS SUMMARY

Bacteriophages, or viruses that infect bacteria, are among the most prevalent biological entities in the world. Many of these phages work to hijack the bacterial host's replication machinery to replicate, leading to host cell death (lytic cycle) or co-existing with the host (lysogenic cycle). The diverse genes phages use to modulate their host's fitness may be relevant in developing antimicrobials, environmental mitigation tactics, and other crucial research spheres. In this article, resident bacteria of Pensacola Beach, Florida, were characterized and cultured, allowing the isolation of four marine phages from environmental samples such as sand, ocean water, and crustacean tissue. The genomes of all four phages were sequenced, and morphological and phylogenetic information has been revealed.

Uninformed Consent? The Impact of Reading Level, Format, and Interactivity of Consent Forms on Participant Comprehension

Kaitlyn A. Carr & Clare Conry-Murray*

Department of Psychology, Saint Joseph's University, Philadelphia, PA

https://doi.org/10.33697/ajur.2025.139

*Student: Kaitlyncarr345@gmail.com**
Mentor: Clareconrymurray@gmail.com

ABSTRACT
The present study examined participants' comprehension of consent form information based on the consent forms' reading level, format, and level of interactivity. Our sample consisted of 228 adult English speakers who were randomly assigned to one of eight groups. Each group saw consent forms with a different combination of reading level, format, and interactivity of questions. All participants were asked to answer 12 multiple-choice questions about two consent forms (framed as two studies). Participants read and self-reported whether they read the form or not. Asking participants questions immediately after exposure to the information in the consent form improved comprehension compared to asking all comprehension questions after reading the entire form. Reading level and bulleted format did not improve comprehension significantly. Furthermore, participants were more likely to report that they read the consent form if they were given interactive questions about the form. Results suggest that interactive questions can be an effective method for improving participant comprehension of the purpose, risks, and benefits of the study or procedure in which they plan to participate.

KEYWORDS
Comprehension; Informed Consent; Format; Reading Level; Interactivity; Participants; Ethics in Research; Consent Form

INTRODUCTION

Research with human participants often requires informed consent. However, when a participant is presented with a consent form prior to participating in a study, they make the personal decision to read it or ignore it, leaving them uninformed. Even when participants read the consent form, it may not be fully understood by research participants. This study examined several factors that may assist in participants' reading and understanding of consent forms: reading levels, bulleted vs paragraph formats, and the interactivity of questions. If we can identify which factors promote reading comprehension, researchers may be able to save participants' time and promote more ethical treatment of participants.

Past research showed that even when participants felt satisfied with the informed consent process, many were unable to identify the risks, benefits, or purpose of the study in which they participated.[1–3] On the other hand, some participants do not read or skim the information in consent forms. Participants may do this when they believe that the forms are too long or because the key points do not stand out in some way.[4] Furthermore, consent forms are becoming increasingly more difficult to understand, as even pediatric forms are typically written at a 10th-grade reading level.[5] This makes both choosing to read the consent form and understanding the information provided difficult, especially for low-literacy populations, the elderly, the cognitively impaired, and patients with certain psychiatric conditions, who are often the focus of clinical research.[6–9]

The difference in attention given to consent forms with different formatting and timing of when comprehension questions are asked could be attributed to habituation. The theory of habituation suggests that people pay less attention to a stimulus that is often presented to them because they identify it as information they have already mentally processed.[10] Participants may become habituated to the material presented in most consent forms due to frequent exposure to "terms and conditions" forms in today's digital age. These "terms and conditions" forms may function similarly to consent forms, as both are required to continue with a desired task. Therefore, participants may comprehend more information when it is given in bullet point form because it draws attention to changes made in consent forms or to what makes this consent form different from others. The same theory could be applied to interactive questions in the way that the questions break the habituation process by directing the participants' attention back to the information that is necessary for them to know and understand.[10–13] A study done by Geier and colleagues also

indicated that participants could pay more attention to the content presented in interactive consent forms, as participants spent more time on average with this version.[12]

The current literature on consent forms has tested the reading level of consent forms and their impact on participant comprehension.[1, 9] Results suggest that consent forms written at a lower reading level are easier for participants to read.[14] However, while one study indicated that forms with lower reading levels also improved participant comprehension,[9] another suggested that reading level showed no difference in participant comprehension.[1] Furthermore, a study done by Joffe[3] and colleagues showed no difference in participant comprehension between the group presented with an easier reading level and a more difficult reading level; however, most participants also reported using outside resources to aid in their comprehension, indicating that many participants are unable to comprehend material in consent forms on their own.[3] This suggests that the reading level of consent forms and their impact on participant comprehension should be studied more in-depth.

Furthermore, several studies have tested the effects of consent form format on participant comprehension and found that shorter, more precise forms that highlighted key points were more likely to be read and understood.[9, 15–19] The use of bullet points was particularly helpful to some groups that were tested on their consent form comprehension.[15, 18–19] However, there is evidence to contrast this finding, in which formatting had no impact on participant comprehension.[12, 20–21] Therefore, concise bullet point formatting of consent forms may make significant differences in participants' comprehension of the information, compared to more lengthy paragraphs.

Another component that should be considered in participants' abilities to read and understand information in consent forms is the level of interactivity they have with the material.[11, 12] Flory and Emanuel's[11] research found that participants had higher levels of comprehension when consent forms offered a comprehension test or feedback from a healthcare professional.[11] Moreover, the level of interactivity and engagement with material on the consent form was positively correlated with the level of comprehension and was more likely to get participants to read the form.[12] Therefore, it is also possible that asking comprehension questions immediately after exposure to important information in consent forms may improve participants' likelihood of reading the forms and their comprehension of them. Although this could be a result of short-term memorization, given that long-term memory for the information in consent forms is unneeded, comprehension immediately after presentation would still be useful for informed consent.

Ultimately, the current literature on consent forms has tested the readability of forms based on reading level, formatting (bullet points, spacing, bolding, *etc.*), and the level of interactivity with the material, and each of these components' impact on participant comprehension. The findings of past studies in these areas are variable, and therefore not conclusive.[1–21] Moreover, there is yet to be a study that has tested interactions among all these conditions on participant comprehension. Developing a new consent form template or criteria may aid in the advancement of ethical research practices and lead to more informed participants. Therefore, the current study examined whether different combinations of reading level, format, and level of interactivity affect consent form comprehension. In other words, which modifications to consent forms make participants more informed? In the current study, participants were placed into one of eight different groups. Each group had a different combination of these three variables. Consistent with previous research,[3, 9] it was hypothesized that participants who read the lower reading level consent forms would have better comprehension scores than participants who read the higher reading level consent forms. Based on past literature, it was predicted that participants in groups that read the bullet point format consent forms would have higher comprehension scores and would be more likely to self-report actually reading the form than those who received the paragraph format consent forms.[9, 15–19] Previous findings also led us to predict a main effect for the level of interactivity, in which participants in the interactive questions group would have higher comprehension scores and would be more likely to self-report actually reading the form than the group without interactive questions.[10–13] We also expected an interaction between format and reading level, in which lower reading level consent forms in bullet point format would show higher scores of comprehension. Lastly, we hypothesized an interaction between the interaction group and format. Specifically, we expected the groups that read the bullet point consent forms with interactive questions to have the highest comprehension scores out of all the groups.

METHODS AND PROCEDURE

Institutional Review Board approval from the university where the study was conducted was obtained for the project title [2174108-1], and all hypotheses and methods were pre-registered in the OSF database prior to collecting data. We recruited 324 English speakers over the age of 18 through social media and the snowball method of convenience sampling. Although we pre-registered a larger sample size, we had to stop collecting data before the planned number of participants was met due to a deadline at the university where the study was conducted. After removing all test trials and participants who were underage or who completed less than half of the survey, the final sample consisted of 228 participants, including 36 men (15.79%), 191 women

(83.77%), and one (0.44%) participant that was either non-binary or another gender. The mean age for the final sample was 43.54 years. In terms of racial and ethnic demographics, 190 (83.33%) participants identified as white, 25 (10.96%) identified as Black or African American, six (2.63%) identified as Hispanic or part of the Latinx community, four (1.75%) identified as Asian or Asian American, and three (1.32%) identified as some other racial or ethnic identity. In terms of education, 59 (25.88%) had a high school diploma or GED equivalent, 22 (9.65%) had an associate's degree, 79 (34.65%) had a bachelor's degree, 48 (21.05%) had a master's degree, and 19 (8.33%) had a doctorate or professional degree. There was also one participant (0.44%) with a cosmetology license.

To hide the true purpose of our study and analyze what participants do naturally when confronted with consent forms, the participants were told that this study would measure their understanding and comprehension of different texts, but not that the texts they needed to comprehend were the consent forms. Instead, they were told that they would complete several shorter studies so that they would believe that each consent form was for a new study. After participants gave consent to take the survey, they were randomly assigned by Qualtrics to one of eight groups. Each group contained a different interaction of our independent variables using a factorial, between groups design 2 (reading level; high or low) x 2 (format; paragraphs or bullet points) x 2 (level of interactivity; questions during or after reading consent form), which resulted in eight conditions. Each condition had two trials with two different consent forms. The content of the two consent forms given to each participant differed, but both consent forms contained the same key qualities (i.e., in terms of designated reading level, format, and interactivity).

The reading level of the consent forms was determined by the Flesch Reading Ease Score (FRES), which was calculated using the number of words, sentences, average syllables per word, and average words per sentence. FRES that range from zero to 50 are considered to be college-level or above and FRES that range from 50 to 100 are considered to be below college level. For the current study, reading scores that were at the college level or above were marked as high, and reading scores below the college level were marked as low. Conditions that contained consent forms written at lower reading levels had FRES of 50.1-66.6. Conditions that contained consent forms written at higher reading levels had FRES of 28.1-40.9. FRES scores within reading level groups were made as close together as possible. The exact scores for each trial in each condition and how they were calculated can be found in the **appendix**. In the group with interactive questions, the questions were presented directly after the part of the consent form that gave the answer to that specific question, as opposed to answering a set of questions after reading the entire consent form. All manipulated consent forms from trial one followed the consent form template recommended by the National Academy of Neuropsychology,[22] and all manipulated consent forms from trial two followed the consent form template recommended by the Institutional Review Board at the university where the study was being conducted.[23] Both "fake" consent forms were labeled with the trial number and the title "Consent to participate in research," to draw attention to the fact that it was a new consent form.

All participants answered questions about their demographics. Next, depending on which group they were randomly assigned to, they either answered comprehension questions about the consent form during or after reading the first manipulated consent form. They were asked six multiple-choice, comprehension questions per trial, and one question per trial asking if they read the consent form. Comprehension questions included information about the purpose, benefits, risks, withdrawal, expected time to completion, and confidentiality. As a distractor from the consent forms, a textbook paragraph related to neuropsychology was then displayed, followed by a shortened, true-or-false reading comprehension test on that textbook paragraph. Participants were then directed to a second trial and were shown a new manipulated consent form about forensic psychology and asked to answer comprehension questions about the consent form from trial two. They then took another shortened, reading comprehension section on the forensic psychology textbook paragraph as a distraction. By creating a second trial, we were able to gain more data as well as control for the subject of each consent form. For example, participants who strongly dislike neuropsychology may answer more questions incorrectly on the first set of questions because they dislike the subject they are about to answer test questions on. Adding a second trial with different subject matter controlled for this effect. Moreover, the overall performance of each participant was less likely to be due to the specific consent forms used in our study or due to the templates recommended by each institution because the consent forms and templates were diverse in their subject matter and template recommendations but contained the same consent form qualities. Overall, the presence of two trials as opposed to one added validity to the experiment. Before ending the survey, participants were debriefed. All experimental materials can be found in the **appendix**.

The dependent variable, which was the comprehension of information, was operationalized by the number of comprehension questions that each participant answered correctly. One point was given for each correct response chosen. For data analysis, the six comprehension questions from each of the two forms were combined to make a composite score for each participant with a maximum range of zero to 12. Four multiple-choice options were given for each comprehension question, including "I don't know/ I am not sure". Any question that was left blank or questions that participants said they did not know or were not sure

about were recorded as incorrect. For true or false comprehension questions, a "Neither true nor false" option was also given to make the number of options for each question equal. Whether participants self-reported actually reading the form or not was also measured.

All data was analyzed on SPSS. A univariate ANOVA was used to analyze data relating to comprehension scores and a chi-square test of independence was used to analyze the data relating to whether participants reported reading the forms.

RESULTS

The 2 (reading level; high or low) x 2 (format; paragraphs or bullet points) x 2 (level of interactivity; questions during or after reading consent form) ANOVA revealed the hypothesized main effect for interactivity of questions on comprehension score, $F(1, 220) = 46.50$, $p < 0.001$, $\eta_p^2 = .174$. Thus, those who received interactive questions scored higher ($M = 9.71$, $SD = 2.43$) than those who received all comprehension questions after reading the entire form ($M = 7.36$, $SD = 2.71$), as we expected. **Figure 1** shows the main effect for interactivity of questions. We also hypothesized a main effect for reading level on total comprehension score. However, our results did not support the hypotheses for reading level, $F(1, 220) = .04$, $p = .846$, $\eta_p^2 = .000$, and format, $F(1, 220) = .04$, $p = .850$, $\eta_p^2 = .000$. Participants who viewed consent forms written at a higher reading level received higher comprehension scores ($M = 8.63$, $SD = 2.90$) than the group who viewed consent forms written at a lower reading level ($M = 8.45$, $SD = 2.77$), which is the opposite of what we hypothesized, but was not statistically significant. For format, participants who viewed bullet point consent forms received lower comprehension scores ($M = 8.50$, $SD = 2.85$) than participants who viewed paragraph consent forms ($M = 8.59$, $SD = 2.82$), but this finding was also not significant.

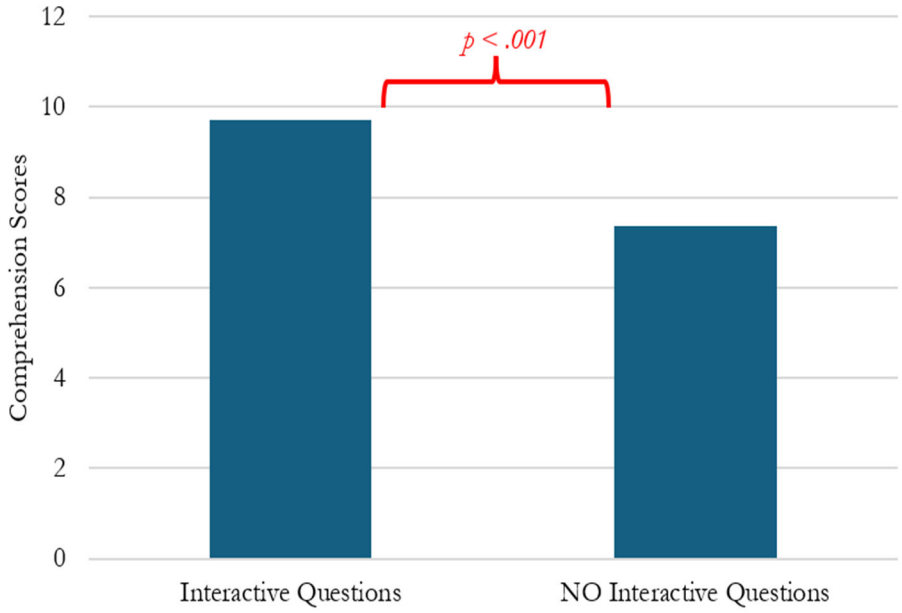


Figure 1. Interactivity of questions on comprehension score.

Furthermore, we hypothesized an interaction between reading level and format, $F(1, 220) = .12$, $p = .734$, $\eta_p^2 = .001$, and between format and interactivity of questions, $F(1, 220) = 2.08$, $p = .151$, $\eta_p^2 = .009$. However, neither of these interactions were significant. Group means and standard deviations can be found in **Table 1** and **Figure 2.**

Reading Level	Format	Interactivity	Comprehension Score
Low	Paragraph	Yes	9.93(2.65)
	Paragraph	No	7.40(2.57)
	Bullet Point	Yes	9.54(2.56)
	Bullet Point	No	7.62(2.72)
High	Paragraph	Yes	10.42(1.78)
	Paragraph	No	6.74(2.65)
	Bullet Point	Yes	9.41(2.67)
	Bullet Point	No	7.64(2.93)

**M(SD) = Comprehension Scores*

Table 1. Mean comprehension scores based on reading level, format, and interactivity of questions.

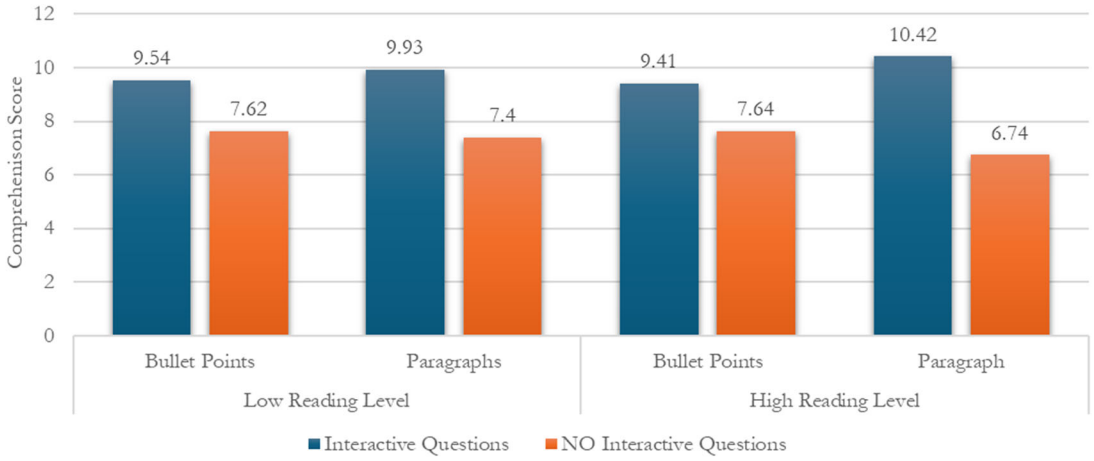


Figure 2. Mean comprehension score by version of consent form.

In terms of the self-reported reading of the consent forms, we hypothesized main effects for interactivity of questions and format. Although we preregistered the data analysis for this dependent variable as an ANOVA, we used a chi-square test of independence, which is more appropriate than an ANOVA, as the dependent variable consisted of a binary "yes" or "no" response. Therefore, we used a chi-square test of independence to examine the relation between the level of interactivity and whether participants self-reported reading the consent forms. The relation between these variables was significant, $\chi 2$ (2) = 11.44, p = .003, *ES* = .116. Those who answered interactive questions were more likely than those who answered all questions after reading the consent form to self-report actually reading the form. While 53% of those who answered interactive questions reported that they did read the form, only 47% of those who answered all comprehension questions after reading the entire form reported that they actually read the form.

Contrary to our expectations, there was not a significant main effect for format on self-reported reading of the consent forms, $\chi 2$ (2) = 2.14, p = .343, *ES* = .061. Those who viewed the paragraph consent form were more likely than those who viewed the bullet point consent form to self-report actually reading the form. Out of those who viewed the paragraph form, 50.13% reported that they did read the form, while only 49.87% of those who viewed the bullet point consent form reported that they actually read the form, which is not what we expected, but was also not statistically significant. A graph with all percentages of self-reported reading of the consent forms by group can be seen in **Figure 3.**

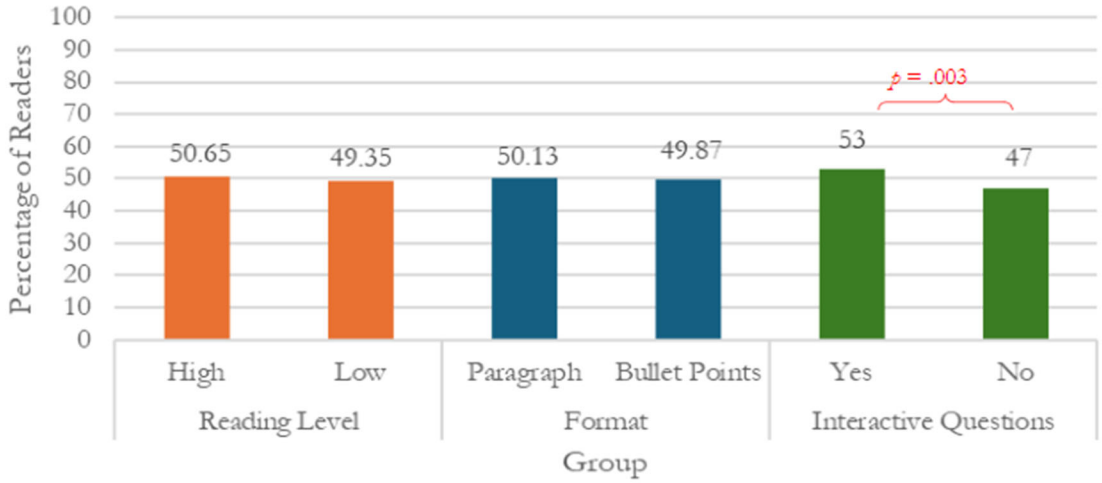


Figure 3. Self-reported reading of the consent forms based on group.

DISCUSSION

This study demonstrated that asking participants questions immediately after exposure to the information that answers the given question made participants more likely to self-report reading the form and resulted in better comprehension of that information when compared to answering all comprehension questions after reading the entire form. However, there were no significant differences in self-reported reading and comprehension between those who received paragraph consent forms and those who received bullet point consent forms, nor were there any significant differences in comprehension based on the reading level at which the consent forms were written.

Our finding that asking interactive questions about the consent form resulted in more self-reported reading and better comprehension was consistent with past studies that tested the interactivity of consent forms on participant comprehension.[11, 12] This is most likely because the interactive questions acted as a reminder to the participant to read the form. If they did not read that section of the form the first time, they may have felt inclined or as though they had more of a responsibility to go back and read it again in order to answer the question correctly. It is also possible that these questions highlighted what was important or different in the consent form, which was also shown to be helpful in a past study.[4] This finding also corresponds with the theory of habituation in the way that participants become habituated to the material in consent forms and pay less attention to them. The interactive questions broke their habituation by highlighting what was different or what was most important for the participants to know.[10–13] Moreover, there could have been a forward testing effect in our study, in which interactive questions facilitated the later learning and retention of consent form information.[25] Past literature has shown that when tested on smaller quantities of information in the interim, the quality of students' learning and comprehension was better and they were more likely to score higher on cumulative tests later on.[25] Therefore, future research should focus on the impact of interactive questions on long-term memory of consent forms as well as how interactive questions on one consent form may aid in the comprehension of consent forms read in the future that do not contain comprehension questions.

Our study confirmed our hypothesis and the findings of past literature on the interactivity of consent forms and their effect on participant comprehension. However, it is important to note that the average comprehension score for participants who received the interactive questions was not a perfect score, meaning that even when participants had the chance to go back and look at the information, they did not understand it. Therefore, future research should focus on what is confusing and what needs to be made clearer in consent forms.

However, the reading level at which the consent forms were written had no significant impact on comprehension scores, which supports the findings of some other studies,[1, 14] but conflicts with our hypothesis and with the findings of another study.[9] This could be due to the fact that most of our participants had at least a high school diploma. Interestingly, the participants given the consent forms written at a higher reading level had slightly higher comprehension scores than those who were given consent forms written at a lower reading level. Given that our results showed very little difference between groups with a slight benefit to those who received the high reading level consent forms indicates that researchers may not need to worry about the reading level at which they write their consent forms. However, the fact that many did not comprehend either consent form regardless of reading level indicates that consent forms may be worded in a confusing way and should be rewritten with even more detail to

relay crucial information to research participants. Future research could focus on identifying which sections of consent forms are most confusing or are most commonly misinterpreted by participants.

Furthermore, no significant differences were found between self-reported reading and comprehension scores of those who read consent forms formatted in paragraphs compared to those who read consent forms formatted with bullet points. This finding did not support our hypothesis and was inconsistent with past literature.[9, 15–19] However, some other studies found similar results to ours.[12, 20–21] Although the bullet points were more concrete and drew attention to what was most important to understand, they also made the consent form look longer, which could have deterred participants from wanting to read the whole form. Past literature has shown that expected cognitive load may result in the same feelings that occur when actually completing that task.[24] Therefore, longer consent forms may also make participants think that the information will be difficult to understand, and make them more likely to skip the form entirely.

There are several limitations to our study. The first is that our findings are not representative of the entire population. Our sample consisted mostly of white, educated, middle-aged women from the United States. Future research should seek a more diverse sample so that it can be generalized to a wider population. This is especially important when measuring the effect of the reading level of the consent forms on participant comprehension, as a lower literacy population or populations with lower levels of education may have scored lower on comprehension questions if they were given a consent form that was written at a higher reading level. Next, although our participants were randomly assigned, we used the snowball sampling method of collecting data. People known to the researchers or those who volunteer for a study may be different from a completely random sample. Moreover, our sample size was somewhat small, and the results could have been different if we obtained a larger sample size. Therefore, future research should also aim for a random sample that is larger than the sample size of this study. Our fourth limitation involves our methods. We manipulated the reading level by making the low reading level anything below a college level and we wrote the high reading level consent forms at a college level or above. This may not be big enough of a difference in Flesch Reading Ease Scores to properly measure this variable. Future researchers should try to create larger differences in consent form reading levels to check the finding that reading level has no impact on participant comprehension.

In conclusion, including comprehension questions in consent forms may be a good way to gauge a participant's or patient's understanding of the purpose, risks, and benefits of studies or procedures in which they are planning to take part. If participants are unable to answer these questions even when they have the chance to look back at the consent form, they are most likely not giving truly informed consent and should be followed up by a conversation with the researcher or health care provider. This practice may make participants more informed, and ultimately research and clinical procedures more ethical.

REFERENCES

1. Davis, T. C., Berkel, H. J., Holcombe, R. F., Pramanik, S., and Divers, S. G. (1998) Informed consent for clinical trials: a comparative study of standard versus simplified forms. *JNCI: Journal of the National Cancer Institute*, *90*(9), 668–674. *https://doi.org/10.1093/jnci/90.9.668*
2. Fortun, P., West, J., Chalkley, L., Shonde, A., and Hawkey, C. (2008) Recall of informed consent information by healthy volunteers in clinical trials. *QJM: An International Journal of Medicine*, *101*(8), 625–629. *https://doi.org/10.1093/qjmed/hcn067*
3. Joffe, S., Cook, E. F., Cleary, P. D., Clark, J. W., and Weeks, J. C. (2001) Quality of informed consent in cancer clinical trials: a cross-sectional survey. *The Lancet*, *358*(9295), 1772–1777. *https://doi.org/10.1016/S0140-6736(01)06805-2*
4. Perrault, E. K., and Keating, D. M. (2018) Seeking ways to inform the uninformed: Improving the informed consent process in online social science research. *Journal of Empirical Research on Human Research Ethics*, *13*(1), 50–60. *https://doi.org/10.1177/1556264617738846*
5. Larson, E., Foe, G., and Lally, R. (2015) Reading Level and Length of Written Research Consent Forms. *Clinical and Translational Science*, *8*(4), 355–356. *https://doi.org/10.1111/cts.12253*
6. Dunn, L. B., Lindamer, L. A., Palmer, B. W., Golshan, S., Schneiderman, L. J., and Jeste, D. V. (2002) Improving Understanding of Research Consent in Middle-Aged and Elderly Patients With Psychotic Disorders. *The American Journal of Geriatric Psychiatry*, *10*(2), 142–150. *https://doi.org/10.1097/00019442-200203000-00005*
7. Jeste, D. V., Palmer, B. W., Golshan, S., Eyler, L. T., Dunn, L. B., Meeks, T., Glorioso, D., Fellows, I., Kraemer, H., and Appelbaum, P. S. (2009) Multimedia Consent for Research in People With Schizophrenia and Normal Subjects: a Randomized Controlled Trial. *Schizophrenia Bulletin*, *35*(4), 719–729. *https://doi.org/10.1093/schbul/sbm148*
8. Kripalani, S., Bengtzen, R., Henderson, L. E., and Jacobson, T. A. (2008) Clinical Research in Low-Literacy Populations: Using Teach-Back to Assess Comprehension of Informed Consent and Privacy Information. *IRB*, *30*(2), 13–19.
9. Young, D. R., Hooker, D. T., and Freeberg, F. E. (1990) Informed consent documents: increasing comprehension by reducing reading level. *IRB: Ethics & Human Research*, *12*(3), 1–5. *https://doi.org/10.2307/3564107*

10. Bravo-Lillo, C., Cranor, L., Komanduri, S., Schechter, S., and Sleeper, M. (2014) Harder to ignore? Revisiting pop-up fatigue and approaches to prevent it. *Symposium On Usable Privacy and Security*, 105–111.
11. Flory, J., and Emanuel, E. (2004) Interventions to improve research participants' understanding in informed consent for research: a systematic review. *Jama*, *292*(13), 1593–1601. *https://doi.org/10.1001/jama.292.13.1593*
12. Geier, C., Adams, R. B., Mitchell, K. M., and Holtz, B. E. (2021) Informed consent for online research—is anybody reading?: assessing comprehension and individual differences in readings of digital consent forms. *Journal of Empirical Research on Human Research Ethics*, *16*(3), 154–164. *https://doi.org/10.1177/15562646211020160*
13. Lindegren, D., Karegar, F., Kane, B., and Pettersson, J. S. (2021) An evaluation of three designs to engage users when providing their consent on smartphones. *Behaviour & Information Technology*, *40*(4), 398–414. *https://doi.org/10.1080/0144929X.2019.1697898*
14. Hadden, K. B., Prince, L. Y., Moore, T. D., James, L. P., Holland, J. R., and Trudeau, C. R. (2017) Improving readability of informed consents for research at an academic medical institution. *Journal of Clinical and Translational Science*, *1*(6), 361–365. *https://doi.org/10.1017/cts.2017.312*
15. Kim, E. J., and Kim, S. H. (2015) Simplification improves understanding of informed consent information in clinical trials regardless of health literacy level. *Clinical Trials*, *12*(3), 232–236. *https://doi.org/10.1177/1740774515571139*
16. LoVerde, M. E., Prochazka, A. V., and Byyny, R. L. (1989) Research consent forms: continued unreadability and increasing length. *Journal of General Internal Medicine : JGIM*, *4*(5), 410–412. *https://doi.org/10.1007/BF02599693*
17. Mann, T. (1994) Informed consent for psychological research: Do subjects comprehend consent forms and understand their legal rights? *Psychological Science, 5*(3), 140–143. *https://doi.org/10.1111/j.1467-9280.1994.tb00650.x*
18. Perrault, E. K., and Nazione, S. A. (2016) Informed consent—Uninformed participants: Shortcomings of online social science consent forms and recommendations for improvement. *Journal of Empirical Research on Human Research Ethics, 11*(3), 274–280. *https://doi.org/10.1177/1556264616654610*
19. Tait, A. R., Voepel-Lewis, T., Malviya, S., and Philipson, S. J. (2005) Improving the Readability and Processability of a Pediatric Informed Consent Document: Effects on Parents' Understanding. *Archives of Pediatrics & Adolescent Medicine*, *159*(4), 347–352. *https://doi.org/10.1001/archpedi.159.4.347*
20. Coyne, C. A., Ronghui, X., Cella, D., Raich, P., Plomer, K., Dignan, M., Wenzel, L., Fairclough, D, Habermann, T., Schnell, L., and Quella, S. (2003) Randomized, Controlled Trial of an Easy-to-Read Informed Consent Statement for Clinical Trial Participation: A Study of the Eastern Cooperative Oncology Group. *Journal of Clinical Oncology*, *21*(5), 836–842. *https://doi.org/10.1200/JCO.2003.07.022*
21. Stunkel, L., Benson, M., McLellan, L., Sinaii, N., Bedarida, G., Emanuel, E., and Grady, C. (2010) Comprehension and Informed Consent: Assessing the Effect of a Short Consent Form. *IRB*, *32*(4), 1–9.
22. National Academy of Neuropsychology. Informed Consent in Clinical Neuropsychology Practice. *https://www.nanonline.org/docs/PAIC/PDFs/NANinformedconsent.pdf* (Accessed May 2024)
23. Saint Joseph's University. IRB Forms & Templates. *https://sites.sju.edu/researchservices/irb-human-subjects-research/irb-forms-templates/* (Accessed May 2024)
24. Barrett, L. F., and Simmons, W. K. (2015) Interoceptive predictions in the brain. *Nature reviews neuroscience*, *16*(7), 419–429. *https://www.nature.com/articles/nrn3950*
25. Yang, C., Potts, R. and Shanks, D.R. (2018) Enhancing learning and retrieval of new information: a review of the forward testing effect. *npj Science Learn* 3(8). *https://doi.org/10.1038/s41539-018-0024-y*

ABOUT THE STUDENT AUTHOR

Kaitlyn Carr is a senior majoring in Psychology and Spanish and minoring in Behavioral Neuroscience. She plans to graduate in the Spring of 2025.

PRESS SUMMARY

This study examines the reading level, format, and interactivity of consent forms, and how these aspects make consent forms more or less easy to read and understand. The results suggested that consent forms that ask comprehension questions at different points in the reading process as opposed to answering a series of comprehension questions after reading an entire form leads participants to better comprehend what they are reading and ultimately agreeing to. These interactive questions also give participants the incentive to actually read the form as opposed to skimming the text or skipping it completely. The authors suggest that interactive questions in consent forms for research studies, clinical trials, and procedures may be a sufficient way to examine participant understanding of what they are consenting to. Additionally, interactive questions and participants' desire to answer the questions correctly may have a similar effect to that of financial compensation for participation in the way that it gives participants an incentive to spend more time and attention on the study and ultimately be more informed in the informed consent process.

A Finite Difference Approach and Its Error Estimate to the Two-Dimensional Poisson Equation for Dirichlet Boundary Conditions

*Matthew Whalen**

Department of Computer Science, University of Kentucky, Lexington, KY

https://doi.org/10.33697/ajur.2025.140

Students: mrwh233@uky.edu, mrwhalen22@gmail.com*
*Mentor: gi6036@wayne.edu**

ABSTRACT
This study introduces a regular five-point finite difference method for approximating the two-dimensional Poisson equation with Dirichlet boundary condition for convex polygonal domains. The Poisson equation frequently emerges in many fields of science and engineering, such as field potentials and heat transfer. As exact solutions are rarely possible, numerical approaches are important to develop efficient and practical modeling. This introductory paper addresses the uniqueness of the problem, finite difference discretization, consistency of the problem, and maximum norm error analysis. We also provide numerical results that not only validate theoretical results but also demonstrate the method's efficiency. Principally, this paper intends to serve as a compilation of the research which underpins the finite difference methods in a way that is unified, consistent, and accessible to undergraduates. Additionally, we have made the MATLAB code for these results publicly available at the end of this paper for reference and practical implementation.

KEYWORDS
Dirichlet; Poisson; Convex; Finite Differences; Norm; Computer Simulation; Numerical Methods

INTRODUCTION

Many physical problems in engineering and physics can be modeled mathematically, giving researchers the chance to precisely model the solutions, often via computers. These modeled problems are often governed by partial differential equations. The two-dimensional Poisson equation is a particular type of partial differential equation that describes field potentials, heat transfer, image restoration and denoising, and many other applications. To be valuable, the two-dimensional Poisson problem needs to be quickly solvable with a high degree of accuracy to ensure valid modeling of these problems. Since the equation is a partial differential equation (PDE), solving problems of its type is challenging to do using standard algebraic methods. Because only a small number of these PDEs have simple algebraic solutions, we will attempt to solve the two-dimensional Poisson equation using sufficiently accurate numerical approximations. We refer the readers to [1–6] for a comprehensive discussion on partial differential equations, numerical analysis and its applications.

Finite-difference,[7, 8] finite element,[9, 10] and finite volume method[11, 12] are three useful methods to numerically solve partial differential equations. This study uses the finite difference method to numerically approximate the solution to the Poisson equation. It works by replacing the continuous derivative operators with approximate finite differences. The finite difference method is simple, effective, and one of the oldest methods to solve the Poisson equation. Although it is one of the oldest methods ever devised, comprehensive information is difficult to find compiled in a single reference. Therefore, this paper provides a complete study of the two-dimensional Poisson equation through a literature review,

proof of uniqueness and consistency results using the 5-point difference scheme, theoretical error analysis and numerical results.

One of the challenges of the two-dimensional version of this problem is the boundary constraints become complicated and take up a large portion of the solution's runtime.[13] These irregular domains tend to show singular behavior near sharp corners when using numerical solutions.[14–17] In the presence of discontinuous coefficients and singular source terms, these problems become more challenging.[18, 19] Due to this, we will use rectangular domains, which simplify the method and allow for easier understanding. For resources on problems involving non-convex domain, please refer to[5, 20–22] for extra study.

The rest of the article is organized as follows: In *Methods and Procedures*, we introduce our model problem, which is the Poisson equation, and prove its uniqueness. Then, we present a five-point finite difference scheme to approximate the solution to the Poisson equation. In *Error Estimates*, we present the theoretical maximum norm error estimates. Code results and discussion are presented in *Results* to validate the theoretical results and the efficiency of the method. Lastly, we conclude with a discussion of potential future works and a reiteration of the impact of this work. Please note that the methods, theorems, and proofs presented in the following sections are drawn from other works, and as such are available in the references listed in each section.

METHODS AND PROCEDURES

Finite difference methods are techniques to find approximate solutions to ordinary differential equations (ODEs) and partial differential equations (PDEs) numerically. They are based on the idea of replacing the ordinary or partial derivatives with a finite difference quotient. In some sense, a finite difference formulation offers a more direct and intuitive approach to the numerical solution of partial differential equations than other formulations.

Poisson equation

In this section we discuss the Poisson equation in two dimensions

$$\begin{aligned} -\Delta u(x,y) &= f(x,y) \text{ in } \Omega \\ u(x,y) &= g(x,y) \text{ on } \partial\Omega \end{aligned} \qquad \textbf{Equation 1.}$$

where the operator $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ stands for the Laplacian operator and f is a given source function. The solution u of **Equation 1** is an unknown scalar potential function. In the Cartesian coordinate system, the two-dimensional Poisson equation can be written as

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(x,y) \qquad \textbf{Equation 2.}$$

where x, y represent independent spatial dimensions. Since the equation does not have a time-dependent component, there are no initial conditions, making it a boundary value problem. The domain denoted as $\Omega = (a,b) \times (c,d)$ representing a rectangle grid of this size and is subject to a Dirichlet boundary condition. The boundary of the domain denoted as $\partial\Omega$. Next, we present the uniqueness theorem for the Poisson's equation. We refer interested readers to[1, 23–26] and references therein.

Uniqueness of Solutions to the Poisson Equation

The general version of the uniqueness theorem can be found in[25]. However, we will state and derive the two-dimensional version of the uniqueness theorem below to verify the uniqueness of the solution to equation 2.

Theorem 1. *Let Ω be an open subset in $\mathbb{R}^2$. Then there exists at most one solution u in Ω to **Equation 1** with Dirichlet boundary conditions.*

Proof. Suppose that p and q are two solutions to **Equation 1**. Let s be the difference between these two solutions such

that,

$$s = p - q. \qquad \textbf{Equation 3.}$$

From **Equation 1** and **Equation 3** we get,

$$\begin{aligned} -\Delta s &= -\Delta(p-q) \;\; in \;\; \Omega \\ s &= p - q = g - g = 0 \;\; on \;\; \partial\Omega. \end{aligned} \qquad \textbf{Equation 4.}$$

Due to the linearity property of the Poisson's equation, and since p and q are solutions to **Equation 1**, we have $-\Delta(p-q) = -\Delta(p) + \Delta(q) = f - f = 0 \;\; in \;\; \Omega$, with $s = 0 \;\; on \;\; \partial\Omega$. This gives that $-\Delta s = 0 \;\; in \;\; \Omega$. Thus, we can see that s is twice differentiable and satisfies the Laplace equation and hence it is a harmonic function. The maximum principle for harmonic functions says that every non-constant s must attain its maximum and minimum values on the boundary $\partial\Omega$. Since $s = 0$ on the boundary $\partial\Omega$, by **Equation 3**, and s attains its maximum and minimum values on the boundary s must be equal to 0 in all of the domain Ω. Since $s = p - q$ we have $p = q$ for all x, y in Ω. Therefore, we can conclude that only one solution exists to Poisson's equation (**Equation 1**). □

Finite Difference Scheme

This section presents the five-point finite difference stencil[27] for the two-dimensional Poisson equation, which involves discretizing **Equation 1** into a system of linear equations solvable through iterative methods. The following three steps outline the process of generating the discretized system of equations.[7]

Step 1: Generate a mesh.

The mesh can be generated in a uniform Cartesian coordinate system as follows:

$$x_i = a + ih_x, \qquad i = 0, 1, 2, 3, ..., n \qquad h_x = \frac{b-a}{n} \qquad \textbf{Equation 5.}$$

$$y_j = c + jh_y, \qquad j = 0, 1, 2, 3, ..., m \qquad h_y = \frac{d-c}{m} \qquad \textbf{Equation 6.}$$

Let $u(x_i, y_j)$ be the exact solution at a typical point (x_i, y_j) in domain Ω and $u_{i,j}$ be the approximate solution at that same point. The solution on the boundary points is given in **Equation 1** and using it our goal is to approximate the solution at the interior points.

Step 2: Discretize derivatives.

From the Taylor series expansion for two variables, we represent the second order partial derivatives at the grid point using the following derivation:

$$u(x_{i+1}, y_j) = u(x_i, y_j) + h_x\frac{\partial u}{\partial x} + \frac{(h_x)^2}{2!}\frac{\partial^2 u}{\partial x^2} + \frac{(h_x)^3}{3!}\frac{\partial^3 u}{\partial x^3} + \frac{(h_x)^4}{4!}\frac{\partial^4 u}{\partial x^4} + ...$$

$$u(x_{i-1}, y_j) = u(x_i, y_j) - h_x\frac{\partial u}{\partial x} + \frac{(h_x)^2}{2!}\frac{\partial^2 u}{\partial x^2} - \frac{(h_x)^3}{3!}\frac{\partial^3 u}{\partial x^3} + \frac{(h_x)^4}{4!}\frac{\partial^4 u}{\partial x^4} + ...$$

Summing the two equations, we get

$$u(x_{i+1}, y_j) + u(x_{i-1}, y_j) = 2u(x_i, y_j) + (h_x)^2\frac{\partial^2 u}{\partial x^2} + \frac{(h_x)^4}{12}\frac{\partial^4 u}{\partial x^4} + ...$$

Solving for $\frac{\partial^2 u}{\partial x^2}$, we get

$$\frac{\partial^2 u}{\partial x^2} = u_{xx} \approx \frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{(h_x)^2} - \frac{(h_x)^2}{12}\frac{\partial^4 u}{\partial x^4} + ... \qquad \textbf{Equation 7.}$$

Performing the same expansion $u(x_i, y_{j+1})$ and $u(x_i, y_{j-1})$, we can get the terms for $\frac{\partial^2 u}{\partial y^2}$:

$$\frac{\partial^2 u}{\partial y^2} = u_{yy} \approx \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{(h_y)^2} - \frac{(h_y)^2}{12}\frac{\partial^4 u}{\partial y^4} + ... \quad \textbf{Equation 8.}$$

By adding **Equation 7** and **Equation 8**, the two-dimensional Poisson equation can be approximated as follows:

$$\begin{aligned} -(f(x_i, y_j) + e_{i,j}) = &\frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{(h_x)^2} \\ &+ \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{(h_y)^2} \quad \textbf{Equation 9.} \\ &\text{where} \quad i = 1, 2, 3, ..., n-1 \quad \text{and} \quad j = 1, 2, 3, ..., m-1 \end{aligned}$$

Where $e_{i,j}$ includes the higher order terms of the Taylor series expansion. This is called the local truncation error, which is accrued during the approximation process when choosing to stop the approximation at a specific term. For our methods, we choose to derive to the fourth order approximation, with the fourth partial term representing the largest term in the rest of the infinite series, and the majority of our error. The error $e_{i,j}$ that we use is defined in **Equation 10**.

$$e_{i,j} \approx \frac{(h_x)^2 \partial^4 u}{12 \partial x^4}(x_i, y_i) + \frac{(h_y)^2 \partial^4 u}{12 \partial y^4}(x_i, y_i) + O(h^4), \text{ where } h = \max(h_x, h_y) \quad \textbf{Equation 10.}$$

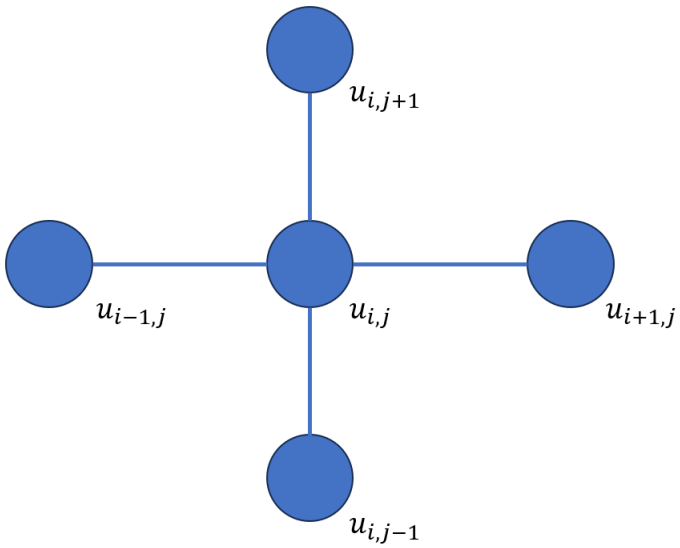


Figure 1. Five-point Finite Difference Stencil.

Since the local truncation error is small for small h values, it is reasonable to ignore $e_{i,j}$ for $h << 0.5$. For our five-point stencil, as can be seen in **Figure 1**, we can ignore the error term $e_{i,j}$ and replace the exact solution $u(x_i, y_j)$ at the grid points with the approximate solution $u_{i,j}$. Then we get the following discretized formula:

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(h_x)^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{(h_y)^2} = -f(x_i, y_j) \quad \textbf{Equation 11.}$$

Equation 11 can be modified as follows:

$$\frac{u_{i+1,j} + u_{i-1,j}}{(h_x)^2} + \frac{u_{i,j+1} + u_{i,j-1}}{(h_y)^2} - \left(\frac{2}{(h_x)^2} + \frac{2}{(h_y)^2}\right) u_{i,j} = -f(x_i, y_j) \quad \textbf{Equation 12.}$$

Note that the finite difference equation at the grid point (x_i, y_j) involves 5 neighboring grid points (x_{i+1}, y_j), (x_{i-1}, y_j), (x_i, y_{j+1}), (x_i, y_{j-1}), and (x_i, y_j) in a five-point stencil. The five-point stencil, shown in **Figure 1**, is simply a finite difference method in which five adjacent points in a grid are used. Without loss of generality, we can assume that $h_x = h_y = h$, so **Equation 12** can be updated to be

$$u_{i+1,j} + u_{i-1,j} - 4u_{i,j} + u_{i,j+1} + u_{i,j-1} = -h^2 f(x_i, y_j). \quad \textbf{Equation 13.}$$

Step 3: Build a system of linear equations.

From **Equation 13** we can build a system of linear equations that can be described by the following matrix system. In this case, we can express the discretized version of **Equation 1** as a matrix equation shown below:

$$-AU = F \quad \textbf{Equation 14.}$$

A represents the stiffness matrix corresponding to the Δ operator and F represents the given source term corresponding to $f(x, y)$, and U represents the discrete solution of **Equation 1**. The new matrix-vector form of **Equation 14** yields the following:

$$A = \frac{1}{h^2}\begin{bmatrix} M & -1 & 0 \\ -1 & \ddots & -1 \\ 0 & -1 & M \end{bmatrix}_{m\times m} \quad \text{where } M = \begin{bmatrix} 4 & -1 & 0 \\ -1 & \ddots & -1 \\ 0 & -1 & 4 \end{bmatrix},$$
$$F = \begin{bmatrix} -f(x_1, y_1) \\ \vdots \\ -f(x_m, y_m) \end{bmatrix}, \text{ and } U = \begin{bmatrix} u_{1,1} \\ \vdots \\ u_{i,j} \end{bmatrix}. \quad \textbf{Equation 15.}$$

Error Estimate

This section discusses how well the numerical approximation determined by the five-point finite difference scheme approximates the exact solution of the two-dimensional Poisson equation for Dirichlet boundary conditions on a convex polygonal domain. In this paper, we use rectangular domains only in order to simplify the computation and make it easier to explain. Other polygonal domains can still be computed in this way, by simply performing a domain transformation form the non-rectangular domain into a domain similar to the one in this paper. More information can be found in.[13, 17] To this end we start with the following three definitions and **Theorem 2**.

Definition 1 (Consistency[2])**.** Let $\Delta_h(u) = -(u_{i-1,j} + u_{i,j-1} - 4u_{i,j} + u_{i,j+1} + u_{i+1,j})$ denote the finite difference approximation associated with the domain Ω_h having mesh size h to the second order partial differential operator $\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$. For a given function $u \in C^\infty(\Omega)$, the truncation error of $\Delta_h(u)$ is given by $e_h(x, y) = (\Delta - \Delta_h)u(x, y)$. The approximation $\Delta_h(u)$ is consistent with Δ if $\lim_{h\to 0} e_h(x, y) = 0$ for all $(x, y) \in \Omega$ and for all $u \in C^\infty(\Omega)$. Moreover, the approximation is consistent to order p if $e_h(x, y) = (h^p)$.

Definition 2 (Convergence[2])**.** Suppose $-\Delta_h\phi(x_j) = f(x_j)$ to be a finite difference approximation to the partial differential equation $-\Delta u(x) = f(x)$ defined on a grid with mesh size h for a simple connect domain $\Omega \subseteq \mathbb{R}^2$. Assume that $\phi(x, y) = u(x, y)$ at all points (x, y) on the boundary $\partial\Omega$. Then the finite difference scheme converges if

$$\max_j |u(x_j) - \phi(x_j)| \to 0 \text{ as } h \to 0.$$

Definition 3 (Maximum norm[7])**.** For any grid function $u : \Omega_h \cup \partial\Omega_h \to \mathbb{R}$, we define the maximum norms as follows:

$$\|u\|_{\infty,\Omega} = \max_{(x_i,y_j)\in\Omega_h} |u_{ij}|,$$
$$\|u\|_{\infty,\partial\Omega} = \max_{(x_i,y_j)\in\partial\Omega_h} |u_{ij}|.$$

In the context of the two-dimensional Poisson equation, the Discrete Maximum Principle is crucial for proving the uniqueness of the solution to the equation. It ensures that, for the discrete version of the problem, the solution behaves similarly to the continuous case where the maximum or minimum of the solution is achieved on the boundary. If the discrete solution u_h satisfies the Discrete Maximum Principle, it is guaranteed that the discrete solution converges to the true continuous solution as the grid resolution increases (in the case of a consistent and stable numerical method).

Theorem 2 (Discrete Maximum Principle[7])**.**

(i) If $\Delta_h u_{i,j} \geq 0$ *for all points* $(x_i, y_j) \in \Omega_h$ *then,* $\max_{(x_i,y_j)\in\Omega_h} u_{i,j} \leq \max_{(x_i,y_j)\in\partial\Omega_h} u_{i,j}$.

(ii) If $\Delta_h u_{i,j} \leq 0$ *for all points* $(x_i, y_j) \in \Omega_h$ *then,* $\min_{(x_i,y_j)\in\Omega_h} u_{i,j} \geq \min_{(x_i,y_j)\in\partial\Omega_h} u_{i,j}$.

Proof. We can prove *(i)* by contradiction. Assume that there exists some $u_{m,n} \in \Omega_h$, where $u_{m,n} = K > 0$, and $K > \max_{\partial\Omega} u_{i,j}$. Given a boundary value $g(x_i, y_j) \geq 0$, we have

$$f_{m-0.5,n}u_{m-1,n} + f_{m+0.5,n}u_{m+1,n} \geq [f_{m-0.5,n} + f_{m+0.5,n} + h^2 g_{m,n}]u_{m,n}.$$

Since h^2 is small, we remove it and the inequality holds. We also replace $u_{m,n}$ with K and get

$$f_{m-0.5,n}u_{m-1,n} + f_{m+0.5,n}u_{m+1,n} \geq [f_{m-0.5,n} + f_{m+0.5,n}]K.$$

Since $u_{m-1,n}$ and $u_{m+1,n}$ are smaller than K by assumption, we have

$$f_{m-0.5,n}u_{m-1,n} + f_{m+0.5,n}u_{m+1,n} \leq [f_{m-0.5,n} + f_{m+0.5,n}]K.$$

With the inequality holding unless $u_{m+1,n} = u_{m-1,n} = K$. If we repeat this argument for all interior points in both grid directions, we can conclude that all boundary points $(x_i, y_j) \in \partial\Omega_h$ are equal to K, which contradicts our assumption of a strict inequality, thus proving *(i)*. The exact same logic using the minimum proves *(ii)*.

□

Theorem 3. *The five-point finite difference analog* $\Delta_h(u) = -(u_{i-1,j} + u_{i,j-1} - 4u_{i,j} + u_{i,j+1} + u_{i+1,j})$ *is consistent to order 2 for the Laplace operator* $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$.[2, 7]

Proof. Let $u \in C^\infty(\Omega)$ and $(x, y) \in \Omega$ be a point such that $(x \pm h, y), (x, y \pm h) \in \Omega \cup \partial\Omega$. Then, by the Taylor Theorem we have

$$u(x+h,y) = u(x,y) + h\frac{\partial u}{\partial x}(x,y) + \frac{h^2}{2!}\frac{\partial^2 u}{\partial x^2}(x,y) + \frac{h^3}{3!}\frac{\partial^3 u}{\partial x^3}(x,y) + \frac{h^4}{4!}\frac{\partial^4 u}{\partial x^4}(\eta^{\pm},y) \quad \textbf{Equation 16.}$$

$$u(x-h,y) = u(x,y) - h\frac{\partial u}{\partial x}(x,y) + \frac{h^2}{2!}\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{h^3}{3!}\frac{\partial^3 u}{\partial x^3}(x,y) + \frac{h^4}{4!}\frac{\partial^4 u}{\partial x^4}(\eta^{\pm},y) \quad \textbf{Equation 17.}$$

where $\eta^{\pm} \in (x-h, x+h)$. By adding **Equation 16** and **Equation 17**, we get

$$\frac{1}{h^2}\left[u(x+h,y) - 2u(x,y) + u(x-h,y)\right] - \frac{\partial^2 u}{\partial x^2}(x,y) = \frac{h^2}{4!}\left[\frac{\partial^4 u}{\partial x^4}(\eta^+,y) + \frac{\partial^4 u}{\partial x^4}(\eta^-,y)\right] \quad \textbf{Equation 18.}$$

Recall that by the intermediate value theorem, we have

$$\left[\frac{\partial^4 u}{\partial x^4}(\eta^+,y) + \frac{\partial^4 u}{\partial x^4}(\eta^-,y)\right] = 2\frac{\partial^4 u}{\partial x^4}(\eta,y)$$

for some value $\eta \in (x-h, x+h)$. Suppose

$$\beta_x^2(x,y) = \frac{1}{h^2}\left[u(x+h,y) - 2u(x,y) + u(x-h,y)\right]$$

and thus

$$\beta_x^2(x,y) = \frac{\partial^2 u}{\partial x^2}(x,y) + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\eta, y). \quad \textbf{Equation 19.}$$

Similarly, we can obtain

$$\beta_y^2(x,y) = \frac{\partial^2 u}{\partial y^2}(x,y) + \frac{h^2}{12}\frac{\partial^4 u}{\partial y^4}(x, \mu). \quad \textbf{Equation 20.}$$

for some $\mu \in (y-h, y+h)$. Finally, by adding **Equation 19** and **Equation 20**, we can conclude that $e_h(x,y) = (\Delta - \Delta_h)u(x,y) = O(h^2)$. Thus, the five-point finite difference scheme is consistent to order 2. □

Proposition 1. *If $\Delta_h u_{ij} = 0$ for $(x_i, y_j) \in \Omega_h$, and $u_{ij} = 0$ for $(x_i, y_j) \in \partial\Omega_h$, then the zero function $u_{ij} = 0$ for all $(x_i, y_j) \in \Omega_h \cup \partial\Omega_h$ is the only solution to the finite difference problem.*

Proposition 2. *If $\Delta_h u_{ij} = f_{ij}$ for $(x_i, y_j) \in \Omega_h$, and $u_{ij} = g_{ij}$ for $(x_i, y_j) \in \partial\Omega_h$, then there exists a unique solution to the finite difference problem for grid functions f_{ij} and g_{ij}.*

Lemma 1. *Suppose that the grid function $u : \Omega_h \cup \partial\Omega_h \to \mathbb{R}$ satisfies the boundary condition $u_{ij} = 0$ for $(x_i, y_j) \in \partial\Omega_h$, then we have the following estimate[7]:*

$$\|u\|_{\infty,\Omega} \leq \frac{1}{8}\|\Delta_h u\|_{\infty,\Omega} \quad \textbf{Equation 21.}$$

By leveraging **Proposition 2** and **Lemma 1**, we can establish the validity of our main theorem. This theorem asserts that the solution derived from the five-point finite difference scheme converges to the exact solution of **Equation 1**. Our proof for this convergence is outlined in the next section.

Theorem 4 ([2, 7]). *Let u be a solution to the Poisson equation (**Equation 1**) and let $\hat{u}$ be the grid function that satisfies the discrete system as follows:*

$$\begin{aligned} -\Delta_h \hat{u}_{ij} &= f_{ij} \text{ for } (x_i, y_j) \in \Omega_h, \\ \hat{u}_{ij} &= g_{ij} \text{ for } (x_i, y_j) \in \partial\Omega_h. \end{aligned} \quad \textbf{Equation 22.}$$

Then, there exists a positive constant C and a constant $K = \left\| \frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right\|_{\infty,\Omega}$ such that

$$\|u - \hat{u}\|_{\infty,\Omega} \leq CKh^2.$$

Proof. For smooth functions f and g the theorem assumes that $u \in C^4(\bar{\Omega})$ since the constant K involves 4th derivatives. By **Equation 19** and **Equation 20** we have the following estimate:

$$(\Delta_h - \Delta)u_{ij} = \frac{h^2}{12}\left[\frac{\partial^4 u}{\partial x^4}(\eta_i, y_j) + \frac{\partial^4 u}{\partial y^4}(x_i, \mu_j)\right] \quad \textbf{Equation 23.}$$

for some $\eta_i \in (x_{i-1}, x_{i+1})$ and $\mu_j \in (y_{j-1}, y_{j+1}$. Since $-\Delta u_i j = f_{ij}$, we have

$$-\Delta_h u_{ij} = f_{ij} - \frac{h^2}{12}\left[\frac{\partial^4 u}{\partial x^4}(\eta_i, y_j) + \frac{\partial^4 u}{\partial y^4}(x_i, \mu_j)\right]. \quad \textbf{Equation 24.}$$

By subtracting **Equation 22** from **Equation 24**, we get

$$\Delta_h(u_{ij} - \hat{u}_{ij}) = \frac{h^2}{12}\left[\frac{\partial^4 u}{\partial x^4}(\eta_i, y_j) + \frac{\partial^4 u}{\partial y^4}(x_i, \mu_j)\right]. \quad \textbf{Equation 25.}$$

Note that $u - \hat{u}$ vanishes on $\partial\Omega_h$ due to the imposed boundary conditions of **Equation 1** and **Equation 22**. If we replace u by $u - \hat{u}$ in **Lemma 1**, we obtain the following:

$$\begin{aligned}\|u - \hat{u}\|_{\infty,\Omega} &\leq \frac{1}{8}\|\Delta_h(u - \hat{u})\|_{\infty,\Omega} \\ &\leq \frac{1}{8} \cdot \frac{h^2}{12}\left\|\frac{\partial^4 u}{\partial x^4}(\eta_i, y_j) + \frac{\partial^4 u}{\partial y^4}(x_i, \mu_j)\right\| \\ &\leq KCh^2\end{aligned}$$

The second equality is derived from **Equation 25** and the third is derived from the definition of *K*. The last equality concludes **Theorem 4**. □

RESULTS

In this section we present a few numerical experiments to illustrate the computational method discussed in this paper. The numerical experiments are performed on a laptop computer with MATLAB R2022a using a MacBook Air with M1 chip. We use the following formula for the order of convergence. We use the following formula for the order of convergence.

$$\text{Order of convergence } R = \frac{\log\left((u_{i+1} - \hat{u}_{i+1})/(u_i - \hat{u}_i)\right)}{\log\left(\frac{1}{2}\right)},$$

where u_{i+1} is the numerical solution in the state $(i+1)^{th}$ and $\hat{u}_{i+1}$ is the analytical solution in the state $(i+1)^{th}$.

Example 1. Consider the Poisson equation:

$$\begin{aligned}-\Delta u(x, y) &= -2\pi^2 \sin(\pi x)\sin(\pi y) \text{ in } \Omega \\ u(x, y) &= 0 \text{ on } \partial\Omega,\end{aligned} \qquad \textbf{Equation 26.}$$

where $\Omega = (0, 2) \times (0, 2)$ and the analytical solution $u(x, y) = \sin(\pi x)\sin(\pi y)$.

Test Case 1. In this example, we compare the numerical solution and the analytical solution with a sequence of different meshes obtained through mesh refinements. We record the maximum norm error between the analytical and numerical solutions in each mesh stage. **Table 1** shows that numerical solutions approximate the true solution better when we decrease the step size. This guarantees that our numerical solution converges to the true solution. **Figure 2** shows the analytical solution to **Equation 26**, and **Figure 3** shows a sequence of numerical solutions to **Equation 26** obtained through six consecutive mesh refinements.

Degrees of Freedom	Max Norm Error $\|u - \hat{u}\|_\infty$
4	0.999984501426793
16	0.177517468002680
64	0.033384181128561
256	0.007287288315757
1024	0.001695751741142
4096	3.989211285880812e-04
16384	8.648598269755947e-05

Table 1. Maximum norm errors for consecutive mesh refinements.

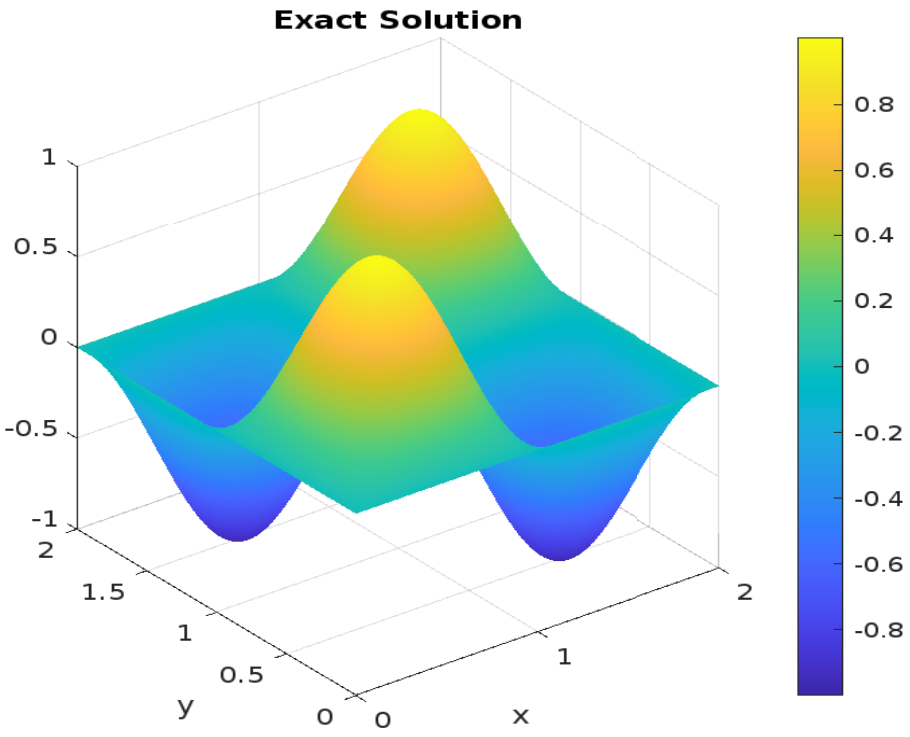
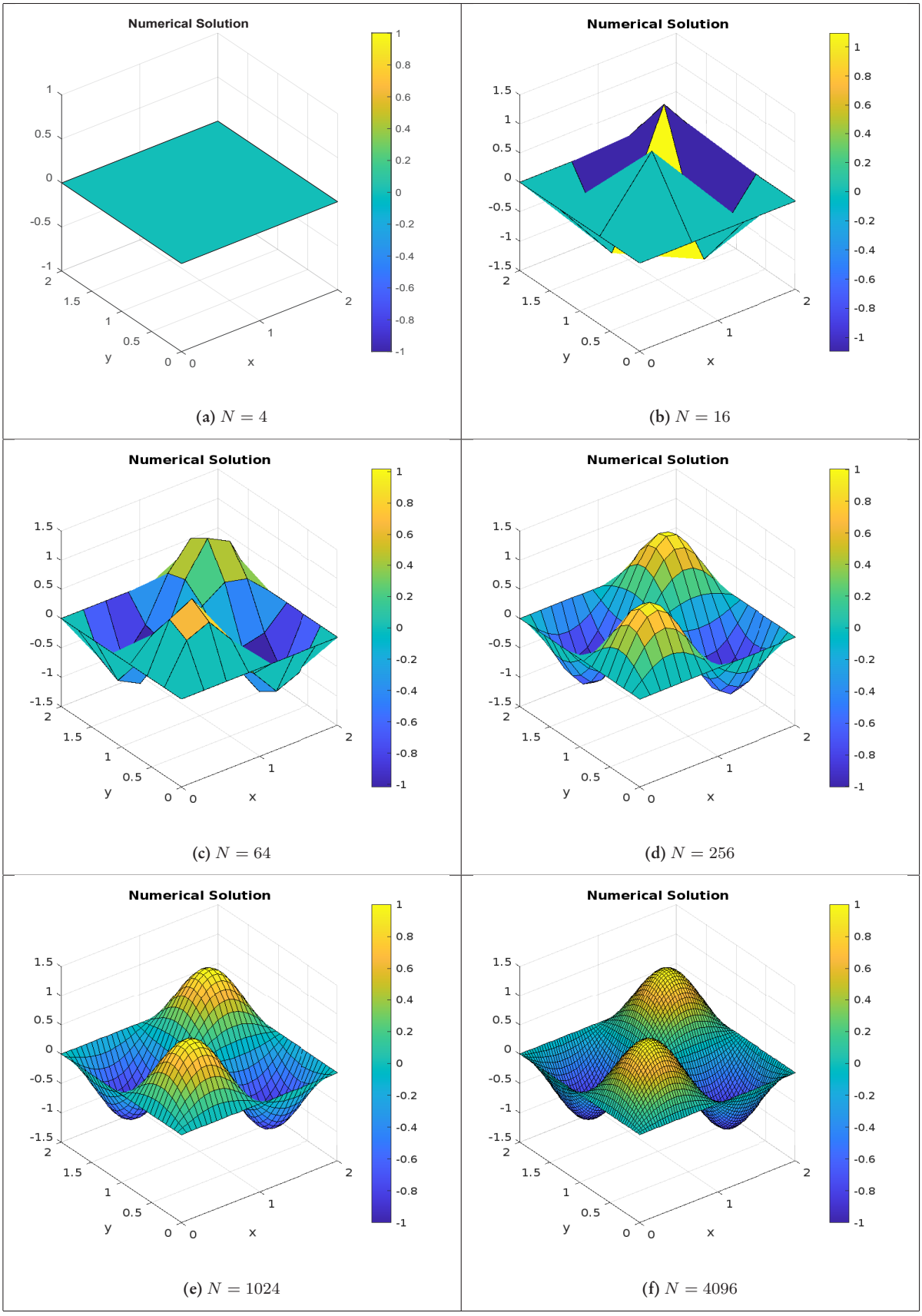


Figure 2. Exact Solution.



(a) $N = 4$ (b) $N = 16$

(c) $N = 64$ (d) $N = 256$

(e) $N = 1024$ (f) $N = 4096$

Figure 3. Numerical Solution for Six Consecutive Meshes.

Test Case 2. Consider the Poisson equation (**Equation 26**) on a domain, $\Omega = (0,1) \times (0,1)$. For this test case, we record the error between numerical solution and analytical solution for each node on the same mesh. The solution values and errors are displayed in **Table 2**. The error is small even with the initial stage of the mesh refinement. **Figure 4** shows the analytical solution and the numerical solution on a mesh with degrees of freedom $N = 25$.

Node (x, y)	Analytical Solution	Numerical Solution	Error
$(0.25, 0.25)$	0.500000000000000	0.526514643772757	0.026514643772757
$(0.5, 0.25)$	0.707106781186547	0.744604150011472	0.037497368824925
$(0.75, 0.25)$	0.500000000000000	0.526514643772757	0.026514643772757
$(0.25, 0.5)$	0.707106781186547	0.744604150011472	0.037497368824925
$(0.5, 0.5)$	1.000000000000000	1.053029287545515	0.053029287545515
$(0.75, 0.5)$	0.707106781186547	0.744604150011472	0.037497368824925
$(0.25, 0.75)$	0.500000000000000	0.526514643772757	0.026514643772757
$(0.5, 0.75)$	0.707106781186547	0.744604150011472	0.037497368824925
$(0.75, 0.75)$	0.500000000000000	0.526514643772757	0.026514643772757

Table 2. Comparison of Analytical and Numerical Solutions on the same Mesh Refinement.

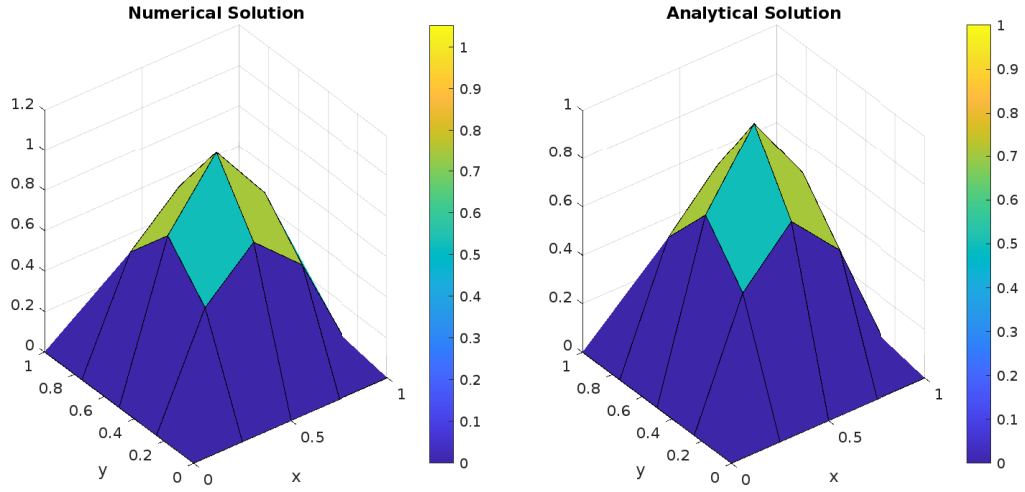


Figure 4. Comparison of Numerical and Analytical Solutions at N=25.

Example 2. In **Example 2**, we report convergence orders of the numerical solution of **Equation 26** obtained through a sequence of mesh refinements. **Table 3** shows that the numerical convergence order $R \approx 2$, which is in strong agreement

with the convergent order shown in **Theorem 3**.

DoF	4	16	64	256	1024	4096	16384
R	-	-	2.4107	2.1957	2.1035	2.0877	2.2056

Table 3. Convergence Orders R for a Sequence of Meshes.

Example 3. In this example, we report the CPU time required by the five-point stencil finite difference method to solve **Equation 26** with $f = 1$ on a square domain $\Omega = (0, 1) \times (0, 1)$. The results are presented in seconds for varying degrees of freedom (DoF) for each mesh in **Table 4**. From **Table 4**, we can observe that the finite difference methods converge to the true solution quickly.

DoF	4	16	64	256	1024	4096	16384
$\mathbf{t_{CPU}}$	0.00280	0.00591	0.00943	0.01425	0.03049	0.86497	44.80878

Table 4. CPU Time in Seconds.

CONCLUSIONS

This study presents a complete analysis of the two-dimensional Poisson equation. We present the derivation of a five-point finite difference scheme and prove its uniqueness and consistency. We also present an infinity norm error estimate and validate those results through numerical simulations. The numerical results presented in the tables and graphs show that the present method approximates the exact solution efficiently with a quadratic convergence, validating our theoretical assumption.

This problem can be extended in several directions. Solving Poisson's equation on a three-dimensional nonconvex polygonal domains with a singularly perturbed parameter added to the Poisson's equation will be an interesting problem which requires more advanced numerical methods such as finite element methods to discretize the Poisson's equation. Another challenging task is to use the point delta function or the line delta function as the source term for the three-dimensional Poisson's equation. Lastly, this can be extended to higher-order PDEs, which have many real-world applications.

The MATLAB code can be accessible via the following link: `https://github.com/mrwhalen22/2D\_poisson\_finite\_difference.git`

ACKNOWLEDGEMENTS

The author thanks their research mentor, Dr. Charuka D. Wickramasinghe, from the University of Kentucky for their guidance on this article during Winter 2023.

REFERENCES

1. Evans, L. C. (2022) Partial differential equations. Vol. 19. *American Mathematical Society*.
2. Lapidus, L. and Pinder G. F. (1999) Numerical solution of partial differential equations in science and engineering. *John Wiley & Sons*.
3. Özişik, M., Necati, Helcio RB Orlande, Marcelo J. Colaço, and Renato M. Cotta. (2017) Finite difference methods in heat transfer. *CRC press*.
4. Phillips, G. M. and Taylor, P. J. eds. (1996) Theory and applications of numerical analysis. *Elsevier*.
5. Wickramasinghe, C. D., & Ahire, P. (2024). A Graded Mesh Refinement For 2d Poisson's Equation on Non-Convex Polygonal Domains. *Asia Pac. J. Math. 2024 11:63. https://doi.org/10.28924/APJM/11-63*.

6. Wickramasinghe, Charuka. 2025. "A Numerical Scheme to Solve Fourth Order Convection Reaction Diffusion Problems With Boundary Layers". Communications in Mathematics and Applications 15 (3). https://www.rgnpublications.com/journals/index.php/cma/article/view/2778.
7. Li, Zhilin, Zhonghua Qiao, and Tao Tang. (2017) Numerical solution of differential equations: introduction to finite difference and finite element methods. *Cambridge University Press*.
8. Yoon, G. J. and Min, C. (2015) Analyses on the finite difference method by Gibou et al. for Poisson equation, J. Comput. Phys. **280**, 184–194. MR3273131.
9. Kim, Nam-Ho, Bhavani V. Sankar, and Ashok V. Kumar. (2018) Introduction to finite element analysis and design. *John Wiley & Sons*.
10. Dilhara Wickramasinghe, C. (2021) "An Efficient Numerical Approximation to Poisson Problem in Two Dimensions."
11. Mazumder, S. (2016) *Numerical methods for partial differential equations*, *Elsevier/Academic Press*, Amsterdam, MR3443468.
12. Moukalled, Fadl, Mangani, L., Darwish, M., Moukalled, F., Mangani, L., and Darwish, M. (2016) The finite volume method. *Springer International Publishing*.
13. Matte, J. -P. and Lafrance, G. (1977) Solution of the discrete Poisson equation with complicated boundaries, *J. Comput. Phys*. **23**, no. 1, 86–91. MR0502093
14. Brandstetter, G. and Govindjee, S. (2015) A high-order immersed boundary discontinuous-Galerkin method for Poisson's equation with discontinuous coefficients and singular sources, *Internat. J. Numer. Methods Engrg*. **101**, no. 11, 847–869. MR3311660.
15. Chen, L. and Li, H. (2010) Superconvergence of gradient recovery schemes on graded meshes for corner singularities, *J. Comput. Math*. **28** (2010), no. 1, 11–31. MR2603578
16. Feng, Q., Han, B., and Minev, P. D. (2021) Sixth order compact finite difference schemes for Poisson interface problems with singular sources, *Comput. Math. Appl*. **99** (2021), 2–25. MR4302074
17. Li, H., Wickramasinghe, C. D., and Peimeng Yin. (2022) "A C^0 finite element method for the biharmonic problem with Dirichlet boundary conditions in a polygonal domain." *arXiv* preprint arXiv:2207.03838.
18. Musolino, P. (2013) A singularly perturbed Dirichlet problem for the Poisson equation in a periodically perforated domain. A functional analytic approach, in *Advances in harmonic analysis and operator theory*, 269–289, Oper. Theory Adv. Appl., 229, Birkhäuser/Springer Basel AG, Basel. MR3060419
19. Wickramasinghe, Charuka Dilhara, "A C^0 Finite Element Method For The Biharmonic Problem In A Polygonal Domain" (2022). Wayne State University Dissertations. 3704. https://digitalcommons.wayne.edu/oa_dissertations/3704.
20. Butler, J. (2021) Numerical Methods for Differential Equations with Python, *https://github.com/john-s-butler-dit/Numerical-Analysis-Python*.
21. Gregory, D. T. and Wickramasinghe, C.D. (2023) An Upwind Finite Difference Method to Singularly Perturbed Convection Diffusion Problems on a Shishkin Mesh. *arXiv* preprint arXiv:2306.03181.
22. Kadalbajoo, M. K. and Patidar, K. C. (2003) "Singularly perturbed problems in partial differential equations: a survey." *Applied Mathematics and Computation* 134, no. 2-3: 371-429.
23. Barton, G. (1989) Elements of Green's functions and propagation: potentials, diffusion, and waves. *Oxford University Press*.
24. Brown, J. and Churchill, R. (2011) Fourier Series and Boundary Value Problems, 8th Edition (*McGraw-Hill Companies*, Inc., New York, NY, 2011).
25. Hössjer, E.(2012) "Uniqueness of Solutions to the Poisson Equation." (2012).
26. Alan, J. (2003) Applied partial differential equations: an introduction. *Academic Press*, 2003.
27. Causon, D. M. and Mingham, C. G. (2010) Introductory finite difference methods for PDEs. *Bookboon*.

ABOUT THE STUDENT AUTHOR

Matthew Whalen completed this work during the summer of 2023, prior to graduating with a bachelor's of Computer Science at the University of Kentucky in May 2024. He is now planning to graduate from the University of Kentucky again in 2025 with a Master's of Business Administration and Management.

PRESS SUMMARY

This paper describes a numerical procedure for simulating solutions to Dirichlet Poisson equations using MATLAB. In addition, it can serve as a uniform summary of related literature needed for such a procedure such that anyone can read and use finite difference schemes. This method of approximating differential equations is valuable in numerous fields of engineering and physics. The main goal of this procedure is to make these equations automatable and solvable using computer technology for applications like fluid and heat simulations. Hopefully, this work can continue to be expanded and adapted to non-Dirichlet boundary condition problems.

A Comparison of Zero-Inflated Models for Modern Biomedical Data

Max Beveridge[a], *Zach Goldstein*[b], *& Hee Cheol Chung*[c]
[a] *Department of Statistics, The George Washington University, Washington, DC*
[b] *Department of Mathematics, Yeshiva University, New York, NY*
[c] *Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC*

https://doi.org/10.33697/ajur.2025.141

Students: maxbeveridge03@gwu.edu, zgoldst3@mail.yu.edu
Mentor: hchung13@charlotte.edu

ABSTRACT
There has been a growing number of datasets exhibiting an excess of zero values that cannot be adequately modeled using standard probability distributions. For example, microbiome data and single-cell RNA sequencing data consist of count measurements in which the proportion of zeros exceeds what can be captured by standard distributions such as the Poisson or negative binomial, while also requiring appropriate modeling of the nonzero counts. Several models have been proposed to address zero-inflated datasets including the zero-inflated negative binomial, hurdle negative binomial model, and the truncated latent Gaussian copula model. This study aims to compare various models and determine which one performs optimally under different conditions using both simulation studies and real data analyses. We are particularly interested in investigating how dependence among the variables, level of zero-inflation or deflation, and variance of the data affects model selection.

KEYWORDS
Zero-Inflated Models; Hurdle Models; Truncated Latent Gaussian Copula Model; Microbiome Data; Gene-Sequencing Data; Zero-Inflation, Negative Binomial; Zero-Deflation

INTRODUCTION

Zero-inflated data refers to datasets with an excess of zeros, where the proportion of zeros cannot be adequately captured by standard probability distributions. Such data frequently arise in various fields, such as health and epidemiology, where large numbers of zeros are often encountered. For example, in substance abuse research, the majority of individuals do not engage in substance abuse, leading to a predominance of zero observations.[1] Similarly, zero-inflated data are common in biomedical research including microbiome studies and single-cell RNA sequencing, where zeros occur due to limited sequencing depth.[2,3] Given the widespread occurrence of zero-inflated data across numerous disciplines, it is essential to model these datasets accurately to ensure valid analyses. Failure to properly account for zero inflation can lead to poor estimation and the potential oversight of statistically significant findings. Accurate modeling of zero-inflated data not only improves the estimation of key parameters but also reduces bias and enhances the understanding of dependence structures.[4] Violating distributional assumptions of statistical tests is one of the "seven deadly sins" of comparative analysis.[5] The consequences of which are biased or incorrect parameter estimates and incorrect p-values. With regard to zero-inflated data, several studies have found that misspecifying the distribution of a general linear model (GLM) when data is zero-inflated leads to invalid statistical inference (e.g., using a Poisson or negative binomial (NB) regression model when the data follows a zero-inflated Poisson or zero-inflated NB distribution).[6]

Zero-inflated models, including zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), hurdle Poisson (HP), and hurdle negative binomial (HNB), have been widely used to model zero-inflated data across fields such as ecology, environmental science (e.g., species counts), economics (e.g., consumer purchases), insurance (e.g., claims data), and criminology (e.g., crime counts in different areas). The key difference between zero-inflated and hurdle models lies in how they handle the excess number of zeros. Zero-inflated models combine a point mass at zero with a standard distribution that also allows non-zero probability at zero. The point mass accounts for structural zeros (inherent zeros), while the non-zero probability from the standard distribution models sampling zeros (zeros that occur by chance). In contrast, hurdle models only account for structural zeros by using a mixture of a point mass at zero

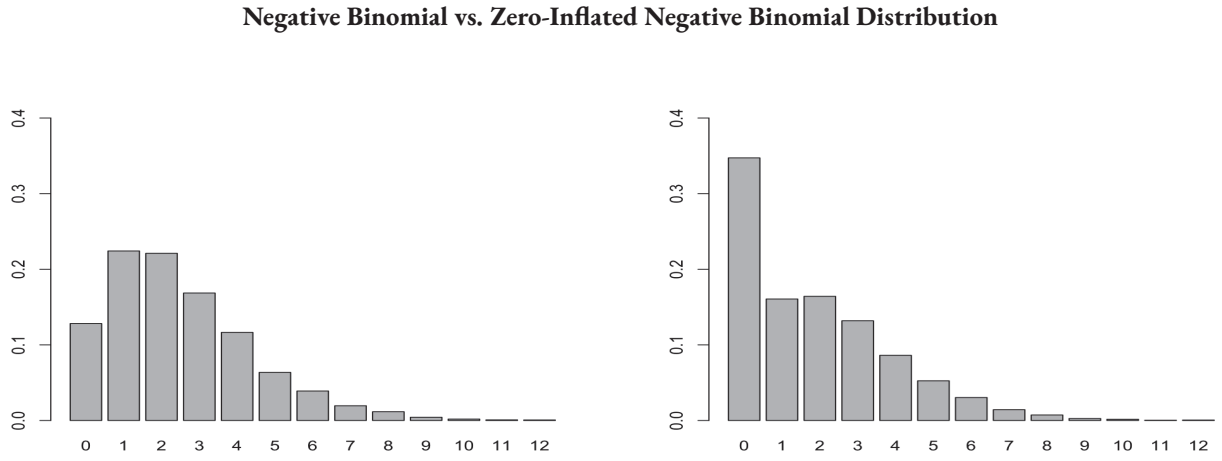


Figure 1. Shown on the left is a standard negative binomial distribution (where $\mu = 2.5$, $r = 5$, and the probability that $Y = 0$ is 0.1317) and on the right is a zero-inflated negative binomial distribution (where $\mu = 2.5$, $r = 5$, and $\pi_Z = 0.25$, and the probability that $Y = 0$ is 0.3487).

In Method and Procedures, we detail each of the zero-inflated models and define terms used in our simulation studies and real data analyses. Then, in Simulation Setting One, we discuss the procedure, results and discussion of comparing the ZINB and HNB models. In Simulation Setting Two and Simulation Setting Three, we discuss the procedure, results, and interpretation of comparing the HNB and TLNPN models (under HNB and TLNPN population data respectively). We then discuss our methods, results, and interpretation of comparing the HNB and TLNPN models using real-world biomedical data in Real Data Analyses. Finally, we will summarize the findings, limitations, and directions for future research in the Conclusion section.

METHOD AND PROCEDURES

In this section, we first detail the zero-inflated models that we will be investigating including the ZINB, HNB, and TLNPN models. We then review important definitions for the proceeding simulation studies and real data analyses.

Models for Zero-Inflated Data

Zero-inflated models account for an excess number of zeros by adjusting the probability of observing zero of a standard probability distribution. In particular, the form of the probability mass function (pmf) of a zero-inflated model is given by:

$$P(Y = y) = \begin{cases} \pi_Z + (1 - \pi_Z)p(y = 0; \mu) & \text{for } y = 0, \\ (1 - \pi_Z)p(y; \mu) & \text{if } y > 0, \end{cases}$$

where $p(\cdot)$ is a pmf of a discrete random variable following a standard distribution, e.g., Poisson or negative binomial distribution, μ is the mean of the distribution, and π_Z is the weight parameter controlling the degree of zero inflation. One of the popularly used zero-inflated models is the zero-inflated negative binomial (ZINB) model with pmf:

$$P(Y = y) = \begin{cases} \pi_Z + (1 - \pi_Z)(\frac{r}{\mu+r})^r & \text{for } y = 0, \\ (1 - \pi_Z)\frac{\Gamma(y+r)}{\Gamma(r)y!}(\frac{\mu}{\mu+r})^y(\frac{r}{\mu+r})^r & \text{if } y > 0, \end{cases}$$

where μ is the mean of the negative binomial model, r is the dispersion parameter, and π_Z is the probability of structural zeros. In zero-inflated models, the probability of observing a zero is given by $\pi_Z + (1 - \pi_Z)p(y = 0; \mu)$. As a result, the probability is bounded below by $p(y = 0; \mu)$, which corresponds to the probability under the standard negative binomial model. Consequently, zero-inflated models cannot account for zero deflation. An illustrative example of the zero-inflated negative binomial distribution is provided in **Figure 1**. In contrast to zero-inflated models, hurdle models are able to account for zero-inflation and zero-deflation.

Hurdle models are distinct from zero-inflated models because they only account for structural zeros and are able to model zero-deflation. Zero-deflation occurs when there are less zero values present than a standard probability distribution would predict. The form of the

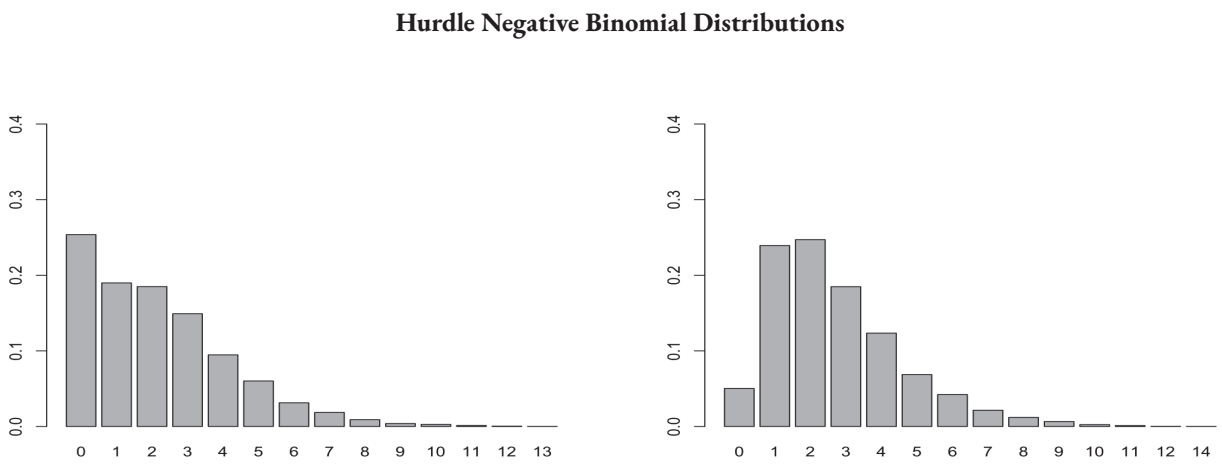


Figure 2. Shown are hurdle negative binomial distributions with $\mu = 2.5$ and $r = 5$. In the left histogram, $\pi_H = 0.25$, so the probability $Y = 0$ is 0.25, and in the right histogram, $\pi_H = 0.05$, so the probability that $Y = 0$ is 0.05. Under a standard negative binomial distribution, the probability that $Y = 0$ is 0.1317.

pmf of a hurdle model is:

$$P(Y = y) = \begin{cases} \pi_H & \text{for } y = 0 \\ (1 - \pi_H)\frac{p(y;\mu)}{1-p(y=0;\mu)} & \text{if } y > 0 \end{cases}$$

where $p(\cdot;\mu)$ is the pmf of a Poisson or negative binomial distribution with mean μ. The parameter π_H is the probability that a structural zero occurs and can take any value from 0 to 1. The hurdle negative binomial (HNB) model is given by:

$$P(Y = y) = \begin{cases} \pi_H & \text{for } y = 0 \\ \frac{1-\pi_H}{1-(\frac{r}{\mu+r})^r}\frac{\Gamma(y+r)}{\Gamma(r)y!}(\frac{\mu}{\mu+r})^y(\frac{r}{\mu+r})^r & \text{if } y > 0. \end{cases}$$

Under the hurdle model, zero occurs with probability π_H, which can be smaller than the probability of $Y = 0$ under the negative binomial model and thus capable of modeling zero-deflated variables. Examples of the hurdle negative binomial distribution can be seen in **Figure 2**. When data involves multiple zero-inflated variables, their associations can be modeled within the generalized linear model framework, assuming covariates are available.

For multiple zero-inflated random variables $Y_1, \ldots, Y_p$, given the covariates $\boldsymbol{x} = (x_1, \ldots, x_{q_1})^\top$ and $\boldsymbol{z} = (z_1, \ldots, z_{q_2})^\top$, which are shared across $Y_1, \ldots, Y_p$, their associations can be modeled within the generalized linear model (GLM) framework. In particular, the ZINB regression model is given by

$$\ln(\mu_j) = \boldsymbol{x}^T\boldsymbol{\beta}_j, \text{ and } \operatorname{logit}(\pi_{Z,j}) = \boldsymbol{z}^T\boldsymbol{\gamma}_j \qquad \textbf{Equation 1.}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{q_1}$ and $\boldsymbol{\gamma}_j \in \mathbb{R}^{q_2}$ are the regression coefficients for the mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ and $\boldsymbol{\pi}_Z = (\pi_{Z,1}, \ldots, \pi_{Z,p})^\top$, respectively, and $\operatorname{logit}(\pi_Z) = \ln\{\pi_Z/(1 - \pi_Z)\}$. For each $j = 1, \ldots, p$, the parameters $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$, and the dispersion parameter r_j, can be estimated using the maximum likelihood estimator. Let $Y_{i1}, \ldots, Y_{ip}, i = 1, \ldots, n$, be a random sample. The log-likelihood function of the jth variable, $L_{ZI,j}$ is defined as $L_{ZI,j} = L_{1,j} + L_{2,j} + L_{3,j} - L_{4,j}$, where

$$L_{1,j} = \sum_{i:y_i=0} \ln\left\{e^{\boldsymbol{z_i}^T\boldsymbol{\gamma}_j} + \left(1 + \frac{\mu_{ij}}{r_j}\right)^{-r_j}\right\}, \quad L_{2,j} = \sum_{i:y_{ij}>0}\sum_{t=0}^{y_{ij}-1} \ln(t + r_j)$$

$$L_{3,j} = \sum_{i:y_i>0} \left\{-\ln(y_{ij}!) - (y_{ij} + r_j)\ln\left(1 + \frac{\mu_{ij}}{r_j}\right) + y_{ij}\ln(r_j^{-1}) + y_{ij}\ln(\mu_{ij})\right\}$$

$$L_{4,j} = \sum_{i=1}^{n} \ln(1 + e^{\boldsymbol{z_i}^T\boldsymbol{\gamma}_j}).$$

Thus, the log-likelihood function of the joint model is given by $L_{ZI} = \sum_j L_{ZI,j}$. The hurdle negative binomial regression model is given by

$$\ln(\mu_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j, \text{ and } \operatorname{logit}(\pi_{H,ij}) = \boldsymbol{x}_i^T \boldsymbol{\gamma}_j \quad \textbf{Equation 2.}$$

The log-likelihood function of the jth variable, $L_{\mathrm{H,j}}$, is given as

$$L_{H,j} = \sum_{i=1}^{n} (I_{y_{ij}=0} \ln(\pi_{H,ij}) + I_{y_i>0}(\ln(1-\pi_{H,ij}) + \ln(h(y_{ij};\mu_{ij},r_j) - \ln(1-(1+r_j\mu_{ij})^{-r_j}))$$

where $h(y_{ij};\mu_{ij},r_j)$ denotes the pmf of the negative binomial distribution with mean μ_{ij} and dispersion parameter r_j.[12] Therefore, the log-likelihood function of the joint hurdle model is given by $L_H = \sum_j L_{H,j}$.

Nevertheless, in real-world applications, such covariates are often not readily available. In these cases, we can only fit the intercept parameters β_0 and γ_0, assuming that all variables are mutually independent. The Gaussian copula model addresses this limitation by utilizing a rank-based correlation estimator. The Gaussian copula model assumes that, for a random vector $\boldsymbol{y} = (Y_1, \ldots, Y_p)^\top$, there exist strictly increasing functions, $g_1, \ldots, g_p$, such that $\boldsymbol{z} = (g_1(Y_1), \ldots, g_p(Y_p))^\top \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It is important to note that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are not identifiable because, for any constants a_j and b_j, the Gaussian copula model still holds with $g_j^* = a_j + b_j g_j$, $j = 1, \ldots, p$, i.e., $(g_1^*(Y_1), \ldots, g_p^*(Y_p))^\top$ follows $N_p(\boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{B\Sigma B})$, where $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$ and $\boldsymbol{B} = \operatorname{diag}\{b_j\}_{j=1}^p$. The identifiability issue is commonly addressed by assuming that $\boldsymbol{\mu} = \mathbf{0}_p$ and $\boldsymbol{\Sigma}$ is a positive definite correlation matrix.[11,13] If g_js are differentiable, then we have an analytic expression as $g_j = \Phi^{-1} \circ F_j$, where F_j and Φ^{-1} are the distribution functions of Y_j and standard Gaussian. The Gaussian copula models are often denoted as $\boldsymbol{y} \sim NPN(\mathbf{0}_p, \boldsymbol{\Sigma}, \boldsymbol{g})$.[14]

The Gaussian copula models assume that Y_j are continuous and are thus not valid for zero-inflated variables. To accommodate zero-inflated and highly skewed variables, the truncated Gaussian copula models[10] have been introduced by incorporating an additional truncation mechanism, as follows:

Definition 1 (Truncated Latent Gaussian Copula Model)**.** *A random vector $\boldsymbol{y} \in \mathbb{R}^p$ satisfies the truncated latent Gaussian Copula model if there exists a random vector $\boldsymbol{y}^* \sim NPN(\mathbf{0}_p, \boldsymbol{\Sigma}, \boldsymbol{g})$ and constants D_j, $j = 1, \ldots, p$ such that $Y_j = I(Y_j^* > D_j)Y_j^*$ where $I(\cdot)$ is an indicator function. We then denote $\boldsymbol{y} \sim TLNPN(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{g}, \boldsymbol{D})$.*

The latent correlation matrix $\boldsymbol{\Sigma}$ of $TLNPN(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{g}, \boldsymbol{D})$ is estimated using Kendall's τ. The sample Kendall's τ between the jth and kth variables is defined as:

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i \le i' \le n} \operatorname{sign}(Y_{ij} - Y_{i'j}) \operatorname{sign}(Y_{ik} - Y_{i'k}).$$

There exists[10,11] an increasing bridge function G defined so $G(\Sigma_{jk}) = E(\hat{\tau}_{jk}) = \tau_{jk}$ where Σ_{jk} is an element of $\boldsymbol{\Sigma}$ corresponding to variables Y_j and Y_k. The bridge function G for two truncated variables is defined as:

$$G_{TT}(\Sigma_{jk}; \Delta_j, \Delta_k) = -2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \boldsymbol{\Sigma}_{4a}) + 2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \boldsymbol{\Sigma}_{4b}),$$

where $\Delta_j = f_j(D_j)$ and $\Phi_4(a_1, a_2, a_3, a_4; \boldsymbol{\Sigma}_4)$ denotes the CDF of 4-dimensional Gaussian with zero mean and correlation matrix $\boldsymbol{\Sigma}_4$ evaluated at $\boldsymbol{a} = (a_1, a_2, a_3, a_4)^\top$. The correlation matrices $\boldsymbol{\Sigma}_{4a}$ and $\boldsymbol{\Sigma}_{4b}$ are given by

$$\boldsymbol{\Sigma}_{4a} = \begin{pmatrix} 1 & 0 & 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} \\ 0 & 1 & \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} & 1 & -\Sigma_{jk} \\ -\Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & -\Sigma_{jk} & 1 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{4b} = \begin{pmatrix} 1 & \Sigma_{jk} & 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} \\ \Sigma_{jk} & 1 & \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & 1 & \Sigma_{jk} \\ 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} & \Sigma_{jk} & 1 \end{pmatrix}.$$

Using the bridge function G, we can consistently[10,11] estimate the latent correlation matrix as $\hat{\Sigma}_{jk} = G^{-1}(\hat{\tau}_{jk}; \hat{\Delta}_j, \hat{\Delta}_k)$, where $\hat{\Delta}_j$ is the moment estimator as $\hat{\Delta}_j = \Phi(\hat{\pi}_j)$ and $\hat{\pi}_j = n^{-1} \sum_{i=1}^{n} 1(Y_{ij} = 0)$ is the sample proportion of zeros of the jth variable. Below, we see how the observed variable Y_j is modeled as a latent standard Gaussian variable, Z_j, truncated by Δ_j, where $\Phi(\Delta_j) = \pi_j$ is the probability of the jth variable taking zero, which is estimated by the sample proportion of zeros,

$$Y_j = I(Y_j^* > D_j)Y_j^* = I\{g(Y_j^*) > g(D_j)\}Y_j^* = I\{Z_j > \Delta_j\}Y_j^* = I\{\Phi(Z_j) > \Phi(\Delta_j)\}Y_j^* = I\{\Phi(Z_j) > \pi_j\}Y_j^*.$$

The latent Gaussian copula model for binary type data was first introduced in 2017 to model dependence among discrete Arabidopsis gene data.[11] The TLNPN model was introduced in 2020 along with the rank-based estimators for the latent correlation matrix, and it was found useful for modeling gene-expression and micro-RNA data.[10] The TLNPN model has shown to be useful when performing discriminant analysis for microbiome data due to its ability to model dependence among zero-inflated variables.[14] At the same time, zero-inflated and hurdle models are also popularly used to model zero-inflated data. However, a lack of research has been done comparing the TLNPN model to the other zero-inflated models and investigating the characteristics of data in which the TLNPN model performs better than the other models.

Definitions for Simulation Studies and Real Data Analyses

We examine the performance and robustness of ZINB, HNB, and TLNPN models using synthetic datasets across various conditions. We simulate data from each of the three populations—ZINB, HNB, and TLNPN—and, in settings two and three, evaluate performance by calculating the Wasserstein distance between test data independently generated from an assumed population model and data generated from the corresponding fitted model. The Wasserstein distance measures the distance between two probability distributions and is used for goodness-of-fit and statistical inference between two probability distributions where a lower distance implies a better fit.[15] Let μ and ν denote the probability measures corresponding to the distributions of the random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Also, let γ be a coupling, which is a probability measure defined on the product space of the probability spaces of $\boldsymbol{x}$ and $\boldsymbol{y}$, with marginals μ and ν. At the population level, the Wasserstein distance is defined as

$$W_p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left\{ \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\gamma} d(\boldsymbol{x}, \boldsymbol{y})^p \right\}^{1/p},$$

where $\Gamma(\mu, \nu)$ is the set of all couplings. At the sample level, the Wasserstein distance between multivariate dataset $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ is defined as

$$W_p(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \inf_{\theta \in \mathcal{S}_n} \left(\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_{\theta(i)}\|^p \right)^{1/p},$$

where θ is a permutation in the symmetric group $\mathcal{S}_n$, the set of all n-permutations. Since $\mathcal{S}_n$ contains $n!$ permutations, we approximate the computation of Wasserstein distances in our numerical study using the network simplex algorithm,[16] implemented in the R package `transport`. As a summary measure of our results, we use arithmetic mean change (AMC). We use AMC because we want to measure the real relative difference between the performance of the models under varying scales of data. Furthermore, AMC is a symmetric measure of relative change, so a 10% improvement and 10% decline in performance from the HNB model to the TLNPN model are both captured by AMC, for example. Let $\omega_{\text{HNB}} = W_p(Y, \hat{Y}_{\text{HNB}})$ and $\omega_{\text{TLNPN}} = W_p(Y, \hat{Y}_{\text{TLNPN}})$ where $\hat{Y}_{\text{TLNPN}}$ is a simulated multivariate dataset generated from the TLNPN model, $\hat{Y}_{\text{HNB}}$ is a simulated multivariate dataset generated from the HNB model, and Y is a multivariate dataset generated from the true model. The AMC of the Wasserstein distance generated between the HNB model and the true model to the Wasserstein distance generated between the TLNPN model and the true model is given by:

$$\text{AMC} = \frac{\omega_{\text{TLNPN}} - \omega_{\text{HNB}}}{(\omega_{\text{TLNPN}} + \omega_{\text{HNB}})/2}. \qquad \textbf{Equation 3.}$$

Therefore, a positive AMC implies that the HNB model performs better, a negative AMC implies the TLNPN model performs better, and AMC=0 implies that the models perform the same.

We explore performance of ZINB, HNB, and TLNPN under three simulation settings. In setting one, we aim to validate previous findings on the robustness of univariate ZINB and HNB models to model misspecification, focusing specifically on varying levels of

zero inflation or deflation and differing proportions of sampling and structural zeros. In this simulation setting, we compare model performance using Akaike Information Criterion (AIC) since they are both discrete distributions and we seek to replicate previous studies.[1] AIC is a measure of model fit based on the likelihood function, with a penalty term that increases linearly with the number of model parameters k. It is defined as

$$\text{AIC} = 2k - 2\ln L,$$

where L is the likelihood function. Smaller AIC values suggest a more favorable model fit. It is a standard measure to compare two models; however, when comparing discrete and continuous models, it is biased towards the continuous model since likelihood values in continuous models are usually higher than probabilities in discrete models. Accordingly, we use AIC to compare the HNB and ZINB models but use Wasserstein distance to compare the HNB and TLNPN models. In simulation setting two, we compare the HNB model to the TLNPN model under HNB population data. We seek to evaluate whether the proportion of zeros or the dependence among the variables has an impact on the relative performance of each model. In simulation setting three, we again compare the HNB and TLNPN models except with TLNPN population data. We again seek to understand how zero-proportion and dependence among variables affects relative model performance.

In the multivariate settings (settings two and three), we apply the following autoregressive (AR) and geometrically decaying eigenvalues (GD) correlation structures to induce dependence between zero-inflated variables. The AR correlation structure is given by:

$$\boldsymbol{\Sigma} = [\rho^{|j-j'|}]_{1\le j,j'\le p}. \qquad \textbf{Equation 4.}$$

The covariance matrix of the GD structure is given by $\boldsymbol{\Sigma} = \boldsymbol{\Gamma N \Gamma}^T$ where $\boldsymbol{N} = \text{diag}\{\nu_j\}_{j=1}^{p}$ is a diagonal matrix with geometrically decaying eigenvalues defined by:

$$\nu_j = \frac{5(\rho^{j-1} - \rho^j)}{1-\rho^p}, \quad j = 1, \ldots, p, \qquad \textbf{Equation 5.}$$

where a lower value of ρ leads to higher correlations (in the absolute value sense) between the covariates, and $\boldsymbol{\Gamma}$ is uniformly generated from the orthogonal group[17] of order p, OG_p, where $OG_p = \{\boldsymbol{O} \mid \boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I}, \boldsymbol{O} \in \mathbb{R}^{p\times p}\}$ is the set of all $p \times p$ orthogonal matrices. The simulation settings are detailed as follows.

SETTING ONE: COMPARING THE UNIVARIATE ZINB AND HNB MODELS

In this setting, we seek to replicate previous findings comparing the univariate ZINB and HNB models under model misspecification and varying zero proportion.[1] We simulated $n = 500$ data points using covariate X_i where $X_i \sim N(0,1)$, $i = 1, \ldots, 500$, $\ln(\mu_i) = \beta_0 + \beta_1 x_i$, $\text{logit}(\pi_{Z,i}) = \gamma_0 + \gamma_1 x_i$ and $\text{logit}(\pi_{H,i}) = \gamma_0 + \gamma_1 x_i$ as in **Equation 1** and **Equation 2**. We performed simulations under three parameter conditions controlling for $\beta_0 = \ln(12)$, $\beta_1 = 2$, $\gamma_1 = 2$, and $r = 0.5$ at 20% zero-proportion, 40% zero-proportion, and 60% zero-proportion, which we adjusted using the γ_0 parameter under both the ZINB model and HNB model. We fix $\beta_1 = \gamma_1 = 2$ to replicate previous findings[1] and fix $\beta_0 = \ln(12)$ and $r = 0.5$ to ensure there is a significant difference in the proportion of structural and sampling zeros under the HNB and ZINB models. We then fit each model to the simulated data and compare the model fit through AIC.

To further investigate the impact that zero-deflation had on relative model performance, we conducted a follow-up simulation study. We simulated $n = 700$ data points using covariate $X_i \sim N(0,1)$, $i = 1, \ldots, 700$, $\ln(\mu_i) = \beta_0 + \beta_1 x_i$, and $\text{logit}(\pi_{H,i}) = \gamma_0 + \gamma_1 x_i$ where $\beta_0 = \ln(6/7)$, $\beta_1 = 0.1$, $\gamma_1 = 0$, and $r = 2$ in **Equation 2**. We varied $\pi_{H,i}$ from 0.08 to 0.7 by adjusting the γ_0 parameter. In this case, under the standard negative binomial distribution with $\mu = \ln(6/7)$ and $r = 2$, the probability of $Y = 0$ is 0.5. We fixed $\beta_0 = \ln(6/7)$ and $r = 2$ in order to make the probability of $Y = 0$ under the standard negative binomial distribution equal to 0.5, so there was a large range of zero-proportions that would be considered zero-deflation. Furthermore, we fixed $\beta_1 = 0.1$ and $\gamma_1 = 0$ because we were mainly interested in the effect of zero-deflation, so we did not want the covariate to have a large impact on the mean or probability of a structural zero. For each iteration, corresponding to a different proportion of zeros, we fit the data using both HNB and ZINB models and compared their AIC values.

We confirmed previous findings and found that when the proportion of sampling and structural zeros differed significantly and β_1 and γ_1 had high values, then the models were sensitive to model misspecification[1] as seen in **Figure 3**. We see that under each propor-

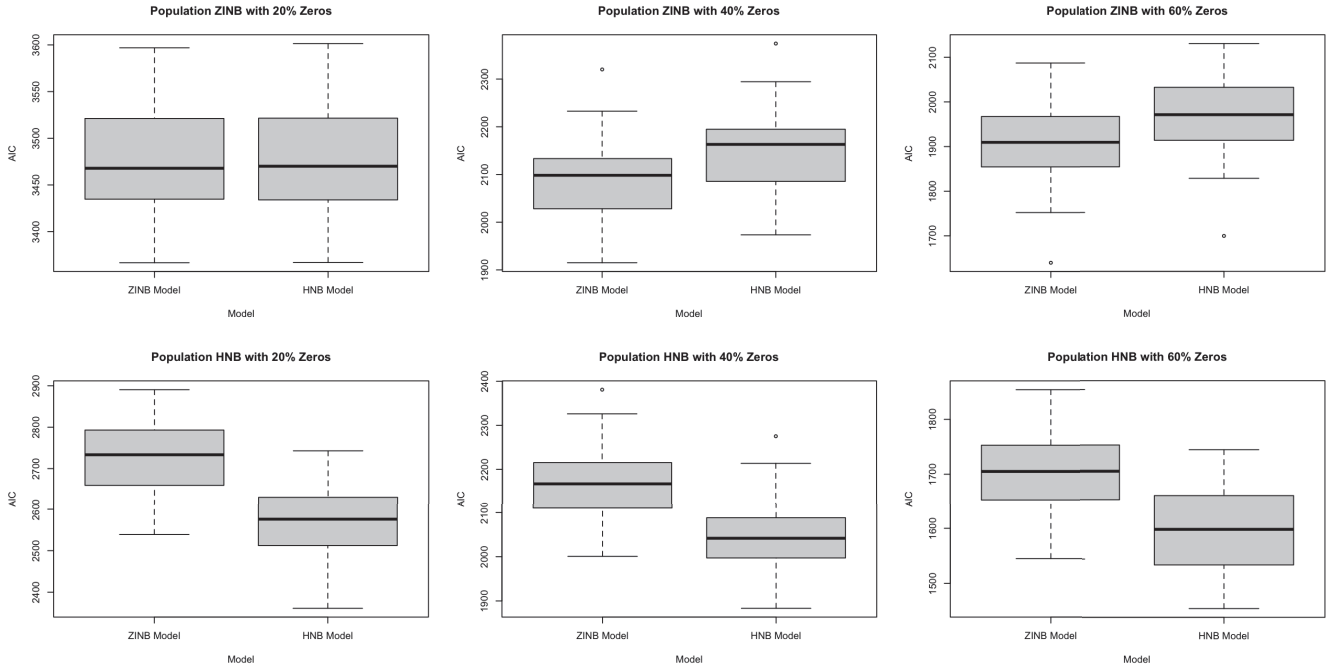


Figure 3. The box plots show how the AIC of the ZINB and HNB models compare with one another under varying conditions. On the top row, the population data was generated from the ZINB model, and on the bottom row, the population data was generated from the HNB model. In this case, $\beta_1 = \gamma_1 = 2$. Each column of box plots corresponds to a different proportion of zeros: 20%, 40%, and 60%. We see in the top row that the ZINB model outperforms the HNB model with the exception of the setting with 20% zero-proportion, and on the bottom row, the HNB model outperforms the ZINB model. We see this result because given a x_i with a large negative magnitude, the ZINB model will predict a sampling zero where the HNB model would not. For example, given $x_i = -5$, $\pi_Z + (1 - \pi_Z)(\frac{r}{\mu_i + r})^r \approx 1$ in **Equation 1**; however, $\pi_H \approx 0$ in **Equation 2**.

tion of zeros (20%, 40%, and 60%), when $\beta_1 = \gamma_1 = 2$, the true model far outperformed the other in terms of AIC (with the exception of when the population data was generated by the ZINB model and the zero-proportion was 20%). In a second simulation study comparing the HNB and ZINB models, our results show that under conditions of zero deflation—where the probability of $Y = 0$ is less than 0.5—the HNB model significantly outperforms the ZINB model in AIC. The difference in AIC values grows exponentially as the proportion of zeros falls below the probability of $Y = 0$ under the standard negative binomial distribution, as illustrated in **Figure 4**. However, it seems that the ZINB model is able to account for moderate zero-deflation without an impact on model performance.

In this setting, we found that when $\beta_1 = \gamma_1 = 2$, the ZINB and HNB were vulnerable to model misspecification as seen in **Figure 3**. We conclude that this trend is due to the fact that the ZINB model will predict a sampling zero given x_i with a large, negative magnitude whereas the HNB model would never predict a sampling zero, causing the difference in AIC. Furthermore, we also observed that in cases of zero-deflation, the HNB model far outperformed the ZINB model, which is displayed in **Figure 4**. However, the ZINB model seemed robust to moderate levels of zero-deflation. This robustness of the ZINB model to moderate zero deflation arises from its ability to adjust parameters such as the dispersion parameter, compensating for a lower-than-expected proportion of zeros. However, as the zero-deflation intensifies, the difference in AIC begins to grow since the ZINB model cannot predict zeros at a probability below that of a standard negative binomial distribution.

SETTING TWO: COMPARING THE HNB AND TLNPN MODELS (WITH HNB POPULATION DATA)

This setting aims to empirically compare the goodness-of-fit of the HNB and TLNPN models when the population data are generated from the HNB model. We seek to investigate how different parameters of the HNB population model affect the relative performance of the TLNPN model to discover both its strengths and weaknesses. We do not consider the ZINB model because it cannot model zero-deflation, which is one of the conditions that we investigate. We also consider the HNB model fitted with and without covariates; however, in most biomedical datasets, covariates are unavailable. We set $n = 1200$ and $p = 5$ and generate covariates $\boldsymbol{x}_i$,

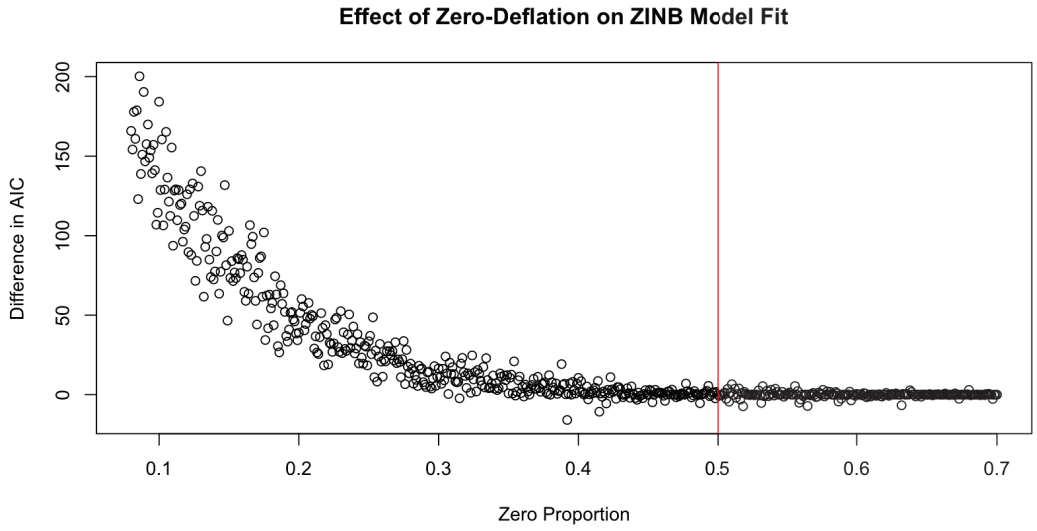


Figure 4. In cases of extreme zero-deflation, the HNB far outperformed the ZINB model. We simulated HNB population data under varying levels of zero-deflation and inflation and fitted both the HNB and ZINB model. We found the difference in AIC between the two models corresponding to each proportion of zeros. The y-axis displays the ZINB model AIC minus the HNB model AIC. Under the standard negative binomial model with $\mu = \ln(6/7)$ and $r = 2$, the probability that $Y = 0$ is 0.5 (the red line). We see that as the proportion of zeros declines below 0.5, the difference in AIC grows at an increasing rate.

$i = 1, \ldots, n$, independently and identically from the multivariate Gaussian with zero mean and covariance matrix given in **Equation 4** and **Equation 5**. We then set $\ln(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$ and $\text{logit}(\pi_{Hij}) = \gamma_0 + \gamma_1 x_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, p$ as in **Equation 2**. We set $\beta_0 = 2.75$ and $r = 6$ for all the simulations, which allows us to evaluate the impact of zero-inflation and zero-deflation. The parameter β_1 controls the impact that the covariate has on μ_{ij} and thus, affects the scale and dependence among the variables. It was set at 0, 1, and 2 in order to evaluate the impact of x_{ij} having no effect on μ_{ij} to x_{ij} having a large effect on μ_{ij}. The parameter γ_0 controls the zero-proportion of the variable and was set at $\ln(1/20)$, $\ln(1/9)$, and $\ln(1/3)$ since we are interested in whether zero-inflation or zero-deflation impacts relative model performance. We also varied the parameter γ_1, which controls how the covariate affects the probability of a structural zero, which impacts the variance of the data. This parameter was set at -0.8, 0.8, and 0. These values were chosen to evaluate the effect of x_{ij} having a negative, positive, and no relationship with $P(Y_{ij} = 0)$, respectively. Finally, we also investigated the impact of the parameter ρ, which controls the amount of correlation between the covariates under both the AR and GD correlation structures, and thus, it affects the dependence among the HNB variables. The parameter ρ was set at 0.01, 0.3, 0.7, and 0.9. We wanted to repeat this simulation study with a GD correlation structure because it can create negative correlation between covariates, unlike the AR correlation structure, and it more closely resembles correlation matrices found in real life.

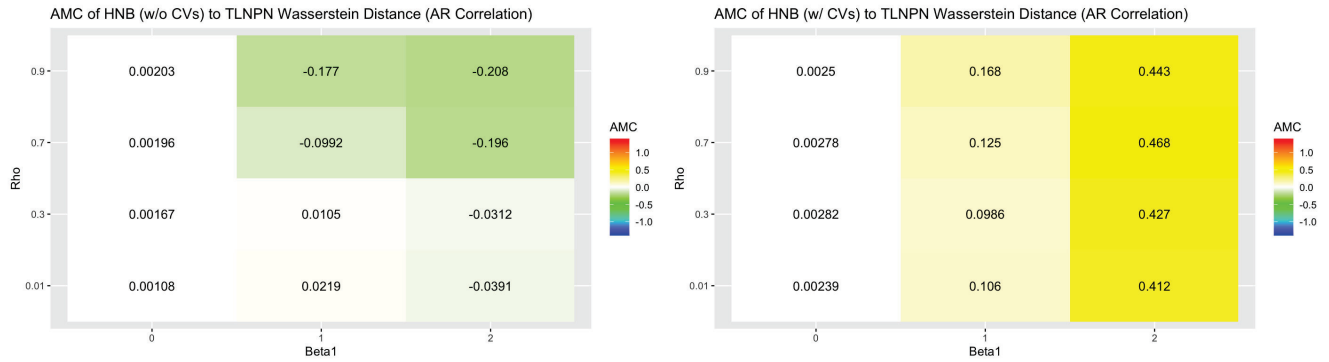


Figure 5. These heatmaps compare the performance of the HNB model to the TLNPN model through our summary measure arithmetic mean change (AMC) given in **Equation 3**. Positive AMC, represented by warm colors, indicates the HNB model performing better, and negative AMC, represented by cool colors, indicates the TLNPN model performing better. When the HNB model is fitted without covariates, we see that as ρ and β_1 increase, the TLNPN model outperforms the HNB model. However, when the HNB model is fitted with covariates, the HNB model outperforms the TLNPN model most as ρ and β_1 increase.

As a goodness-of-fit measure, we consider 5-fold cross-validated prediction error. In particular, we randomly split $n = 1200$ observations into five equal-sized folds, using four of the folds for training and keeping the remaining one for testing. The HNB and TLNPN models are fitted to the training data, and from each fitted model, we simulate a dataset of 240 observations for prediction and obtain the Wasserstein distance between the simulated data from the fitted model and the test data. The described process is repeated for each fold, and we average the resulting five Wasserstein distances to obtain the cross-validated prediction error, which will be used as our performance measure. We repeat this procedure for 30 replicated datasets and summarize the results in **Figure 5** through **Figure 12**.

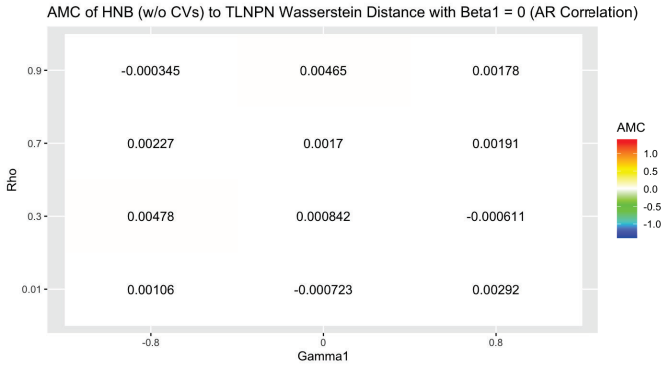


Figure 6. When $\beta_1 = 0$, the HNB model and TLNPN model perform the same. We see this result because when $\beta_1 = 0$, then the HNB variables are practically independent (the γ_1 parameter could still incur a minor amount of dependence between the variables). Therefore, the TLNPN model does not have an advantage in fitting the multivariate distribution as there is very little dependence among the variables for it to model.

We first investigate the interaction between β_1 and ρ and its impact on relative model performance. When we account for whether covariates were included in the fitting of the HNB model, we find starkly different results as shown in **Figure 5**. We see that when covariates aren't considered in the fitting of the HNB model, we see results consistent with our predictions that a high ρ and high β_1 parameter would improve the relative performance of the TLNPN model. We also see that when $\beta_1 = 0$ or $\rho = 0.01$, the models perform nearly the same. However, when covariates are considered in fitting the HNB model, we see notably different results. Still, the TLNPN and HNB models perform nearly the same when $\beta_1 = 0$ with the AMC of the Wasserstein distances being approximately 0 regardless of ρ. However, as β_1 increases, the relative performance of the TLNPN model against the HNB model worsens. Furthermore, it seems that as ρ increases, the TLNPN model performs worse against the HNB model fitted with covariates. We conducted a follow-up simulation study to investigate this trend.

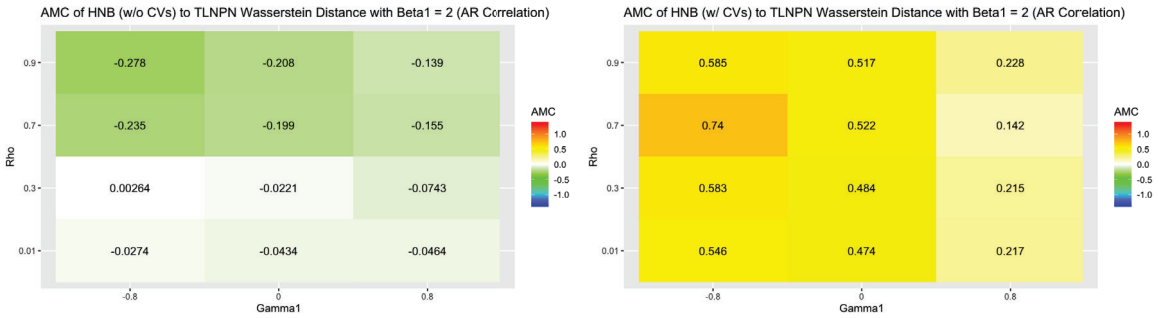


Figure 7. The heatmaps show the relative performance of the TLNPN model to the HNB model under different values of ρ and γ_1. We see that when covariates are not considered when fitting the HNB model, then the TLNPN model performs best relative to the HNB model when $\gamma_1 = -0.8$ and $\rho = 0.9$ (left). When $\gamma_1 = -0.8$, as the covariate, x_{ij}, increases, the mean of the model increases while the probability of a structural zero decreases. Since the covariates for each variable are correlated, this results in stronger correlation among the zero-inflated variables. However, when $\gamma_1 = 0.8$ an observation is more likely to contain a variable with a high value and one that is equal to 0, reducing the dependence among the variables leading to an increase in the AMC. We also see that when covariates are considered in fitting the HNB model, then the TLNPN model performs worse relative to the HNB model as γ_1 decreases and generally as ρ increases (right).

The follow-up simulation study followed the same process as the first except we also measured the marginal Wasserstein distance between the test data and the simulated data and the correlation matrix of the test data and the correlation matrix of the simulated data. In this study, we considered the HNB model fitted with covariates. We found that as β_1 increases from 1 to 2 (when $\rho = 0.9$), the

TLNPN model will further underestimate the correlation between the zero-inflated variables as shown in **Figure 8**. We also found that the one-dimensional Wasserstein distances of the marginal TLNPN data to the marginal test data were greater than that of those generated by HNB data as displayed in **Figure 9**.

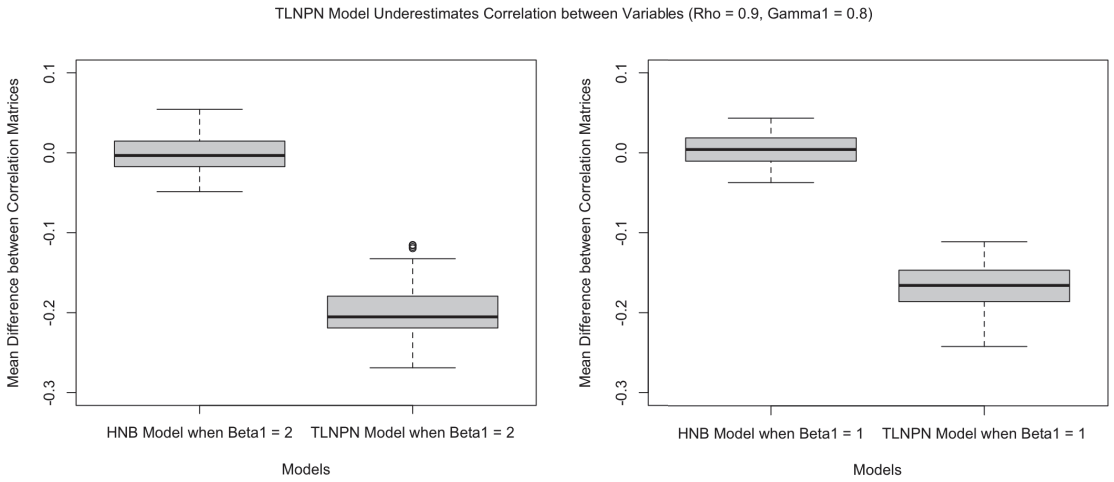


Figure 8. In these box plots, the y-axis measures the mean difference between the values of the correlation matrices between the data produced by the model and the test data. In this figure, the HNB model was fitted with covariates. These box plots show that as β_1 increases, the TLNPN model increasingly underestimates correlation between variables. We suspect this occurs because the latent Gaussian variables in the TLNPN model must fit a latent correlation matrix such that the joint distribution will have data points with variables equal to large positive values and variables equal to 0 on account of the zero-inflation.

We found that the zero-proportion of the variables, controlled by γ_0, did not have an impact on relative model performance between the HNB and TLNPN models. Furthermore, we also investigated the interaction between ρ and γ_1 with regard to the relative model fit between the TLNPN and HNB models when controlling for β_1. We see in **Figure 6** and **Figure 23** that regardless of whether covariates are included in the HNB model fitting, when $\beta_1 = 0$, γ_1 has no impact on the relative performance of the TLNPN model. When $\beta_1 = 2$, we see a trend in **Figure 7**. When covariates are not considered, the TLNPN model performs best when $\gamma_1 = -0.8$ and $\rho = 0.9$. Additionally, as ρ increases, the performance of the TLNPN model against the HNB model improves. When covariates are considered, the trend is reversed where the TLNPN performs worse when $\gamma_1 = -0.8$ and ρ is high. We conducted a follow-up simulation study to investigate this trend.

From our previous follow-up simulation study, we see that as γ_1 increases, the one-dimensional Wasserstein distance between the TLNPN data and the test data decreases as seen in **Figure 9**. To explain the differences in the one-dimensional Wasserstein distances, we conducted another follow-up simulation study in which we compared the distributions of the test data to the HNB and TLNPN simulated data when $\gamma_1 = -2$ and when $\gamma_1 = 2$, and the results are displayed in **Figure 10**. We see that when $\gamma_1 = -2$, we find much higher residuals between the TLNPN and test data; however, when $\gamma_1 = 2$, these residuals decrease dramatically.

The results of this simulation study presented thus far have been generated from covariates following an AR correlation structure; we found similar results from covariates generated from the GD correlation structure. Again, note that in a GD correlation structure, as ρ decreases, the correlation (in an absolute sense) between the covariates tends to increase.

We see in **Figure 11**, that again, there is an interaction between ρ and β_1 such that when ρ is small and β_1 is large, the TLNPN model outperforms the HNB model fitted without covariates. We again see similar results when the HNB model is fitted with covariates where the HNB model outperforms the TLNPN model most when $\beta_1 = 2$ and $\rho = 0.01$. For both scenarios, we see that when $\beta_1 = 0$, the models perform nearly identically.

As with the AR correlation structure, we again see in **Figure 12** an interaction between β_1 and γ_1 where the effect of γ_1 only becomes clear when $\beta_1 \neq 0$ since the models perform nearly identically when $\beta_1 = 0$ regardless of ρ or γ_1. When $\beta_1 = 2$, we begin to see a familiar pattern. When covariates are not considered when fitting the HNB model, the TLNPN model outperforms the HNB model

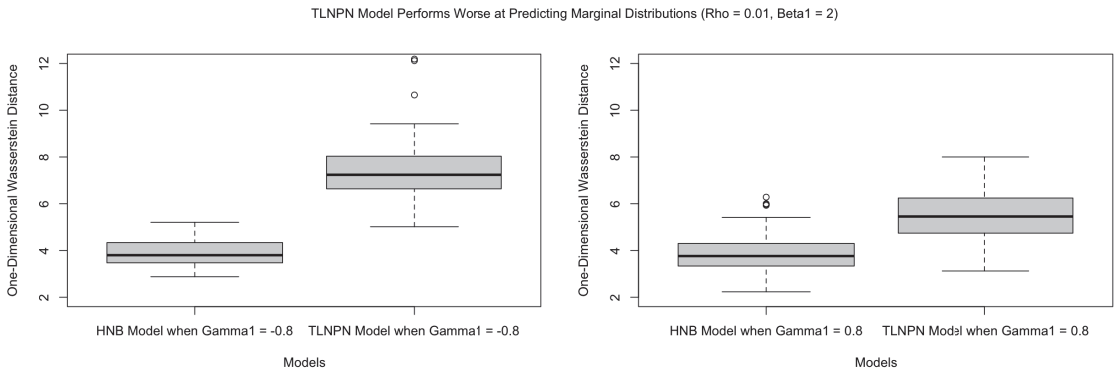


Figure 9. These box plots show the impact that γ_1 has on the marginal performance of the TLNPN model. The y-axis displays the one-dimensional Wasserstein distance between the first variable of the test data and the first variable of the simulated data. We see that as γ_1 increases, the Wasserstein distance between the TLNPN data and test data decreases. This trend is a result of higher, more extreme values become less probable as γ_1 increases, which the TLNPN model struggles to predict since the variance of the HNB model increases as μ_{ij} increases. The HNB model was fitted with covariates, and we see that the marginal Wasserstein distance between the HNB simulated data and the test data stays approximately the same as γ_1 increases.

most when $\rho = 0.01$ and $\gamma_1 = -0.8$ where as ρ and γ_1 decrease, the better the TLNPN model performs relative to the HNB model. However, when covariates are considered when fitting the HNB model, the relative performance of the TLNPN model declines as ρ and γ_1 decrease.

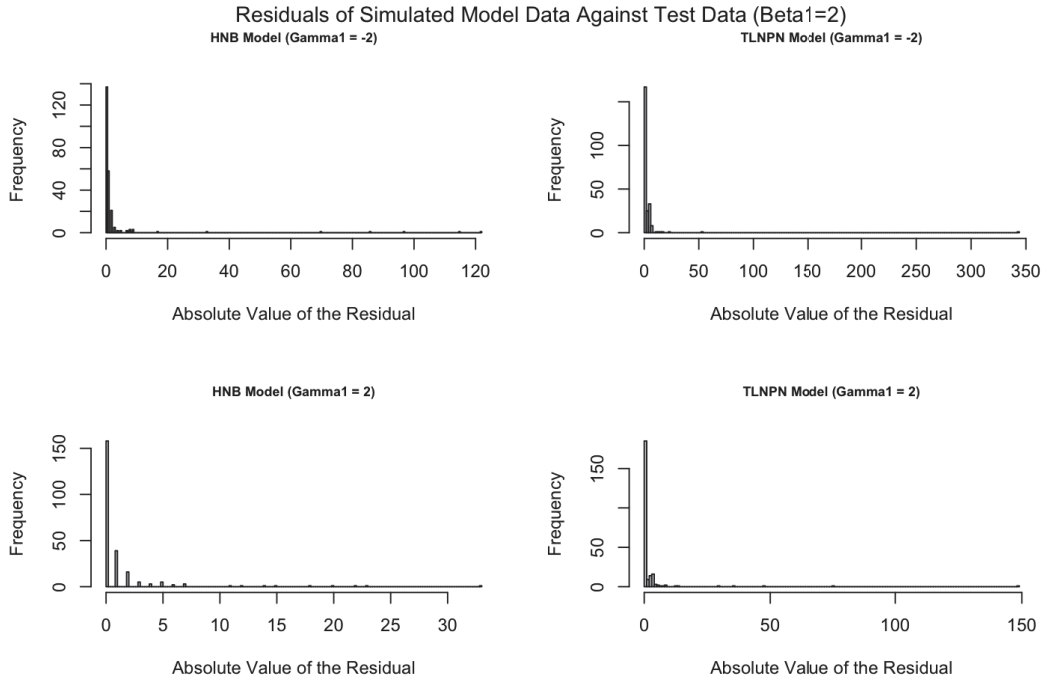


Figure 10. These histograms display how γ_1 affects the marginal performance of the HNB and TLNPN models in the first and second column respectively. The first row shows the difference between the sorted values of the simulated and test data when $\gamma_1 = -2$, and the second row shows the case when $\gamma_1 = 2$. As γ_1 increases, the graphs show the residuals decreasing, particularly for the TLNPN data, which occurs because when γ_1 increases, given $\beta_1 > 0$, extreme values become less likely because as the covariate increases, both μ_{ij} and $\pi_{H,ij}$ increase.

The first main result of our second simulation setting is displayed in **Figure 5** where we see that when the HNB model is fitted without covariates, the TLNPN model performs best when $\beta_1 = 2$ and $\rho = 0.9$ under the AR correlation structure. We attribute this to the ability of the TLNPN model to account for the correlation between the latent Gaussian variables, which can be influenced through the correlation between the covariates. However, for the correlation between the covariates to have an impact on the latent correlation, the β_1 parameter, which controls the impact the covariates have on the mean, has to be nonzero. We see that regardless of

ρ, when $\beta_1 = 0$, the models perform nearly identically because the impact of the correlation between the covariates has no bearing on the correlation of the latent variable since the covariates have no effect on the mean. Similarly, when $\beta_1 = 2$ and $\rho = 0.01$, we see that the models perform nearly the same, because there is a lack of dependence among the covariates and thus, the variables. Therefore, under the AR correlation structure, when both ρ and β_1 are large, we see the best relative performance of the TLNPN model because the variables are more highly dependent on each other since the covariates are highly correlated and the covariates have a large impact on μ.

In **Figure 5**, we also observe that when the HNB model is fitted with covariates, it outperforms the TLNPN model most when both β_1 and ρ are high. We conducted a follow-up simulation study to investigate this trend; the results of which are summarized in **Figure 8** and **Figure 9**. We see in these figures that the TLNPN model underestimates the correlation among the HNB variables and performs worse than the HNB model fitted with covariates on the marginal level. We attribute this trend to the TLNPN model estimating the latent correlation between the latent Gaussian variables through a formula that utilizes Kendall's τ. In contrast, the HNB model was able to much more accurately capture the correlation between the variables. We attribute this to the HNB model using the covariates of the test data when simulating data from the fitted model, resulting in more accurate predictions than those of the TLNPN model, particularly for extremely high values. The TLNPN model can only predict values within its training dataset, which makes it vulnerable to modeling datasets with extreme, outlier values, which are much more probable when $\beta_1 = 2$ as compared to 1 or 0.

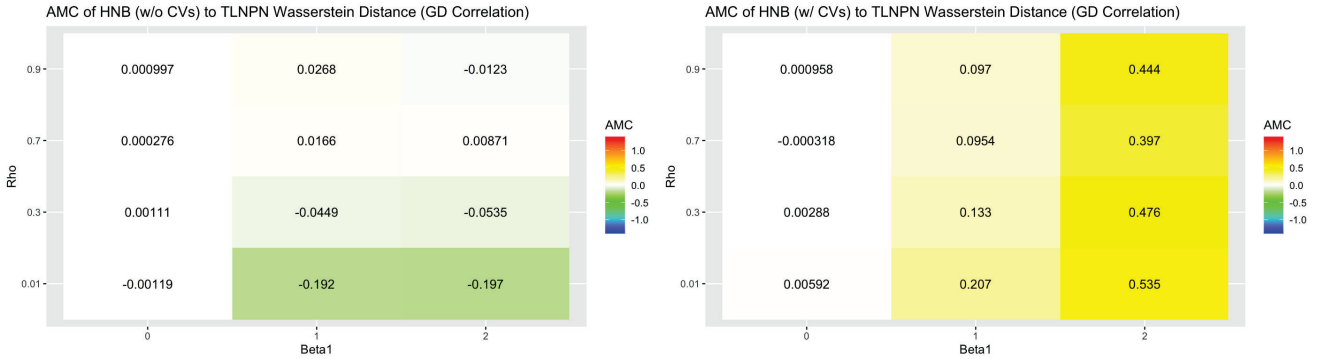


Figure 11. We see in the first heat map that the TLNPN model outperforms the HNB model (fitted without covariates) most when $\rho = 0.01$ and $\beta_1 = 2$. This occurs because under the GD correlation structure, when $\beta_1 = 2$ and $\rho = 0.01$, there is a stronger dependence among the zero-inflated variables. In the second heat map, the HNB model is fitted with covariates. We see that the TLNPN model performs worst relative to the HNB model when $\beta_1 = 2$ and $\rho = 0.01$ since the TLNPN model underestimates the correlation among the variables and performs worse at modeling the marginal distributions.

In **Figure 6**, we see that when $\beta_1 = 0$, the γ_1 parameter seems to have no effect on relative model performance. This is a result of the lack of dependence and extreme values among the variables that results when $\beta_1 = 0$, so the parameter γ_1 can only have a limited impact on the dependence and variance of the variables. However, when $\beta_1 = 2$, we see a clear pattern emerge both when the HNB model is fitted with covariates and when it is not as shown in **Figure 7**. When covariates are not considered when fitting the HNB model, the TLNPN model outperforms the HNB model most when $\rho = 0.9$ and $\gamma_1 = -0.8$. We attribute this trend to the fact that β_1 is always non-negative in our simulations, therefore, if an increase in the covariate both increases the mean and decreases the probability of a zero, then the resulting latent correlation, calculated from Kendall's τ will be much stronger, which the TLNPN model accommodates.

Despite this, when covariates are considered when fitting the HNB model, the pattern reverses, and the TLNPN model performs worse when $\gamma_1 = -0.8$ and ρ is large. We conducted a follow-up simulation study to investigate, and the result is displayed in **Figure 10** where we see that as γ_1 increases, the residuals between the simulated data and the test data greatly reduces, particularly for the TLNPN data. Therefore, there are two trends at work that cause the TLNPN model to perform worse against the HNB model fitted with covariates when $\gamma_1 = -0.8$ as compared to when $\gamma_1 = 0.8$ (given ρ is high and $\beta_1 > 0$). One, when $\gamma_1 = -0.8$, the probability of a structural zero decreases as the covariate, and therefore the mean of the distribution, increases. This increases the correlation between the zero-inflated variables, and the HNB model more accurately describes the correlation structure between the zero-inflated variables than the TLNPN model, which tends to underestimate the correlation between the variables. Two, as the mean of the HNB

distribution increases, the variance increases as well, which will increase the Wasserstein distance between test and simulated data. However, the HNB model is better equipped to predict higher values compared to the TLNPN model because that model uses the same covariates as the test data. We see that when $\gamma_1 = 0.8$, there is an improvement in the relative performance of the TLNPN model against the HNB model fitted with covariates modeling the marginal distribution compared to when $\gamma_1 = -0.8$ as seen in **Figure 9**. Therefore, the TLNPN model performs relatively better when $\gamma_1 = 0.8$ because it makes the occurrence of an extremely high data point less probable, which we see in **Figure 10** where as γ_1 increases, the size of the residuals between the marginal distributions decreases dramatically. We also found that the level of zero-inflation or deflation had no effect on the relative model performance. We conclude that this results from the ability of both models to account for both zero-inflation and zero-deflation.

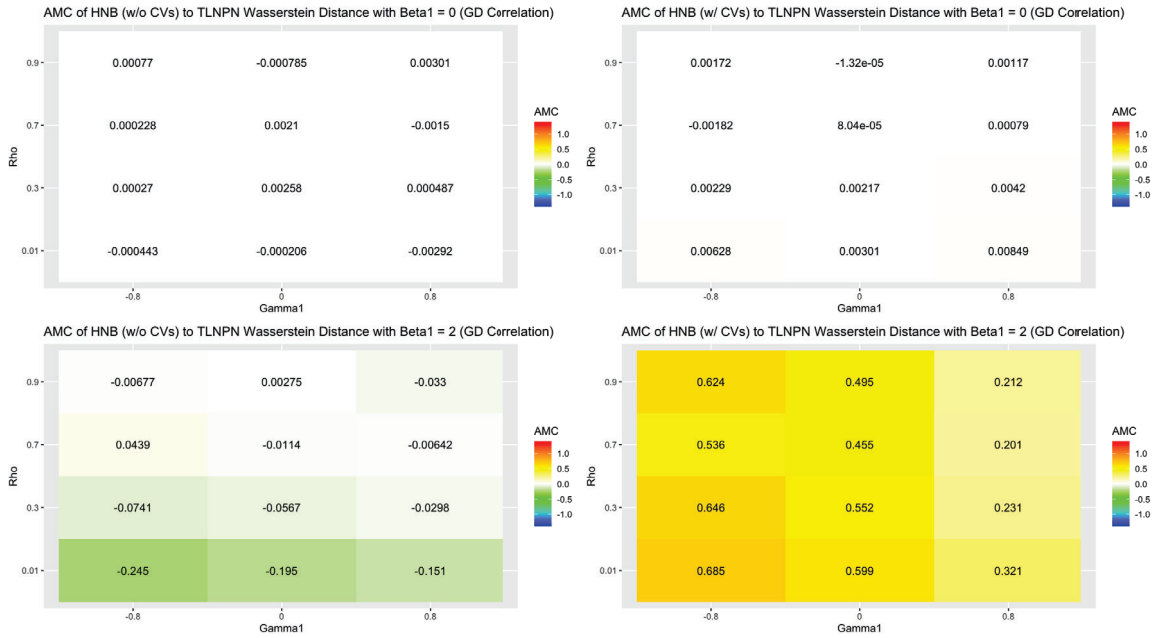


Figure 12. In the left column, the HNB model is fitted without covariates, and in the right column, the HNB model is fitted with covariates. On the top row, we consider when $\beta_1 = 0$, and on the bottom row, we consider when $\beta_1 = 2$. We see on the top row, that the TLNPN model and HNB model perform nearly identically when $\beta_1 = 0$ as the variables are practically independent and have much less variance compared to when $\beta_1 = 2$. We see on the lower left heat map that the TLNPN model outperforms the HNB model fitted without covariates most when $\rho = 0.01$ and $\gamma_1 = -0.8$ as a result of the increased dependence among the variables. We see in the lower right heat map, the TLNPN model performs worst relative to the HNB model when $\rho = 0.01$ and $\gamma_1 = -0.8$ since the TLNPN model underestimates the correlation among the variables and performs worse at modeling the marginal distributions.

We performed the simulation study where the covariates were generated from a GD correlation structure. Under the GD correlation structure, we see results investigating the interaction between ρ and β_1 in **Figure 11** where when the HNB model is fitted without covariates, the TLNPN model outperforms the HNB model when $\rho = 0.01$ and $\beta_1 = 2$. We conclude that this results from the dependence among the HNB variables that results when the covariates are highly correlated and have a large impact on μ_{ij}, which the TLNPN model accounts for, but the HNB does not. We see that when the HNB model is fitted with covariates, the HNB model outperforms the TLNPN model most when $\beta_1 = 2$ and $\rho = 0.01$ since it more accurately describes the dependence structure and can better predict large values. For both scenarios, we see that when $\beta_1 = 0$, the models perform nearly identically as the zero-inflated variables have almost no dependence among each other, and both models fit the marginal distributions similarly well.

We also investigated the interaction between ρ and γ_1 under the GD correlation structure, which is presented in **Figure 12**. The interpretation of these results is the same as the interpretation of the results when the covariates are generated from an AR correlation structure; when $\beta_1 = 0$, then γ_1 has very little impact on the dependence structure of the zero-inflated variables, so it doesn't have an impact on the relative performance between the TLNPN and the HNB models regardless of whether the HNB model was fitted with covariates. However, when $\beta_1 > 0$ and $\gamma_1 < 0$, then the correlation between the zero-inflated variables will strengthen since the higher the mean of the distribution is, the lower the probability of a structural zero. When the HNB model is fitted without covariates, the TLNPN model outperforms the HNB model the most when $\rho = 0.01$ and $\gamma_1 = -0.8$; however, when the HNB model is fitted with

covariates, the opposite is true due to the HNB model better describing the dependence structure and marginal distributions of the zero-inflated variables.

SETTING THREE: COMPARING THE HNB AND TLNPN MODELS (WITH TLNPN POPULATION DATA)

In this simulation study, we again compare the performance of the TLNPN model to the HNB model, but use the TLNPN model as the true model using Quantitative Microbiome Profiling (QMP) data.[2] We evaluated the performance of both models under different conditions. The motivation behind this simulation study is to evaluate how the two models compare under varying parameters of the TLNPN population model, which include zero-proportion, latent correlation, and variance of the training data. We consider both AR and GD correlation structures for the latent correlation matrix. We consider the GD correlation structure in order to simulate correlation matrices that are commonly found in real-world biomedical datasets, and we consider the AR correlation structure in order to more directly evaluate the impact of ρ on the relative performance of the models. We used the empirical CDF of both the original (untransformed) data and the square root transformation of the data as the training data of the population TLNPN model since the QMP dataset is extreme scale data, and we seek to investigate whether the scale and skewness of the data impacts relative model performance. We also vary the proportion of zeros in the TLNPN data (ZP) to evaluate whether zero-proportion has an effect on relative model performance. Finally, we set the correlation parameter, ρ, at different values to evaluate if the dependence of the variables has an impact on relative model performance. For this study, we generated the Gaussian-level variables with a correlation matrix of $\mathbf{\Sigma}$, so $\mathbf{x}_i = (X_{i1}, X_{i2}, \ldots, X_{i5})^\top \sim N_p(\mathbf{o}, \mathbf{\Sigma}), i = 1, \ldots, n$ where $n = 1200$. Let $\hat{F}_j$ be the empirical CDF of the jth variable of the QMP data. We generate data such that $y_{ij} = \hat{F}_j^{-1} \circ \Phi(x_{ij}), i = 1, \ldots, n$ and $j = 1, \ldots, p$ where $p = 5$. In this study, we control for the amount of zeros by subsetting on five variables in the QMP dataset that had the desired zero-proportions, then selecting those to generate the TLNPN data. In particular, for the jth variable, $\hat{\pi}_j = \Phi(\Delta_j)$ where $D_j = g_j^{-1}(\Delta_j)$ and $g_j^{-1} = F_j^{-1} \circ \Phi$, which is how each D_j is selected this in simulation study. We then split the population data into five folds, which we used for five-fold cross validation. We fitted the TLNPN and HNB models to the training data and generated data from each model, and find the respective Wasserstein distances between the simulated data and the test data. The average between the five-folds is then found and recorded.

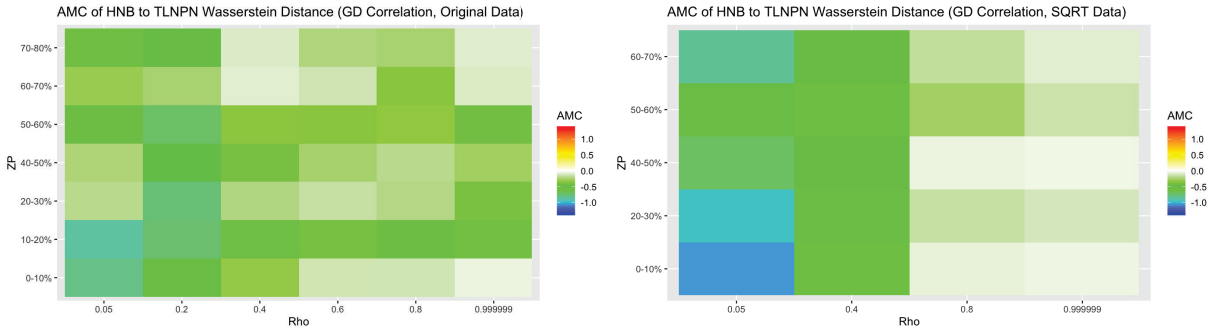


Figure 13. In these heat maps, we compare the performance of the TLNPN and HNB models under TLNPN population data. On the left, the TLNPN population model was trained on the untransformed QMP data. On the right, the TLNPN population model was trained on the square root of the QMP data. We see that in both cases, the TLNPN universally performs better regardless of the zero-proportion of the data or ρ. Although there is no clear pattern in the first heat map, we see in the second heat map, a clear pattern emerge: as ρ decreases and as zero-proportion decreases (when $\rho = 0.05$), the TLNPN model improves its performance relative to the HNB model.

In this simulation study, we investigate the impact that the zero-proportion and correlation of the TLNPN variables have on relative model performance. Our results for the GD correlation structure are presented in **Figure 13**. We see that when evaluating the models under the TLNPN variables distributed as the original, untransformed data, there is no clear pattern; the TLNPN model outperforms the HNB model in all cases, but it's not clear how zero-proportion or ρ impacts the AMC of the HNB Wasserstein distance to the TLNPN Wasserstein distance. However, when we use the square root of the data, we find a much clearer pattern. The lower ρ is when using the GD correlation structure, the better the TLNPN model performs relative to the HNB model.

In **Figure 14**, we present our results for the AR correlation structure. Here, our results are quite similar: the higher ρ is (meaning the higher the correlation between the latent Gaussian variables is), then the better the TLNPN model performs relative to the HNB model. Again, we see that zero-proportion does not reliably impact relative model fit except when $\rho = 0.999999$.

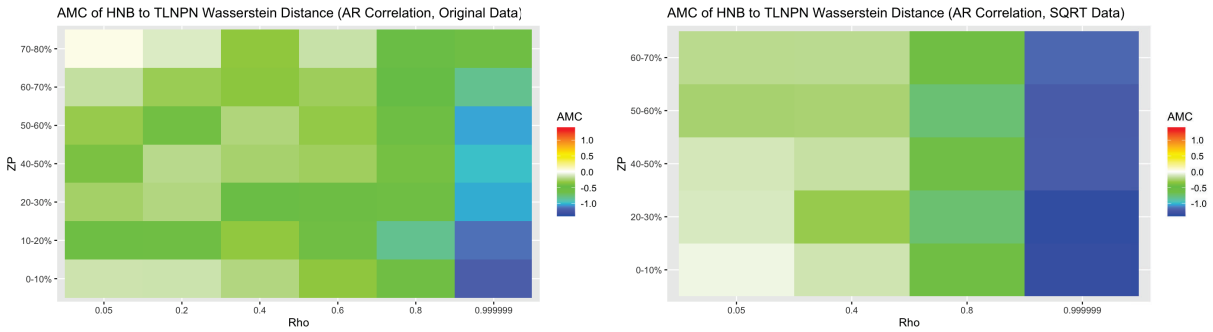


Figure 14. In these heat maps, we compare the performance of the TLNPN model to the HNB model where the population data was generated from the TLNPN model and the latent correlation matrix follows an AR correlation structure. On the left, the population TLNPN data was trained on the untransformed data, and on the right, the population TLNPN data was trained on the square root of the QMP data. We see in both heat maps, the TLNPN model universally outperforms the HNB model, and the TLNPN model performs best when $\rho = 0.999999$ (i.e., when there is strong dependence among the variables). Additionally, we see that the zero-proportion of the variables does not have a reliable effect on relative model performance except when $\rho = 0.999999$ where as zero-proportion decreases, the AMC decreases.

In **Figure 13**, we display results from the third simulation setting where we investigate the effect of zero-proportion, ρ (under the GD correlation structure), and the square-root data transformation on relative model performance. We see that in the square-root transformation, the TLNPN model outperforms the HNB model most when $\rho = 0.05$. We can attribute this to the higher correlation between the zero-inflated variables, which the TLNPN model is able to account for as opposed to the HNB model. However, it still seems that zero-proportion does not have an effect on the relative model performance. We attribute this to the fact that both models can handle zero-inflation or deflation. We display our results for the AR correlation structure in **Figure 14**, and see in both the untransformed CDF and square-root transformation CDF, the TLNPN model improves its performance against the HNB model as ρ increases. We again conclude that this is due to the TLNPN model's ability to model dependence among the zero-inflated variables.

REAL DATA ANALYSES

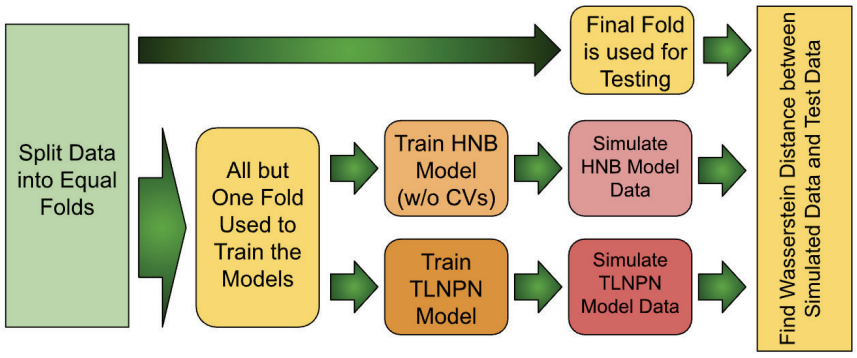


Figure 15. This figure displays a schematic illustration of real data analysis procedure.

In this section, we compare the Hurdle model and the truncated latent Gaussian copula model in their ability to describe real data examples. Our real data studies used datasets from a gut bacteria article[2] and gene-sequencing data (*https://www.10xgenomics.com*). We used regular validation, three-fold for the Quantitative Microbiome Profiling Data and five-fold for the gene sequencing data. We trained the HNB and TLNPN models on all but one fold, then simulated data from those models and found the Wasserstein distance between the simulated data from each model with the final fold. We considered 50 random splits, and our analysis process is graphically summarized in **Figure 15**. We summarize the relative performance of the models using AMC as given by **Equation 3**.

Quantitative Microbiome Profiling Data

As an example of real world zero-inflated data, we use Quantitative Microbiome Profiling (QMP) data.[2] This data measures the number of 101 different genera of gut bacteria in 135 people (29 with Crohn's Disease and 106 controls). We found that a limitation of fitting the HNB model was that the most popular R function used to fit the HNB model (from the `pscl` package) was limited to integers less than or equal to $2^{31} - 1$.[18] We had to rescale the data by taking the power of 0.851 of each data point and then rounding, which

makes the maximum value of the data $2^{31} - 1$. The first, second, third, and fourth quartiles of the zero-proportion of the 101 variables are 3.7%, 28.9%, 57.8%, and 79.3% respectively, and the data displays a high amount of skewness. The result of the analysis is displayed in **Figure 16**.

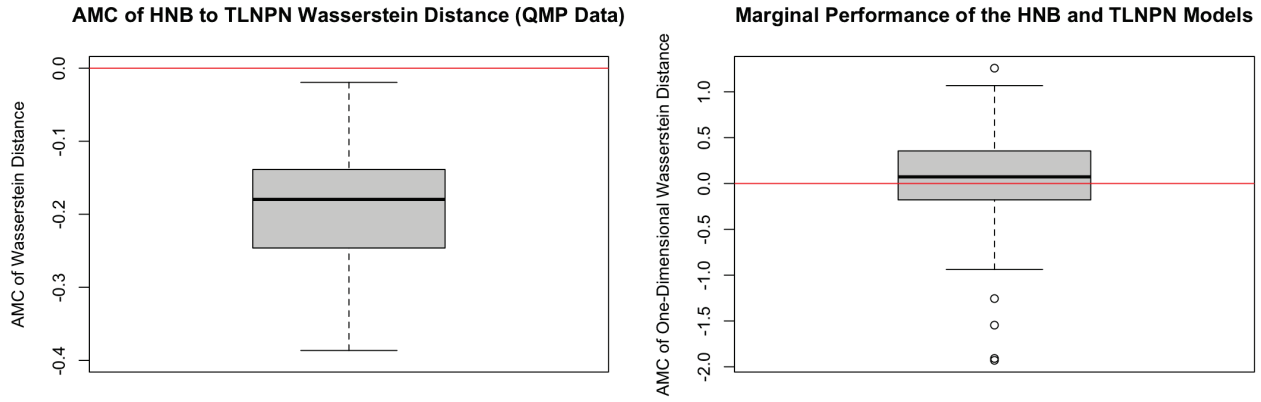


Figure 16. The left box plot shows the Arithmetic Mean Changes (AMC), defined in **Equation 3** of the HNB Wasserstein distance to the TLNPN Wasserstein distance, based on 50 random splits of QMP data. The red line at zero marks the reference: points above indicate a better HNB fit, and points below indicate a better TLNPN fit. The left panel shows that the TLNPN data had a lower 101-dimensional Wasserstein distance with the test data than the HNB data. We found the one-dimensional Wasserstein distance between each of the variables of the test data and each of the variables of the simulated data from the models. The right panel shows the AMC of the HNB to the TLNPN one-dimensional Wasserstein distance for each variable. Both models performed similarly on marginal distributions, but TLNPN model consistently outperformed HNB model on the joint distribution.

We found that the TLNPN model outperformed the HNB model in terms of p-dimensional Wasserstein distance in every replication. We also see in **Figure 16** a box plot of the AMC of the one-dimensional Wasserstein distances of the each of the 101 variables from the HNB model to the TLNPN model.

Our first real data analysis compared the performance of the TLNPN model against the HNB model using the QMP dataset; the results are displayed in **Figure 16** where we see that the TLNPN model outperformed the HNB model with regard to p-dimensional Wasserstein distance under every random split but performed similarly, on average, on a marginal level. We conclude that the difference between the Wasserstein distances was a result of the TLNPN model accounting for dependence among the variables whereas the HNB model does not model dependence among the variables. We see that the two models perform, on average, about the same when modeling the marginal distributions, which rules out the explanation that the TLNPN model outperformed the HNB model due to marginal fit.

To further emphasize this difference, we use the example of the second and fifth variables, which were significantly correlated with each other, and compare how the HNB and TLNPN models modeled their joint distribution as compared with the test data. We see in **Figure 17** that there is a dependence between the variables in the test data, which the TLNPN data captures, but the HNB data does not, leading to a higher Wasserstein distance for the HNB simulated data.

Single Cell RNA Sequencing Data

As another example of real data analysis, we use single-cell RNA sequencing data from the lymphoblastoid cell line; the original data can be found on the 10x Genomics Datasets website (*https://www.10xgenomics.com*). This dataset measures $p = 329$ genes from $n = 265$ cells. The first, second, third, and fourth quartiles of the zero-proportion of the variables in the RNA dataset are 41.1%, 65.7%, 81.1%, and 89.8% respectively. In this case, we did not have to transform the data as all data points were already well below 2^{31} and 95% are below 10. However, we found in **Figure 18** that the TLNPN model and HNB model performed similarly over the 50 replications for the p-dimensional Wasserstein distance. Furthermore, we found that the HNB and TLNPN models performed similarly modeling the marginal distributions of the p variables.

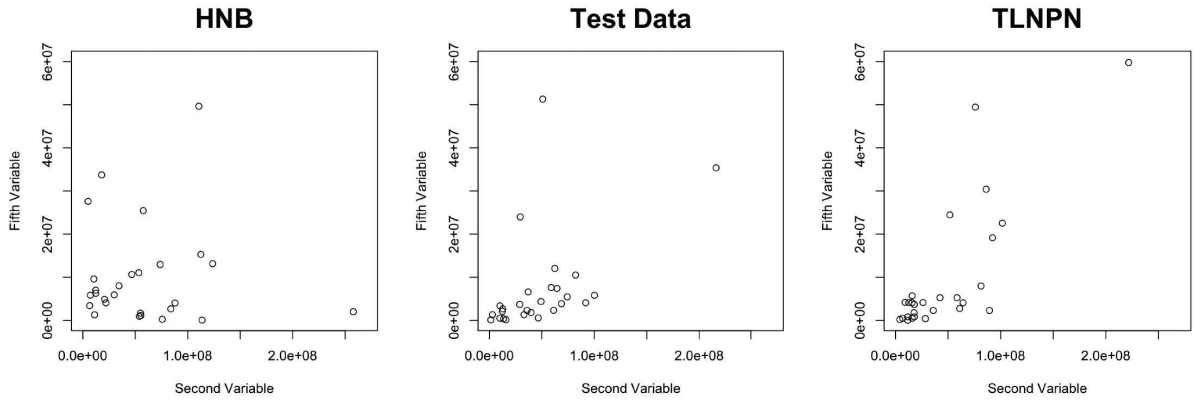


Figure 17. Here, we compare the joint distributions of second and fifth variables of the HNB, TLNPN, and QMP test data to show how the TLNPN model is able to model the dependence of the test data that the HNB model cannot.

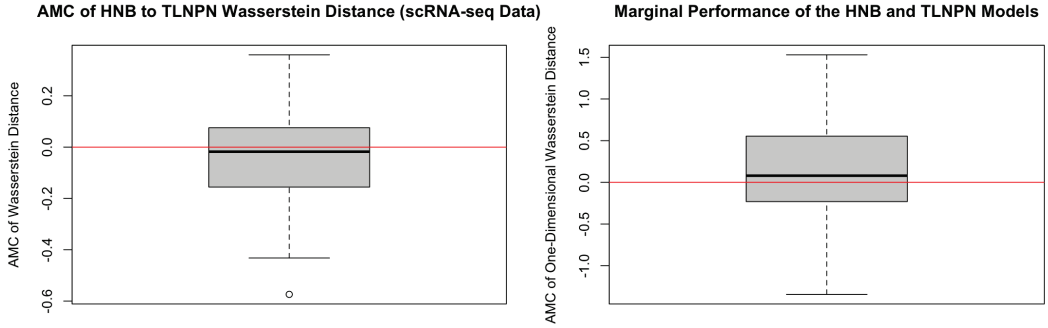


Figure 18. The first box plot compares the overall performance of the TLNPN model to the HNB model under the RNA sequencing data. The y-axis displays the AMC of the Wasserstein distance generated by the HNB model to that of the TLNPN model, defined in **Equation 3**. The results are based on 50 random splits of the gene sequencing data. The red line at zero marks the reference: points above indicate better HNB fit, and points below indicate better TLNPN fit. The second box plot shows how the models compared when modeling marginal distributions. Overall, the HNB and TLNPN models performed similarly both for the joint distribution and the marginal distributions.

In this real-data analysis, we compared the TLNPN model to the HNB model using single cell RNA sequencing data; the results of which are displayed in **Figure 18**. We see that the models perform similarly on a multivariate and marginal level. Based on these results, we conjecture that the characteristics of this dataset, the small scale and a lack of highly correlated variables, resulted in the similar performance of the HNB and TLNPN models. The TLNPN model is more robust to extreme values than the HNB model fitted without covariates, but this dataset had very little skewness and variance in comparison to the QMP dataset, which contributed to the TLNPN model performing similarly to the HNB model.

CONCLUSION

In this work, we sought to compare models for zero-inflated data through both simulation and real data studies that mimicked and used modern biomedical data. Zero-inflated and hurdles models have been popularly used in this field and we sought to compare them with the newly introduced truncated latent Gaussian copula model. The recent emergence of the TLNPN model created a gap in the literature comparing the TLNPN model to the established zero-inflated and hurdle models, and this paper sought to compare these models under different circumstances. Furthermore, in this work, we sought to find the weaknesses of the TLNPN model such as its underestimation of correlation among zero-inflated variables and its struggles in modeling marginal distributions. We found in the simulation studies and real data analyses that the main considerations for deciding to fit either the TLNPN or the HNB models were access to covariates, variance of the data, and dependence among the variables.

The obvious advantage of using the TLNPN model is that it can account for dependence among variables without having access to

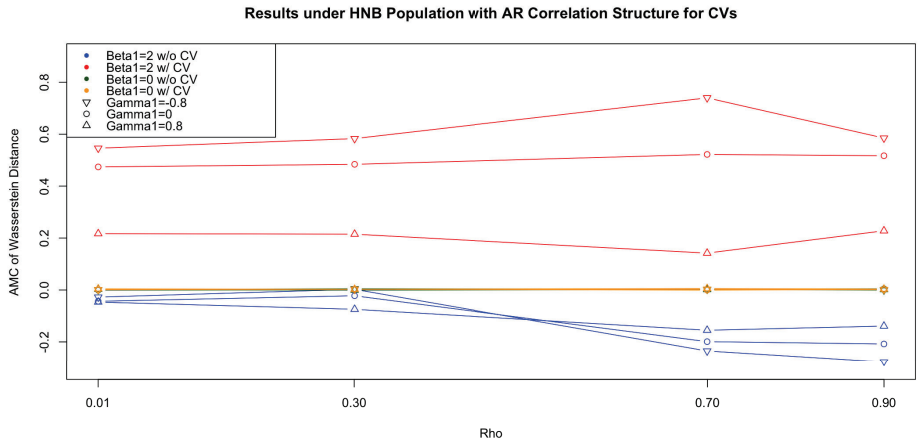


Figure 19. We summarize the results of Simulation Setting Two under HNB population data where the covariates follow an AR correlation structure. We see that when $\beta_1 = 2$, the HNB model fitted with covariates outperforms the TLNPN model; as γ_1 decreases, the AMC further increases. When the HNB model is fitted without covariates, the opposite trend emerges where the TLNPN model outperforms the HNB model, and the AMC decreases as γ_1 decreases (when $\rho = 0.7$ or $\rho = 0.9$). We also see that when $\beta_1 = 0$, the two models perform nearly identically regardless of ρ, γ_1, or whether covariates are fitted in the HNB model.

the covariates in contrast to the HNB model. However, the TLNPN requires a large amount of training data to accurately estimate the correlation of the Gaussian-level variables, and furthermore, in cases of strong dependence among the variables, the TLNPN model tends to underestimate the correlation between the zero-inflated variables when the true model is a HNB model. Furthermore, when the HNB model has access to the covariates, it tends to model the dependence structure between the variables much more accurately. Nevertheless, when no covariate is available, the TLNPN model typically outperforms the HNB model in fitting multivariate distributions of highly dependent zero-inflated variables. We see in **Figure 19** where the HNB model fitted with covariates outperforms the TLNPN model when $\beta_1 = 2$ on account of its ability to better model dependence among the zero-inflated variables and the marginal distributions. However, when $\beta_1 = 2$ and the HNB model is fitted without covariates, the TLNPN model outperforms the HNB model.

Another drawback of the TLNPN model is that when the true population is HNB with a high β_1 parameter, the TLNPN model struggles to model large, outlier values compared to the HNB model fitted with covariates. The TLNPN model will never predict a data point outside of its original training data because of its use of the empirical CDF, so the training data must be similar to the testing data for it to perform well. However, on a marginal level, the HNB model itself can be vulnerable to overdispersed and highly skewed data, which the TLNPN model is better at fitting marginally. Furthermore, a computational limitation of the HNB model is that the main function used to fit the data to a HNB model can only handle integer values below 2^{31}, so datasets with large values may need to be rescaled, as modern biomedical datasets often contain extremely large measurements. This can pose challenges for statistical analysis, as results may vary depending on the rescaling method used.

Future research could investigate how the TLNPN model performs against other models, particularly zero-inflated Poisson and hurdle Poisson models with an overdispersion parameter. Furthermore, investigation of incorporating covariates into the TLNPN model will be an interesting research direction to pursue.

ACKNOWLEDGEMENTS

This research was supported by the 2024 Mathematics Research Experiences for Undergraduates (REU) program at the University of North Carolina at Charlotte under NSF-REU grant DMS-2150179.

REFERENCES

1. Feng, C. X. (2021) A comparison of zero-inflated and hurdle models for modeling zero-inflated count data, *J Stat Distrib App* 8, 8. *https://doi.org/10.1186/s40488-021-00121-4*
2. Vandeputte, D., Kathagen, G. and D'hoe, K., and Vieira-Silva, S., and Valles-Colomer, M., and Sabino, J., and Wang, J., and Tito, R. Y., and De Commer, L., and Darzi, Y., Vermeire, S., Falony, G., and Raes, J. (2017) Quantitative microbiome profiling links gut community variation to microbial load, *Nature* 551, 507–511. *https://doi.org/10.1038/nature24460*
3. Li, H. (2015) Microbiome, metagenomics, and high-dimensional compositional data analysis, *Annu. Rev. Stat. Appl.* 2, 73–94. *https://doi.org/10.1146/annurev-statistics-010814-020351*
4. Peruman-Chaney, S., Morgan, C., McDowall, D., and Aban, I., (2013) Zero-inflated and overdispersed: what's one to do?, *JSCS* 2, 59–67. *https://doi.org/10.1080/00949655.2012.668550*
5. Freckleton, R. P., (2009) The seven deadly sins of comparative analysis, *J. Evol. Biol.* 22, 1367–1375. *https://doi.org/10.1111/j.1420-9101.2009.01757.x*
6. Campbell, H. (2021) The consequences of checking for zero-inflation and overdispersion in the analysis of count data, *MEE* 12, 665–680. *https://doi.org/10.1111/2041-210X.13559*
7. Hua, H., Wan, T., and Crits-Christoph, P. (2014) Structural zeroes and zero-inflated models, *Shanghai Arch. Psychiatry* 26, 236–242. *10.3969/j.issn.1002-0829.2014.04.008*
8. Rose, C. E., Martin, S. W., Wannemuehler, K. A., and Plikaytis., B. D. (2006) On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data, *J. Biopharm. Stat.* 16, 463–481. *https://doi.org/10.1080/10543400600719384*
9. Dong, C., Clarke, D., Yan, X., Khattak, A., and Huang, B. (2014) Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections, *Accident Analysis and Prevention* 70, 320–329. *https://doi.org/10.1016/j.aap.2014.04.018*
10. Yoon, G., Carroll, R. J., and Gaynanova, I. (2020) Sparse semiparametric canonical correlation analysis for data of mixed types, *Biometrika* 107, 609–625. *https://doi.org/10.1093/biomet/asaa007*
11. Fan, J., Liu, H., Ning, Y., and Zou, Hui (2017) High dimensional semiparametric latent graphical model for mixed data, *J. R. Stat. Soc. Ser. B Methodol.* 79, 405–421. *https://doi.org/10.1111/rssb.12168*
12. Saffari, S. E., Adnan, R., and Greene, W. (2012) Hurdle negative binomial regression model with right censored count data, *SORT* 36, 181–194.
13. Liu H., Lafferty J., and Wasserman L. (2009) The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs, *Journal of Machine Learning Research* 10, 2295–2328. *https://www.jmlr.org/papers/volume10/liu09a/liu09a.pdf*
14. Chung H. C., Ni, Y., and Gaynanova, I. (2022) Sparse semiparametric discriminant analysis for high-dimensional zero-inflated data, *arXiv*. *https://doi.org/10.48550/arXiv.2208.03734*
15. Panaretos, V. and Zemel, Y. (2018) Statistical Aspects of Wasserstein Distances, *Annu. Rev. Stat. Appl.* 6, 401–431. *https://doi.org/10.1146/annurev-statistics-030718-104938*
16. Schuhmacher D, Bähre B, Bonneel N, Gottschlich C, Hartmann V, Heinemann F, Schmitzer B, Schrieber J (2024). transport: Computation of Optimal Transport Plans and Wasserstein Distances. R package version 0.15-4, https://cran.r-project.org/package=transport.
17. Chikuse, Y. (2012). Statistics on special manifolds (Vol. 174). Springer Science & Business Media.
18. Zeileis, A., Kleiber, C., and Jackman, S. (2008) Regression Models for Count Data in R. *J. Stat. Softw.*, 27, 1–25. *https://doi.org/10.18637/jss.v027.i08*

ABOUT THE STUDENT AUTHORS

Max Beveridge will graduate from the George Washington University in May of 2025 with a Bachelor of Science in Statistics and International Affairs. Zachary Goldstein will graduate from Yeshiva University with a Bachelor of Arts in Pure/Applied Mathematics and Mathematical Economics in May of 2026.

PRESS SUMMARY

Many modern biomedical datasets have variables that are zero-inflated, and modeling these zeros correctly is critical for accurate statistical analysis. We evaluate three models (zero-inflated negative binomial, hurdle negative binomial, and the truncated latent Gaussian copula models) to see which performs the best under varying conditions. Specifically, we seek to evaluate whether the level of dependence among the variables impacts which model performs the best.

Amplifying Disparities: The Inequitable Burden of Transportation Noise in Rural and Minority Communities

Erin Koster[*a] & *Michelle Stuhlmacher*[a]
[a]*Department of Geography and GIS, DePaul University, Chicago, IL*

https://doi.org/10.33697/ajur.2025.142

Student: ekoster1@depaul.edu
Mentor: michelle.stuhlmacher@depaul.edu

ABSTRACT
Exposure to noise pollution has been linked to a variety of negative health effects including heart disease and hypertension. Exposure to noise is not distributed equally; minority communities are often adjacent to highways and airports and have been found to have disproportionately high levels of transportation noise exposure. What is not yet understood, however, is if transportation noise exposure is increasing or decreasing over time and for whom. In this research we examine the change in transportation noise between 2016 and 2018 in the U.S. with an emphasis on the types of communities impacted. We utilize modeled transportation noise data from the U.S. Department of Transportation and conduct bi-variate regressions with demographic data at the census tract level. Our results show that transportation noise pollution is increasing nationwide, with minority and rural communities disproportionately affected by this increase. We close with a discussion of the policy recommendations for combating the growing inequality in transportation noise exposure.

KEYWORDS
Noise Pollution; Transportation; Environmental Justice; Rural Communities; Minority Communities; Public Health; Geographic Information Systems; Spatial Data; Geography; Data Science

INTRODUCTION

Exposure to elevated levels of anthropogenic noise (*i.e.*, noise pollution) has been identified by the World Health Organization as the second most important environmental risk factor for public health after air quality.[1] Noise pollution has been linked to cardiovascular disease, sleep disturbance, cognitive impairment in children, permanent hearing loss and tinnitus, hypertension, endocrine disruption, and a variety of other chronic conditions.[2, 3] Noise can lead to a detrimental psychological impact and reduce the quality of life for those exposed, as noise annoyance is a direct mediator between exposure to noise and psychological stress.[4, 5]

Noise from cars, trucks, trains, and airplanes makes up a considerable proportion of the anthropogenic noise exposure. In the United States, transportation-based noise pollution is an environmental justice concern.[6, 7] Similar to other environmental injustices, noisy land uses—such as highways, industrial corridors, and airports—have been systematically placed in minority and low-income neighborhoods through processes such as redlining, post-war urban renewal highway construction, and the NIMBY-ism (Not In My Back Yard) of wealthy, White communities.[8–10] Recent research has found that formerly redlined neighborhoods (D-grade) experience three times the level of noise exposure compared to A-graded neighborhoods.[8] Redlined communities were often located in dense urban environments, near transportation corridors, and found to be the subject of discriminatory loaning practices even prior to the formal implementation of these policies.[11] Geographically, the burden of transportation noise has been pushed onto low-income and minority communities, and the close proximity to unwanted land uses like major highways and airports exposes these communities to higher levels of transportation noise.[12, 13] For example, overall noise exposure has been found to be 4-7 decibels higher in minority communities compared to White.[8, 14]

While there is strong evidence of race and socioeconomic disparity in noise exposure,[12, 14] what remains to be understood is where noise exposure is changing over time and for whom. As the population grows, transportation noise is expected to increase, resulting in an increased likelihood for dangerously loud or long exposures to noise.[15, 16] Additionally, a transition towards electric vehicles is projected to lower transportation noise levels in areas where the speed limit is below 30-50 km/hr (19-30 mi/hr), but for rural areas with higher speed limits, the potential noise decrease for electric vehicles will be negligible because rolling noise makes up most of the noise generated at faster speeds.[17] Lack of investment in alternative transportation methods, such as high-speed rail travel, may also contribute to this projected increase.[18] If the spatial distribution of noise increases mirrors the current inequality of noise exposure, minority and rural populations are most likely to be exposed to increased noise.

To fill this gap in understanding of noise exposure change over time, we examine how increases in noise pollution are distributed across American communities, with a particular focus on rural and minority communities because the environmental injustices experienced by these communities are compounded by a lack of access to high-quality health care.[2] For instance, Black and Hispanic individuals are more likely than Whites to report forgoing medical care because of cost[19] and minorities living in rural areas are less likely to have access to routine healthcare compared to both rural Whites and urban minority populations.[20, 21] Using census-tract level data from the American Community Survey, we empirically test transportation noise change over time by asking:

1. Are transportation noise levels increasing in the United States between 2016 and 2018?
2. Do rural, urban, or suburban areas experience the largest change in noise levels?
3. Which communities are disproportionately affected by increasing noise levels?

METHODS AND PROCEDURES

Data Sources

American Community Survey (ACS) five-year data for 2020 (2016-2020) at the census tract level at the census tract level provided the socioeconomic and demographic data. These were downloaded from the Integrated Public Use Microdata Series National Historic Geographic Information System (IPUMS NHGIS).[22] Census tract level data are appropriate for this study due to the nationwide scale of analysis and the uncertainty in smaller populations level estimates (such as census block groups).[23]

From the ACS, we selected several race, ethnicity, income, and inequality variables that are commonly used in environmental justice literature to understand differential impacts and inequalities between demographic groups.[14, 24, 25] Specifically, the variables used for this analysis were the number of households, the number of households that speak limited English, the number of people with a bachelor's degree or greater, total population, Hispanic population, White population, Black population, Native American (American Indian and Alaska Native) population, Asian population, Native Hawaiian and Other Pacific Islander population, population that identifies with another race, and population that identifies as two or more races, median household income, and the Gini coefficient. The Gini coefficient measures income inequality in a range from 0 (perfect equality) to 1 (perfect inequality).[26]

Road and aviation noise for the contiguous U.S. were downloaded from the U.S. Department of Transportation (DOT), Bureau of Transportation Statistics for the years 2016 and 2018.[27] At the time of research, modeled noise maps were available for 2016, 2018, and 2020; the years 2016 and 2018 were chosen for analysis to avoid the decline in traffic and transportation noise levels that occurred as a result of COVID-19 travel restrictions. The dataset is modeled noise levels measured in A-weighted decibels (dBA) averaged into a single "annual day".[28] dBA weights sound frequencies to the same sensitivity as the average human ear.[29] We selected this noise dataset because it is one of the few national level quantifications of transportation noise over multiple points in time. There are, however, limitations to the model. First, the model does not take into account areas where buildings and surrounding materials may naturally dampen sound.[23] Second, the model assumes all areas have acoustically soft ground which may result in an over-prediction or under-prediction of noise levels.[28] We utilized a change-over-time approach in order to compensate for many of these limitations because the same assumptions underly both years and we are examining only the differences between the two models (where noise was modeled to increase or decrease).

Data Preparation

Census Tracts. We classified each census tract in the contiguous U.S. on a rural to urban spectrum based on its population density, with different classifications shown in Oklahoma City, Oklahoma (**Figure 1**). The urban-rural classification method comes from the Housing Assistance Council's (HAC) tract designation.[30] This classification method was used in order to provide a more detailed definition of rural and urban America, as the U.S. Census classification of rural and urban simply defines rural as anywhere that is "not urban".[31, 32] Simplified rural-urban classifications do not capture the nuanced spatial transitions between cities, suburbs, and rural areas- a pattern recognized in new urbanist planning theories which conceptualize these transitions along a "urban-rural continuum" that delineates zones from the most rural to the most urban.[33] The Housing Assistance Committee classification scheme has been used in research determining the distribution of rural demographics, such as by the First Nations Development Institute.[34] Based on their definition, census tracts were divided into one of five different categories: tracts with less than sixteen housing units per square mile are designated as rural, tracts with 16 to 64 housing units per square mile are designated as exurban/small towns, tracts with 64 to 640 tracts housing units per square mile are designated as outer suburban, tracts with 641 to 1,600 housing units per square mile are designated as inner suburban, and tracts with more than 1,600 housing units per square mile are designated as urban.[34] In our analysis, the exurban and small town are a combined category because the degree of commuting was not considered. A potential imitation of this classification is the reliance on population density to determine classification. While non-populated tracts were removed from the dataset, some highly built-up areas may be classified as rural or exurban if they have low population density (*i.e.*, industrial corridors).

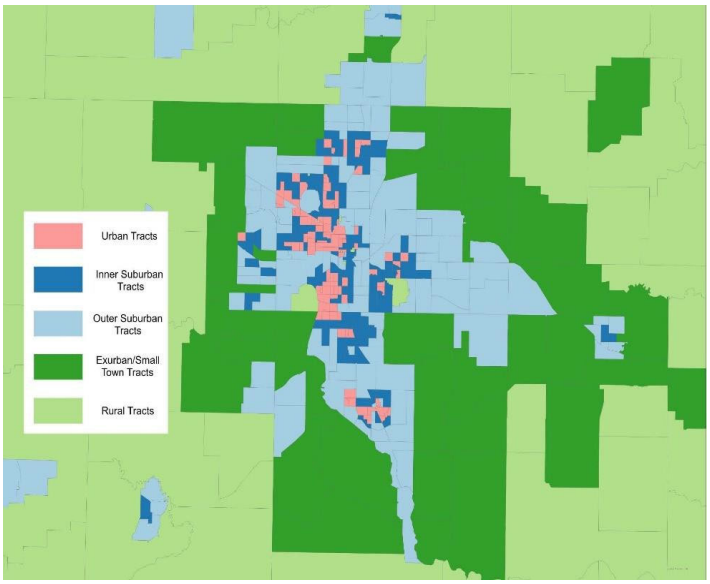


Figure 1. Oklahoma City, Oklahoma's urban to rural designation for census tracts based on the Housing Assistance Committee's Rural-Urban Classification Scheme.

Noise Data. The noise raster for each state was downloaded and mosaiced together to create two nationwide rasters for 2016 and 2018 respectively. In the raw noise raster, areas below 45dBA are "NoData".[28] Because we were interested in change over time, the "NoData" values in the 2018 and 2016 noise rasters were replaced with the value 44dBA, one dBA lower than the smallest recorded noise measurement in the DOT rasters. This allows for a more accurate evaluation of noise level changes when differencing the 2016 and 2018, while remaining conservative in estimation. Next the 2016 raster was subtracted from the 2018 raster to create a single raster showing the noise difference over time. Last, zonal statistics were used to calculate the average change in noise levels for each census tract.

Statistical Analysis. Census tracts with no population were removed[14] and then the Spearman rank-order correlation test was used to analyze the average level of noise change in each census tract compared to ACS variables. We tested the distribution of the data, and the Spearman test was chosen because the data had a non-normal distribution.[35] R version 4.2.1 and packages sf, dplyr, Hmisc, and raster[36–39] were used for analysis along with ArcGIS Pro version 3.1.3 and the ArcPy package.[40, 41]

RESULTS

Changes in transportation noise.

Across all census tracts we found an overall increase in transportation noise between 2016 and 2018. Approximately 9% of census tracts experienced an increase in noise greater than 0.5dBA, which corresponds to ~29.8 million Americans (**Table 1**). We used 0.5 dBA as a threshold because the average increase across tracts was 0.5 dBA. In rural and exurban tracts, noise levels changed from a 5.23 dBA decrease and a 18.3 dBA increase, while urban tracts ranged from a 10.6 dBA decrease to a 17.52 dBA increase.

	Increased (> 0.5 dBA)	**Similar** (-0.5- 0.5 dBA)	**Decreased** (< -0.5 dBA)
All Tracts (n = 83,301)	9.26%	85.21%	5.53%
Urban Tracts (n = 25,901)	16.25%	74.32%	9.42%
Inner Suburban Tracts (n = 19,155)	11.47%	81.70%	6.83%
Outer Suburban Tracts (n = 20,986)	5.17%	91.42%	3.41%
Exurban/Small Town Tracts (n = 9,021)	0.86%	98.48%	0.65%
Rural Tracts (n = 8,238)	1.70%	97.27%	1.03%

Table 1. Percent of census tract where average noise levels increased, decreased, or stayed the same between 2016 and 2018 for all tracts and by level of urbanization.

The proportion of tracts with increased noise was higher for urban (16%) and suburban (11%) census tracts. Approximately 16.6 million urban Americans experienced an increase in average noise levels. Rural and exurban/small town tracts saw the smallest increases (less than 1% and 1.7% respectively), this represents approximately 349,000 rural and exurban Americans.

Characteristics of communities most affected by transportation noise increases.
Table 2 presents the correlations between noise increases and the demographic and neighborhood characteristics for all census tracts in the contiguous United States.

	Coefficient	p-value
Gini Index (income inequality)	0.016	***
Percent college educated	0.083	***
Percent of households with limited English	0.14	***
Percent Hispanic population	0.14	***
Percent White population	-0.11	***
Percent Black population	0.067	***
Percent Native American population	-0.018	***
Percent Asian Population	0.13	***
Percent Hawaiian/Pacific Islander	0.020	***
Median household income	0.039	***

Notes. ***$p <$ 0.001, **$p < 0.01$, *$p < 0.05$, ^$p < 0.1$**
Table 2. Spearman correlation coefficients and p-values for the relationship between socioeconomic variables and the change in traffic noise for all tracts between 2016 and 2018.

	A. Urban Tracts		B. Rural and Exurban/Small Town Tracts	
	Coefficient	p-value	Coefficient	p-value
Gini Index (income inequality)	0.0017		0.029	***
Percent college educated	-0.0021		0.024	**
Percent of households with limited English	0.041	***	0.05	***
Percent Hispanic population	0.041	***	0.11	***
Percent White population	0.0080		-0.049	***
Percent Black population	-0.016	*	-0.053	***
Percent Native American population	-0.019	**	0.079	***
Percent Asian Population	0.0062		0.025	**
Percent Hawaiian/Pacific Islander	0.0034		0.057	***
Median household income	0.0030		-0.042	***

Notes. ***$p <$ 0.001, **$p < 0.01$, *$p < 0.05$, ^$p < 0.1$**
Table 3. Spearman correlation coefficients and p-values for the relationship between socioeconomic variables and the change in traffic noise for urban tracts (A) and rural and exurban (B) between 2016 and 2018.

We find a small, but statistically significant relationships for all variables considered. A greater proportion of White and Native American populations was associated with a smaller increase in transportation noise over the two-year period. In tracts where the percent Hispanic, Black, Asian, and Hawaiian/Pacific Islander populations were larger, the average change in noise was higher. Greater income inequality (Gini index) and a higher proportion of households with limited English were positively correlated with higher average noise change. Tracts with higher median household income and a higher proportion of the population with at least

a bachelor's degree were also positively correlated with noise increases. Among these factors, the relationship between an increased proportion of the Hispanic population and greater average noise levels was the strongest.

Noise, and those impacted, in urban and rural environments may have different explanatory variables, so Spearman tests were conducted separately for the urban tracts (n = 25,901) and the rural and exurban/small town tracts (n = 17,259) separately (**Table 3**). We found both shared and distinct trends in noise change across urban and rural/exurban America, with rural and exurban tracts having greater statistical significance than urban tracts.
Across both urban and rural tracts, a greater proportion of households that speak limited English and higher percentages of Hispanic population were correlated with increasing noise levels. The percent Black population was correlated with a decrease in noise levels across both urban and rural tracts. In rural and exurban tracts (but not urban tracts), a greater percent White population, greater median household income, and lower income inequality was correlated with decreasing noise levels. A greater proportion of college educated individuals was associated with increased noise levels. The percent Native American population and the Hawaiian/Pacific Islander population are both correlated with increasing noise levels in rural tracts, while the Native American population is correlated with decreasing noise levels in urban tracts.

The proportion of households that speak limited English, the percent Hispanic population, and percent Black population are the only demographic variables associated with statistically significance increases in average noise across all our analyzes (national, urban, and rural; **Tables 2 and 3**). To dig into the impact of noise increases on racial and ethnic minorities, we summed the number of Hispanic, Black, Asian, Native American, and Native Hawaiian and other Pacific Islander residents in census tracts experiencing noise increases (**Figure 2**).

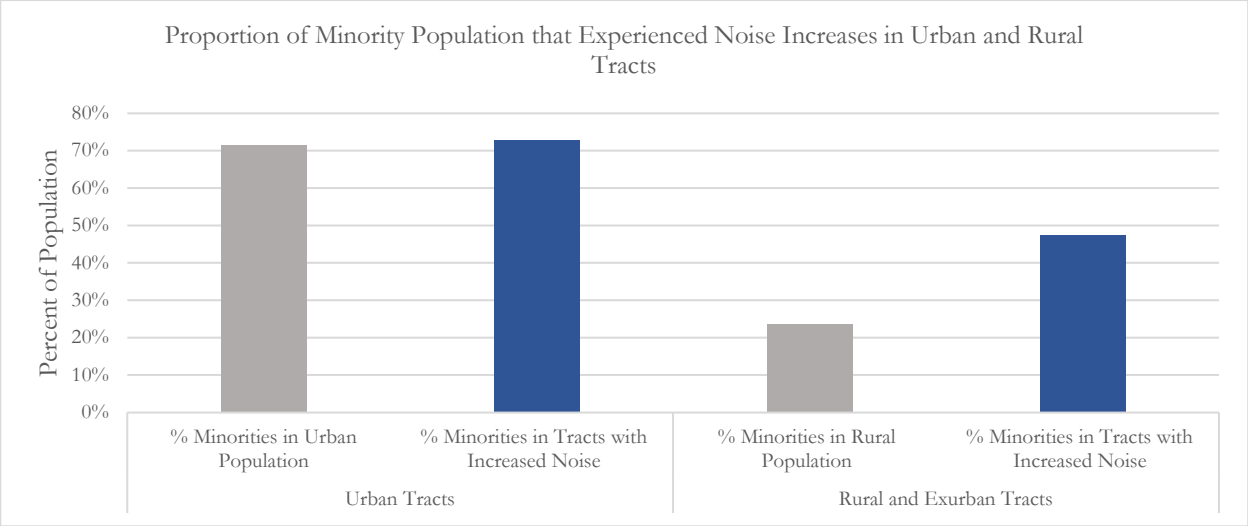


Figure 2. Bar chart displaying the proportion of minority population that experienced noise increases in urban compared to rural/exurban tracts.

In urban areas, racial and ethnic minorities make up 71.64% of the population, and account for 72.77% of the total urban population affected by increased noise levels. In rural and exurban areas, the difference between the proportions is larger: racial and ethnic minorities make up only 23.55% of the population, but account for 47.45% of the total rural and exurban population affected by increased noise levels.

DISCUSSION

Areas Impacted by Increasing Noise Levels.
Nationwide, we found that noise levels between 2016 and 2018 increased by about 0.5 dBA on average. More census tracts experienced an increase in noise levels than a decrease in noise levels across all levels of urbanization, but the majority of tracts that experienced a noise increase were in urban areas (**Table 1**). This finding is consistent with Seto and Huang's[42] result that more populated areas tended to experience higher noise levels. This increase in transportation noise is likely attributable to the growing population and increased use of air and highway transportation, particularly in urbanized areas.[16]

Socioeconomic Status and Noise Levels.
The increases in noise, however, were not distributed equally. We found that socioeconomic status significantly impacts exposure to increasing levels of transportation noise at the national scale. Nationwide, higher levels of economic inequality and a higher

proportion of households that speak limited English were associated with an increase in noise levels. This is consistent with Collins et al.'s[13] results that lower socioeconomic status children were more likely to be exposed to high levels of noise while at school. Other indicators of socioeconomic status—median household income and percent college educated—had the inverse relationship, however. In the national level analysis, higher median household income and proportion of the population with at least a bachelor's degree were correlated with increasing noise levels (the same was also true for percent college educated in rural/exurban tracts). This trend may be a result of louder neighborhoods being viewed as livelier and more desirable places to live. More opportunities for education and employment may attract more residents, resulting in higher population densities with higher noise levels. A similar trend was found by Havard et al.,[43] in which higher road traffic noise exposure increased with home value and education level in many areas of Paris, France. For rural and exurban tracts, however, a higher median income is associated with lower noise average noise changes.

Minority Demographics and Noise Levels.
Consistent with Collins et al.'s[12] and Casey et al.'s[14] findings focused on the populations most burdened by noise, we find that census tracts with a higher proportion of racial and ethnic minority populations are often associated with greater noise increases. Nationwide, census tracts with a higher proportion of Hispanic, Black, Asian and Hawaiian/Pacific Islander populations are associated with higher levels of noise change (**Table 2**). In contrast, census tracts with a larger proportion of White residents are correlated with a decrease in noise for the nationwide and rural area analysis (**Table 2, 3**).

While we found that greater minority populations were often correlated with noise levels increases, there are a few exceptions. At the nationwide level, we found that a greater proportion of the Black/African American population was associated with increased noise levels (**Table 2**), which is consistent with previous research that has found that neighborhoods with a higher percent Black population tend to have higher noise levels compared to surrounding neighborhoods.[14] The trend reversed, however, when looking at the urban and rural/exurban tracts separately—a higher proportion of Black population was associated with decreased noise (**Table 3**). The difference between these two results may be attributable to systemic disinvestment in Black communities that has resulted in a lack of job opportunities, deindustrialization, and other factors that result in population loss.[44] For example, many of Chicago's South Side neighborhoods have experienced net population decline between 1990 and 2016.[44] While many Black and minority neighborhoods have louder noise levels compared to wealthier and whiter neighborhoods;[12] systemic disinvestment has resulted in general population decline which could account for the decrease in transportation noise levels over time. The trends for Native American populations also vary depending on the region type. A greater proportion of Native American residents was associated with higher average noise change in rural/exurban tracts (**Table 3**), but noise decreases at the nationwide level (**Table 2**) and in urban tracts (**Table 3**). These findings are generally consistent with Casey et al.'s[14] study that found the Native American demographic to be correlated with lower noise levels, but inconsistent with Collins et al.'s[12] findings that Native American populations were correlated with increased aviation noise. The differences may be due in part to the uneven distribution of census tracts with high proportion of Native American residents caused by the reservation system.[34]

Rural Minorities and Noise Levels.
Despite the above exceptions, our findings suggest that rural minorities were substantially more likely than urban minorities to be exposed to increasing transportation noise levels. Racial and ethnic minorities make up 23.55% of the population in rural and small-town census tracts, yet they account for 47.45% of the rural and small-town population impacted by increasing noise levels (**Figure 2**). This finding contrasts with Casey et al.'s[14] result that the correlation between noise and higher minority populations is generally consistent between urban and rural/suburban areas. This may be because they examined overall noise while this research examines how noise levels changed between 2016 and 2018.

Rural and small-town minority populations face unique challenges compounded by a variety of factors, which will likely worsen as transportation noise levels continue to increase. First, agricultural noise is not included in the modeled noise data used for this analysis but is a major source of noise exposure for some rural minority populations. A sample of 150 migrant agricultural workers found that over half had some degree of hearing loss.[45] Second, rural and minority populations have decreased access to health care and other community health challenges. Rural Americans suffer from higher rates of death than their non-rural counterparts, in part due to compounding social factors including lack of access to nutritious food, social isolation, and rural poverty.[46] Being non-White further amplifies this issue. For example, Black rural residents are two to three times more likely than White rural residents to die from heart disease, one of the chronic health issues linked to long term exposure to noise pollution.[3, 47] Current noise research rarely focuses on rural areas, and as a result we see the impact of noise on rural minority populations as an important area of future examination.[47, 48]

Policy Recommendations
Noise control policies, regulations, and enforcement vary across the world, but are generally lacking within the U.S. at both the federal and state level. In the U.S. Noise Pollution and Abatement Act of 1970, the federal government delegated power for noise

regulation to the Environmental Protection Agency's Office of Noise Abatement and Control.[49] This department was defunded in 1982, leaving states to draft and enforce their own laws regarding environmental noise pollution.[50] Policies regarding noise control vary by state. Some states, such as Minnesota, have implemented noise control laws with explicit limits on noise in residential areas, while others, like Alabama, have very little state-level legislation and delegate the task to municipalities.[51, 52] Through the analysis conducted for this research, we have identified the census tracts across the U.S. that are most affected by increasing levels of transportation noise and, therefore, have greater need for policy interventions. These census tracts are ones that may merit further analysis and attention because noise can vary dramatically within a small spatial scale. Furthering this goal, we have made the census tract-level results publicly available in a web map to allow residents and policy makers to look-up their neighborhood (**Figure 3**). The web map is accessible at this link.

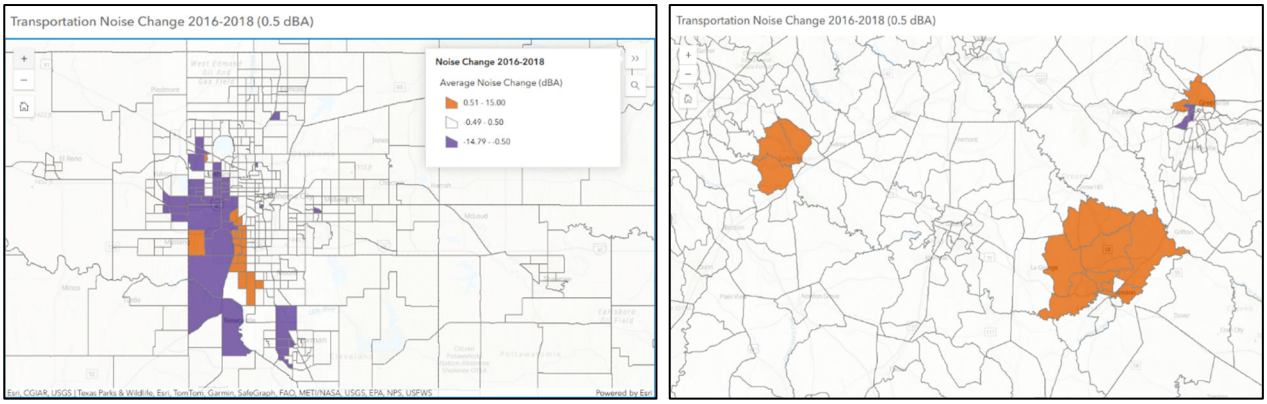


Figure 3. Screenshots of web map displaying Oklahoma City, Oklahoma (left) and Lenoir County, North Carolina (right). The web map shares the census tract-level results of transportation noise change publicly.

Using the web map, we identified that the rural and exurban census tracts with the greatest transportation noise increase during our study period were located near airports. Lenoir County in North Carolina was the rural tract with the highest noise increase of 7.77 dBA followed by Comanche County in Oklahoma with an increase of 6.93 dBA **(Figure 3).** The airports located near these areas of high noise increases (Kinston Regional Jetport and Lawton-Fort Sill Regional Airport) appeared to operate using traditional flight paths in which the same areas are continually subjected to high levels of noise. Measures such as modifying take-off and landing paths or optimizing aircraft type can reduce noise exposure by as much as 26.61 depending on the season and exact procedures implemented.[53]

For ground-based transportation, a potential multi-faceted intervention is the addition of green infrastructure directly adjacent to roadways and rail lines.[54, 55] The leaves and branches of vegetation scatter noise while the bark of trees (as well as soil) absorb noise.[56, 57] The implementation of green infrastructure near transportation hubs as well as in the communities most affected by the stressors of transportation noise will assist in mitigating the disproportional effects of transportation noise and provide a suite of co-benefits such as urban heat island reduction and carbon sequestration.[55, 58, 59] In addition to prioritizing communities with high noise levels and low levels of greenery, areas where speed limits are above 30-50 km/hr (19-30 mi/hr) should also be a focus because noise levels will stay elevated in these areas even as society undergoes a potential transition to quieter electric vehicles.[17] The newly proposed field of transportation forestry is well poised to contribute to these types of interventions. Transportation forestry combines the expertise of a variety of fields to site, select, plant, and maintain green infrastructure along transportation networks in a manner that improves safety and advances environmental justice.[60]

Additionally, city and town design that accommodates alternative (and largely noiseless) forms of transportation—such as walking or biking—can reduce transportation noise in higher density areas. Cities like Amsterdam have seen positive outcomes, including noise reduction, from investment in pedestrian-friendly infrastructure.[61] This is also the case in San Francisco, where city design has a strong influence on residents' choices to make trips by bicycle.[62] Market responsive planning and zoning can facilitate urban landscapes that promote land use diversity and "healthy cities".[62]

Overall, this and other research on the unequal distribution of noise pollution[8, 12–14] points to the need to establish a national framework for measuring transportation noise, setting scientifically informed targets, and ensuring enforcement to reach the targets. While municipal or state level policy action can move the needle, a national level approach could lead to greater equity than the current patchwork of state and municipal noise regulations. For example, the European Union's Environmental Noise Directive created four actions arenas: 1) noise mapping and assessments of noise health impacts, 2) sharing information with the public on noise and it's impacts, 3) preventing and reducing noise, and 4) preserving areas with low noise levels.[56] Each member

state is then responsible for the conducting activities and reporting within these arenas.[63] A similar framework in the U.S. could encourage policy makers, urban planners, transportation engineers, and other urban professionals to prioritize noise and design compliant systems that do not unequally burden communities.

CONCLUSIONS

We find that transportation noise levels between 2016 and 2018 are increasing, with minority and rural communities disproportionately affected by the increased noise. Without purposeful intervention, our analysis suggests that the burden of noise pollution will continue to grow for these communities, compounding other environmental justice concerns such as lack of healthcare access. Many potential interventions to improve noise exposure already exist—such as modifying flight patterns and implementing green infrastructure along transportation networks—what is needed is a policy framework that will guide efforts to limit noise exposure, especially for vulnerable communities.

AKNOWLEDGEMENTS

Funding for this project was generously provided by the Summer Undergraduate Research Grant, sponsored by DePaul University's College of Liberal Arts and Social Sciences.

REFERENCES

1. World Health Organization Europe (2010, April 27) Noise. *World Health Organization https://www.who.int/europe/newsroom/fact-sheets/item/noise*

2. Babisch, W., Bäckman, A., Basner, M., Berglund, B., Fledmann, F.-J., Deshaies, P., Côté, C., Clark, C., Classen, T., Brown, A. L., Botteldooren, D., Bonnefoy, X., Boldo, E., Boegli, H., Bluhm, G., & Berry, B. (2011) Burden of disease from environmental noise: Quantification of Healthy Life Years Lost in Europe. *World Health Organization, Regional Office for Europe*

3. Hammer, M. S., Swinburn, T. K., & Neitzel, R. L. (2014) Environmental noise pollution in the United States: Developing an effective public health response. *Environmental Health Perspectives*, 122(2), 115–119. *https://doi.org/10.1289/ehp.1307272*

4. Sofer, M., Potcher, O., Gnaim, N., & Sofer M, Potchter O, Gnaim N, & Gnaim, J. M. (2012) Environmental nuisances from industrial activities in residential areas of Arab municipalities in Israel. *Applied Geography*, 35(1–2), 353–362. *https://doi.org/10.1016/j.apgeog.2012.08.005*

5. Tao, Y., T Chai, Y., Kou, L., & Kwan, M.-P. (2020) Understanding noise exposure, noise annoyance, and psychological stress: Incorporating individual mobility and the temporality of the exposure-effect relationship. *Applied Geography*, 125, 102283. *https://doi.org/10.1016/j.apgeog.2020.102283*

6. Bullard, R. D. (2000) Dumping In Dixie: Race, Class, And Environmental Quality, Third Edition. *Routledge & CRC Press. https://www.routledge.com/Dumping-In-Dixie-Race-Class-And-Environmental-Quality-Third-Edition/Bullard/p/book/9780813367927*

7. Sobotta, R. R., Campbell H. E., Owens B. J. (2007) Aviation Noise and Environmental Justice: The Barrio Barrier. Journal of Regional Science, 47(1), 125–54. *https://doi.org/10.1111/j.1467-9787.2007.00503.x*

8. Collins, T. W., & Grineski, S. E. (2024) Race, historical redlining, and contemporary transportation noise disparities in the United States. *Journal of Exposure Science & Environmental Epidemiology*. *https://doi.org/10.1038/s41370-024-00682-x*

9. Estrada, G. (2005) If you build it, they will move: The Los Angeles freeway system and the displacement of Mexican East Los Angeles, 1944-1972. *Southern California Quarterly*, 87(3), 287–315. *https://doi.org/10.2307/41172272*

10. Wexler, M. N. (1996) A sociological framing of the NIMBY (Not In My Backyard) syndrome. *International Review of Modern Sociology*, 26(1), 91–110. *http://www.jstor.org/stable/41421101*

11. Xu, W. (2022) Legacies of Institutionalized Redlining: A Comparison Between Speculative and Implemented Mortgage Risk Maps in Chicago, Illinois. *Housing Policy Debate*, 32(2), 249–74. *https://doi.org/10.1080/10511482.2020.1858924*

12. Collins, T. W., Nadybal, S., & Grineski, S. E. (2020) Sonic injustice: Disparate residential exposures to transport noise from road and aviation sources in the continental United States. *Journal of Transport Geography*, 82. *https://doi.org/10.1016/j.jtrangeo.2019.102604*

13. Collins, T. W., Grineski, S. E., & Nadybal, S. (2019) Social disparities in exposure to noise at public schools in the contiguous United States. *Environmental Research,* 175, 257-265. . https://doi.org/10.1016/j.envres.2019.05.024

14. Casey, J. A., Morello-Frosch, R., Mennitt, D. J., Fristrup, K., Ogburn, E. L., & James, P. (2017) Race/ethnicity, socioeconomic status, residential segregation, and spatial variation in noise exposure in the contiguous United States. *Environmental Health Perspectives*, 125(7), 077017. *https://doi.org/10.1289/EHP898*

15. Chepesiuk, R. (2005) Decibel Hell: The Effects of Living in a Noisy World. Environmental Health Perspectives. 113(1) *https://doi.org/10.1289/ehp.113-a34*.

16. Goines, L., & Hagler, L. (2007) Noise pollution: A modern plague. *Southern Medical Journal*, 100(3), 287–294. *https://doi.org/10.1097/SMJ.0b013e3180318be5*

17. Campello-Vicente, H., Peral-Orts, R., Campillo-Davo, N., & Velasco-Sanchez, E. (2017) The effect of electric vehicles on urban noise maps. *Applied Acoustics*, 116, 59–64. *https://doi.org/10.1016/j.apacoust.2016.09.018*

18. Anciaes, P., Alhassan, J-K. (2024) Chapter Seven - Economic and social impacts of public transport investments: A scoping

literature review. *Advances in Transport Policy and Planning*, 13, 227-254. *https://doi.org/10.1016/bs.atpp.2023.12.002*

19. Caraballo, C., Massey, D., Mahajan, S., Lu, Y., Annapureddy, A. R., Roy, B., Riley, C., Murugiah, K., Valero-Elizondo, J., Onuma, O., Nunez-Smith, M., Forman, H. P., Nasir, K., Herrin, J., & Krumholz, H. M. (2020) Racial and ethnic disparities in access to health care among adults in the United States: A 20-year national health interview survey analysis, 1999-2018. *medRxiv: The Preprint Server for Health Sciences*, 2020.10.30.20223420. *https://doi.org/10.1101/2020.10.30.20223420*

20. Caldwell J. T., Ford, C. L., Wallace, S. P., Wang, M. C., & Takahashi, L. M. (2016) Intersection of living in a rural versus urban area and race/ethnicity in explaining access to health care in the United States. *American Journal of Public Health*, 106(8), 1463–1469. *https://doi.org/10.2105/AJPH.2016.303212*

21. Planey, A. M., Perry, J. R., Kent, E. E., Thomas, S. R., Friedman, H., Randolph, R. K., & Holmes, G. M. (2022) Since 1990, rural hospital closures have increasingly occurred in counties that are more urbanized, diverse, and economically unequal. *NC Rural Health Research Program, UNC Sheps Center. https://www.shepscenter.unc.edu/product/rural-hospital-closures-have-increasingly-occurred-in-counties-that-are-more-urbanized-diverse-and-economically-unequal/*

22. Manson, S, Schroeder, J., Van Riper, D., Kugler, T., & Ruggles, S. (2022) National Historical Geographic Information System: Version 17.0 (17.0) American Community Survey: 5-Year Data (2016-2020), Tracts. *Minneapolis, MN: IPUMS. https://doi.org/10.18128/D050.V17.0*

23. Spielman, S. E., Folch, D., Nagle, N. (2014) Patterns and causes of uncertainty in the American Community Survey. Applied Geography, 46, 147-157. *https://doi.org/10.1016/j.apgeog.2013.11.002*

24. Stuhlmacher, M., & Kim, Y. (2024) The unequal connectivity of parks and tree canopy in 10 U.S. cities. *Urban Forestry & Urban Greening*, 96, 128332. *https://doi.org/10.1016/j.ufug.2024.128332*

25. Min, E., Piazza, M., Galaviz, V. E., Saganić, E., Schmeltz, M., Freelander, L., Farquhar, S. A., Karr, C. J., Gruen, D., Banerjee, D., Yost, M., & Seto, E. Y. W. (2021) Quantifying the distribution of environmental health threats and hazards in Washington state using a cumulative environmental inequality index. *Environmental Justice,* 14(4), 298-314. *https://doi.org/10.1089/env.2021.0021*

26. Hasell, J. (2023, July 10) Measuring inequality: What is the Gini coefficient? *Our World in Data. https://ourworldindata.org/what-is-the-gini-coefficient#article-citation* (accessed Jun 2023)

27. United States Department of Transportation. (2022, November 24) National Transportation Noise Map. *Bureau of Transportation Statistics. https://www.bts.gov/geospatial/national-transportation-noise-map* (accessed Jun 2023)

28. Volpe National Transportation Systems Center, Environmental Measurement and Modeling Division, Environmental Science and Engineering Division. (2020) National Transportation Noise Map Documentation. https://doi.org/10.21949/1520433

29. European Union. (2017, February 14) A-weighted decibel. *European Environment Agency Glossary. https://www.eea.europa.eu/help/glossary/eea-glossary/a-weighted-decibel#:~:text=A%20sound%20level%20meter%20that,used%20for%20measuring%20loud%20noise.* (accessed Jun 2023)

30. Housing Assistance Council. (2014, January) HAC's rural and small town typology database. *https://ruralhome.org/wpcontent/uploads/storage/documents/policy_comments/dts/TECHNICAL_DOCUMENTATION_HAC_Rural__Small_Town_Definition.pdf* (accessed Jun 2023)

31. Health Resources and Service Administration. (n.d.) Defining rural population. *https://www.hrsa.gov/rural-health/about-us/what-is-rural#:~:text=The%20Census%20does%20not%20define,UCs%20of%202%2C500%20%2D%2049%2C999%20people* (accessed Jun 2023)

32. United States Census Bureau. (2023, September 26) Urban and rural. Census.gov. *https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html*

33. Geddes, P. (1915) *Cities in Evolution: An Introduction to the Town Planning Movement and to the Study of Civics*. Williams & Norgate.

34. Dewees, S., & Marks, B. (2017) Twice Invisible: Understanding Rural Native America. *First Nations Development Institute. https://www.usetinc.org/wp-content/uploads/bvenuti/WWS/2017/May%202017/May%208/Twice%20Invisible%20-%20Research%20Note.pdf* (accessed Jun 2023)

35. Bishara, A. J., & Hittner, J. B. (2012) Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods,* 17(3), 399–417. *https://doi.org/10.1037/a0028087*

36. Harrell, F., & Dupont, C. (2023) Hmisc: Harrell miscellaneous R package (5.1) [Computer software]. *https://hbiostat.org/R/Hmisc/* (accessed Jun 2023)

37. Hijmans, R. J., Etten, J. van, Sumner, M., Cheng, J., Baston, D., Bevan, A., Bivand, R., Busetto, L., Canty, M., Fasoli, B., Forrest, D., Ghosh, A., Golicher, D., Gray, J., Greenberg, J. A., Hiemstra, P., Hingee, K., Geosciences, I. for M. A., Karney, C., Mattiuzzi, M., Mosher, S., Naimi, B., Nowosad, J., Pebesma, E., Lamigueiro, O., Racine, E., Rowlingson, B., Shortridge, A., Venables, B., Wueest, R. (2021) raster: Geographic data analysis and modeling (3.4-13) [Computer software]. *https://CRAN.R-project.org/package=raster (accessed Jun 2023)raster: Geographic Data Analysis and Modeling [Internet]. 2025 [cited 2025 Mar 27]. Available from: https://cran.r-project.org/web/packages/raster/index.html*

38. Pebesma, E., Bivand, R., Racine, E., Sumner, M., Cook, I., Keitt, T., Lovelace, R., Wickham, H., Ooms, J., Müller, K.,

Pedersen, T. L., Baston, D., & Dunnington, D. (2021) sf: Simple features for R (1.0-1) [Computer software]. *https://CRAN.R-project.org/package=sf* (accessed Jun 2023)
39. Wickham, H., François, R., Henry, L., Müller, K., & RStudio. (2021) dplyr: A grammar of data manipulation (1.0.7) [Computer software]. *https://CRAN.R-project.org/package=dplyr* (accessed Jun 2023)
40. Environmental Systems Research Institute. (2023) ArcGIS Pro (Version 3.1.3) [Software]. *https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview* (accessed Jun 2023)
41. Tateosian, L. (2015) Calling tools with Arcpy. *Python For ArcGIS,* 95–117. https://doi.org/10.1007/978-3-319-18398-5_6T
42. Seto, E., & Huang, C.-H. (2023) The National Transportation Noise Exposure Map. *https://doi.org/10.1101/2023.02.02.23285396*
43. Havard, S., Reich, B. J., Bean, K., & Chaix, B. (2011) Social inequalities in residential exposure to road traffic noise: An environmental justice analysis based on the RECORD Cohort Study. *Occupational and Environmental Medicine,* 68(5), 366–374. *https://doi.org/10.1136/oem.2010.060640*
44. Scarborough, W., Arenas, I., Lewis, A. E., Córdova, T., Ewing, E., Lee, L. Y., Loury, A., Pattillo, M., Ransby, B., Stovall, D., & Sutton, S. (2020) (rep.). Between the Great Migration and Growing Exodus: The Future of Black Chicago? *UIC Institute for Research on Race and Public Policy.* Retrieved September 9, 2023, from *https://news.wttw.com/sites/default/files/article/file-attachments/Black%20Exodus%20IRRPP_StateOfRacialJustice_FutureOfBlackChicago.pdf*
45. Rabinowitz, P. M., Sircar, K. D., Tarabar, S., Galusha, D., & Slade, M. D. (2005) Hearing loss in migrant agricultural workers. *Journal of Agromedicine,* 10(4), 9–17. *https://doi.org/10.1300/J096v10n04_04*
46. Nielsen, M., D'Agostino, D., & Gregory, P. (2017) Addressing rural health challenges head-on. *Missouri Medicine,* 114(5), 363–366. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6140198/*
47. Grundy, S., & Prusaczyk, B. (2022) The complex intersection of race and rurality: The detrimental effects of race-neutral rural health policies. *Health Equity,* 6(1), 334–337. *https://doi.org/10.1089/heq.2021.0136*
48. Alston, L., Field, M., Brew, F., Payne, W., Aras, D., & Versace, V. L. (2022) Addressing the lack of research in rural communities through building rural health service research: Establishment of a research unit in Colac, a medium rural town. *The Australian Journal of Rural Health,* 30(4), 536–539. *https://doi.org/10.1111/ajr.12860*
49. U.S. General Services Administration. (n.d.) Noise control act of 1972 history: Public law 92. *https://www.gsa.gov/cdnstatic/Noise_Control_Act_of_1972.pdf* (accessed Jun 2023)
50. United States Environmental Protection Agency. (n.d.) EPA history: Noise and the noise control act. *US EPA.* *https://www.epa.gov/history/epa-history-noise-and-noise-control-act* (accessed Jun 2023)
51. Minnesota Pollution Control Agency. (2015) A guide to noise control in Minnesota. *https://www.pca.state.mn.us/sites/default/files/p-gen6-01.pdf* (accessed Jun 2023)
52. Alabama Department of Environmental Management. (n.d.) Frequently asked questions. *https://adem.alabama.gov/inside/faq.cnt* (accessed Jun 2023)
53. Xie, J., Zhu, L., & Lee, H. M. (2023) Aircraft noise reduction strategies and analysis of the effects. *International Journal of Environmental Research and Public Health*, 20(2), 1352. *https://doi.org/10.3390/ijerph20021352*
54. Stuhlmacher, M., Woods, J., Yang, L., Sarigai, S. (2024) How does the composition and configuration of green space influence urban noise?: A systematic literature review. *Current Landscape Ecology Reports,* 9, 73-87. *https://doi.org/10.1007/s40823-024-00099-0*
55. Wickramathilaka, N., Ujang, U., Azri, S., & Choon, T. L. (2022) Influence of urban green spaces on road traffic noise levels: A review. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4/W3-2022, 195–201. *https://doi.org/10.5194/isprs-archives-xlviii-4-w3-2022-195-2022*
56. Li, M., Van Renterghem, T., Kang, J., Verheyen, K., & Botteldooren, D. (2020) Sound absorption by tree bark. *Applied Acoustics,* 165, 107328. *https://doi.org/10.1016/j.apacoust.2020.107328*
57. Pérez, G., Coma, J., Barreneche, C., de Gracia, A., Urrestarazu, M., Burés, S., & Cabeza, L. F. (2016) Acoustic insulation capacity of vertical greenery systems for buildings. *Applied Acoustics,* 110, 218–226. *https://doi.org/10.1016/j.apacoust.2016.03.040*
58. Nowak, D. J., & Crane, D. E. (2002) Carbon storage and sequestration by urban trees in the USA. *Environmental Pollution,* 116(3), 381–389. *https://doi.org/10.1016/s0269-7491(01)00214-7*
59. Zhou, W., Wang, J., & Cadenasso, M. L. (2017) Effects of the spatial configuration of trees on urban heat mitigation: A comparative study. *Remote Sensing of Environment,* 195, 1–12. *https://doi.org/10.1016/j.rse.2017.03.043*
60. Browning, M., Wolf, K., Murray-Tuite, P., Browning, L., Chowdhury, M., Chen, C.-F., Coyle, D., Dzhambov, A. M., Fan, C., Hauer, R., Khreis, H., Mack, A., McAnirlin, O., Marshall, L., Middel, A., Nieuwenhijsen, M., Nicolette, M., Rigolon, A., Ogle, J., Ossola, A., Timilsina, N., Thurson, K., White, D., Zhang, K.., Yeager, R. (2024) The Value and Urgency of Transportation Forestry. *https://doi.org/10.32942/x2zs54*
61. Pucher, J., Dijkstra, L. (2003) Promoting Safe Walking and Cycling to Improve Public Health: Lessons From The Netherlands and Germany. Am J Public Health. 93(9), 1509-1516. *https://doi.org/10.2105/ajph.93.9.1509*
62. Cervero, R., Duncan, M. (2003) Walking, Bicycling, and Urban Landscapes: Evidence From the San Francisco Bay Area. *Am*

J Public Health. 93(9), 1478–1483. *https://doi.org/10.2105/ajph.93.9.1478*
63. European Union. (25 June 2002) Environmental Noise Directive. *European Commission. https://environment.ec.europa.eu/topics/noise/environmental-noise-directive_en* (accessed Jun 2023)

ABOUT THE STUDENT AUTHOR

Erin Koster is a senior at DePaul University, graduating November 2024. She studies geography with a concentration in geographic information systems (GIS). She has served as a teaching assistant for introductory GIS and remote sensing courses, a research assistant on a NASA environmental justice grant creating datasets for the ChiVes community dashboard, and a GIS team lead for the Student Urban Research Corps within DePaul University's Urban Collaborative. She is also an alumnus of the NASA DEVELOP program.

PRESS SUMMARY

Noise is an increasing public health concern and is related to serious medical issues like hearing loss and sleep disorders. Our research finds that on average, noise pollution because of transportation has increased between the years 2016 and 2018. The change in noise levels overtime has not previously been studied on a nationwide scale in the United States. Minority and rural populations are disproportionately affected by the transportation noise increase.

Utilizing a Large Language Model for Training Students in Personal Care Product Formulation

*McKinnley Bilbao, Caitlin West, Tomas Carmona, Morgan Covarrubia, Alex Goslin, Katherine Judge, Garland Munn, Hazel Ticas, Abe Tonioli, Collin Tuttle, & Daniel Scott**

Department of Chemistry, Utah Valley University, Orem, UT

https://doi.org/10.33697/ajur.2025.143

Students: 10802009@uvu.edu, 10749330@uvu.edu, 10887896@uvu.edu, morgan.covarrubias@uvu.edu, 10874541@uvu.edu, 10876851@uvu.edu, garland.munn11@hotmail.com, ehazel.ticas@gmail.com, abetonioli@gmail.com, 10959584@uvu.edu
*Mentor: daniel.scott@uvu.edu**

ABSTRACT
This study examines the use of a large language model (LLM), specifically ChatGPT 3.5, to train novice formulators in the development of personal care products. The aim is to assess the LLM's ability to guide students as they formulate a 10-minute hydrating face mask. The research explores how effectively students can rely on the LLM for ingredient substitutions and recipe adjustments during an iterative formulation process, with the goal of producing a high-quality or improved product. Results indicate that while ChatGPT 3.5 demonstrates above-average chemistry knowledge and can provide useful suggestions when prompted clearly, it has significant limitations. These include unreliable memory in extended conversations and difficulty with precise mathematical calculations, particularly for ingredient adjustments. For example, the LLM's limited memory hindered its ability to incorporate information from earlier iterations, often resulting in redundant or inconsistent recommendations. To address these calculation errors, in-house code was developed to ensure formulation accuracy. Additionally, the LLM's contribution to cost optimization was minimal, and it struggled to identify complex formulation components that trained formulators would typically recognize. Although the LLM supported rapid initial product development, it was less effective in more advanced stages, including cost optimization and refining complex components.

KEYWORDS
Machine Learning; ChatGPT; Cosmetics; Formulation; Novice; Formulators; Face; Mask; Education; LLMs

INTRODUCTION

The landscape of personal care product development is complex and multifaceted, involving a delicate balance of ingredient choices,[1] manufacturing processes, and consumer demands. Successful formulation requires careful consideration of efficacy, safety, and consumer appeal. With rising costs and growing demand for innovation, there is an increasing need for methods that streamline development while optimizing both performance and cost. Emerging technologies, particularly machine learning, offer promising solutions by introducing novel approaches to formulation and development. Prior studies have demonstrated the role of artificial intelligence (AI) in optimizing formulations for experienced professionals.[1, 2] AI-driven ingredient selection has been explored in both pharmaceutical and personal care industries, where it has improved efficiency and accelerated development. However, these applications have primarily focused on expert users rather than novice formulators. This research shifts that focus to educational contexts, exploring whether large language models (LLMs) can assist students with limited formulation experience. This aligns with research by Webb *et al.*, which suggests that LLMs can facilitate learning by engaging users in structured problem-solving.[3] Building on these findings, this work evaluates whether LLMs can enhance formulation training and highlights both their potential and limitations.

While Artrith *et al.* discuss the challenges of applying machine learning in chemistry, particularly data reliability and reproducibility,[4] these issues remain underexplored in the context of student training. Machine learning models require robust datasets and specialized algorithms,[4-7] but the extent to which LLMs can assist novice users in experimental design and ingredient optimization has not been well studied. This project aims to bridge that gap by assessing how LLM-generated suggestions align with experimental outcomes and whether structured prompting strategies can mitigate common limitations. LLMs are proficient in tasks involving pattern recognition, analogy, and abstract reasoning.[3] These models can process large volumes of information and generate coherent outputs aligned with user instructions. However, the rapid shifts in personal care formulation—driven by ingredient availability, regulatory changes, and consumer trends—make integration of LLMs challenging. Effective use of LLMs

depends heavily on structured prompting.[8, 9] This research examines how novice formulators adapt prompt engineering techniques to improve the usefulness of LLM-generated suggestions.

This study explores the use of LLMs to propose changes to an initial product formula for a 10-minute hydrating facemask and examines their effectiveness as a tool for novice formulators. Through an iterative process of LLM-guided ingredient substitutions and experimental testing, it was assessed how well LLMs can support formulation training in an educational setting. The findings provide insight into the viability of LLMs as a guide for student formulators in producing quality formulations while identifying their inherent limitations. The significance of this work extends beyond the formulation of a single product, offering valuable insights for prospective formulators, formulation chemists, and educators involved in training students in laboratory settings.

METHODS AND PROCEDURES

Machine Learning Preface

The use of LLMs in scientific research has prompted calls for rigorous standards to ensure repeatability and maintain scientific integrity.[4] In typical experiments, the dependent variable is observed, while the independent variable is manipulated. However, LLM-based experiments introduce a challenge: if the dependent variable is the input prompt and the independent variable is the LLM output, the reproducibility of results may be questioned due to the variability and lack of standardization in LLM responses. Artrith *et al.* emphasize that machine learning-driven experiments in chemistry must adhere to rigorous documentation protocols to ensure reproducibility. This is particularly relevant for LLM-based formulation studies, where variations in responses can impact experimental reliability.[4]

In this study, the LLM's responses to prompts were not considered direct data points but were treated as independent variables. The dependent variable was the resulting product specifications derived from the efforts of the students following the LLM-generated suggestions. As such, the primary measure of success was not the content of the LLM responses but rather the quality and speed of product development when LLM-generated directions were followed. Webb *et al.* highlight how LLMs can enhance problem-solving by engaging users in structured iterations, making them well-suited for assisting novice formulators in experimental adjustments.[3] Thus, the fitness of the final product is the data used to evaluate the LLM's utility as a tool for enhancing the formulation process, particularly in the hands of novice student formulators.

The formulation process, in this context, is defined as transforming an initial recipe into a final product. An improvement in this process is observed when a product with better specifications is created in a shorter time using LLM-generated instructions. Given the limitations of LLMs in numerical accuracy,[10] qualitative assessments are necessary to determine the practicality of LLM-generated formulation strategies. This approach aligns with common industry practices where sensory characteristics and stability are key determinants of product success. Although much of this work involves qualitative or semi-quantitative assessments, such evaluations are common in the development of personal care products. The specifications of the resulting product, as the formulation process proceeds, serve as an indicator of the LLM's effectiveness in training novice formulators.

Machine Learning for Formula Generation

The LLM utilized was ChatGPT, which is currently free and open source to the public.[11] Although other LLMs were considered, the most likely contenders considered were: 1) Agent GPT: *https://agentgpt.reworkd.ai/*, 2) Literally Anything *https://www.literallyanything.io/*, and 3) Ora: *https://ora.ai/dashboard*. ChatGPT 3.5 was chosen because of its popularity, familiarity, and because it is currently always free.[12] It is unlikely, however, that ChatGPT 3.5 is inherently superior to the other three platforms, or potentially many others, for this specific task.

ChatGPT 3.5, based on the GPT-3.5 architecture, has a stated prompt limit of approximately 600 words.[13] As a natural language processing model, it excels at generating human-like text in multiple languages.[14] Its capabilities include content creation, answering conceptual questions, assembling explanations, and assisting with basic coding tasks.[11] ChatGPT 3.5's knowledge base is limited to information available up to January 2022, and it lacks the ability to process human emotions or experiences.[11] Because its responses are generated from pattern recognition rather than reasoning, independent verification of all output is essential. One of the model's most significant shortcomings in chemical formulation is its inaccuracy in performing calculations and its tendency to fabricate references and content.[11]

To compensate for these weaknesses and maintain repeatability when making decisions about formula changes based on LLM feedback, Artrith *et al.* recommend following a checklist when using LLM tools in chemistry. They recommend that each researcher establish a mechanism for clear results and reporting when using LLM models like ChatGPT.[4] These mechanisms could be considered when documenting any formulation adjustments made by LLM. This includes, most importantly, fact-

checking suggestions by searching for the chemical names in publicly available libraries to ensure the safety of the components and weighing the environmental and consumer experience impact before adopting any changes.

Formula Synthesis

This study began with a base formula for a face mask provided by a local professional cosmetic chemist. The face mask was designed to be applied for 10 minutes and washed off with water. While not intended as an overnight mask, the formula was safe enough that leaving it on for longer would not pose any issues. Additionally, the product was versatile enough to be used as a moisturizing lotion for hands or body, rather than solely as a face mask.

With this initial formula, three sets of LLM experiments were proposed. The first experiment employed an LLM to generate a modified version of this formula, utilizing existing materials, to act as the starting formula for subsequent experiments. Identifying the role of each ingredient in the original formula was a subset of this experiment. Identifying formula replacements that met desired specifications was the focus of the second set of experiments. Following the models' recommendations, products were formulated using the suggested ingredients, and the resulting formulas were analyzed for properties such as specification adherence, stability, and microbial resistance. The third set of experiments was focused on lowering the cost of the product while maintaining the quality of the desired specifications. The prompts in this set of experiments attempted to blend price findings through research outside the LLMs with the feedback provided by the LLMs.

Figure 1 illustrates the progression of these ingredient modifications across multiple iterations. The initial base formula was adjusted based on LLM recommendations to match available ingredients, and subsequent iterations incorporated additional refinements. The flowchart highlights how certain substitutions, such as emulsifiers and humectants, were optimized in response to experimental observations.

When an LLM generates false information or miscalculates data it is commonly known as a hallucination. Multiple codes were written to correct mathematical hallucinations associated with the formula suggestions. Castro *et al.* demonstrated that ChatGPT does not always provide curated data and that it struggles with understanding the context of chemistry prompts.[15] During experimentation, the LLM-created recipe iterations were provided in a format that did not translate directly to lab work. To test these recipe iterations, a code was created that translated the LLM recipe formula into percentages and grams to ensure the required mass of the suggested substitution and to verify the desired total mass of material was created.

Testing Different Iterations

Ten novice student formulators, divided into groups of two or three, were tasked with creating multiple iterations of a recipe using an LLM. Different groups were responsible for generating each iteration, with some recipes duplicated for comparison purposes. During this process, different formulators were trained in the production of the base formula and how to query LLM to make changes to the formula based on specification enhancements or cost. Gradual refinements in wording, structure, and information included in prompts led to the development of a prompt template which helped maintain consistency in the information fed to the LLM across different inquiries. The template, found in **Table 1**, remained fluid and was updated as necessary throughout the experimentation.

Once an iteration was completed, it was passed to another individual for specification testing. To ensure consistency and maintain specification integrity, one individual consistently performed this testing. While this aspect introduced some subjectivity, efforts were made to standardize the process by providing clear guidelines for evaluation to ensure consistency. The qualitative properties evaluated included odor, skin hydration, and after-wash effect, along with any other noticeable differences between recipes. Additionally, the pH and viscosity of the iterations were measured. The iteration viscosities were measured using an NDJ-5S digital viscometer, with readings taken immediately after the probe was immersed in the sample and again after two minutes. The pH was measured using a pH indicator paper. The cost per item and batch were recorded for each iteration produced.

Following qualitative testing, each formula underwent quality control assessments for microbial growth and stability. Upon completing these tests, each iteration was subject to one of two pathways: (1) if undesirable changes were detected, the formula was returned for further LLM adjustments; or (2) if the changes were favorable, the formula was presented to the remaining formulators (n=10) and a small set of consumers (n=4) for additional feedback. This feedback informed future decisions and directed further LLM iterations, continuously refining the recipe.

Formula Testing

To ensure the formulas created were stable and microbe resistant, challenges were performed regarding the stability and microbe propagation of the formula. These were used to get a general direction needed in terms of ingredients used in an LLM prompt. The stability test was a one-week test at 50°C with no added humidity. The product was well covered, and dehydration resulted in

condensation of water on the top of the container and an expansion of the container lid, indicating a slight increase in the pressure on the sample and a decrease in the product's water content as the challenge proceeded. The micro challenge was performed by inoculating 5 mL of sterile Luria–Bertani (LB) medium with approximately 50 μL of the product to be challenged. The resulting mixture was swirled at 37°C for 24 hours, and the sample's clarity and odor were observed. Bacterial growth was also detected in control samples to verify that the method was reliable. If there was any question about whether microbial growth was occurring, the clarity of the solution was quantified using a visible spectrometer to monitor the solution's absorbance at 600 nm.

General Template Used for LLM Prompts
The following text is the recipe for a gel-based cosmetic face mask: Purified water: 70.65% EDTA: 0.15% Glycerin: 6.00% Butylene Glycol: 3.00% Xanthan gum: 0.6% Guar Gum: 0.6% AntiMicrobial Banana mixture: 0.90% Sodium Lauryl Sulfate: 1.00% Mango Butter: 3.00% Olive Oil: 5.00% Coconut Oil: 6.00% Vitamin E: 0.50% Polyglyceryl Oleate: 2.00% Papaya Banana: 0.10% Protein-Hyaluronate blend: 0.50% The desired yield of this recipe is 50 grams. Suggest alterations to this recipe based on the following criteria: *USER TEXT HERE*

Table 1. Standardized prompt template used to generate LLM-driven formulation suggestions. Ingredient percentages reflect a typical starting point, and the "USER TEXT HERE" placeholder indicates where students inserted specific formulation goals (e.g., ingredient substitution, cost reduction). This consistent structure improved the quality and relevance of LLM responses across iterative experiments.

RESULTS

The LLM was successful in adjusting a base recipe of a cosmetic formula to achieve a stable starting point from a provided library. **Figure 1** demonstrates the ingredient evolution from the base recipe to an optimized formulation. This process involved systematic substitutions and adjustments in emulsifiers, stabilizers, and active ingredients, ultimately leading to a more refined final product. In total, approximately 19 LLM-driven iterations of the altered formula were produced, with 17 of the iterations chosen for further analysis. These iterations demonstrated LLM's effective proposition of adjustments to a recipe, leading to a sellable product after two iterations. Additionally, iteration 17 formulated using LLM suggestions created a product that closely aligned with the ideal specifications (see **Table 2**). It is important to note that formulators relied on a template prompt to achieve these results (see **Table 1**)**.** This improved both the memory and relevance of the LLM response by maintaining consistent information input. Although generated iterations typically produced a change in the product specifications, this change wasn't always an improvement on previous iterations. The quantitative and qualitative data collected from the iterations are summarized in **Table 2**. Stability testing revealed some stability issues in the iterations produced. For example, sample five exhibited significant separation after the stability challenge (see **Figure 2**). The absence of certain materials may have influenced the results, as not all formulations suggested by the LLM could be tested. Still, this process provided significant value by highlighting areas for specific improvements in the formulation, demonstrating how the LLM can accelerate product development, even with a restricted material set.

The LLM's ability to retain information from previous interactions, even within the same chat session, was limited. This prevented the LLM from leveraging information from previous discussions or remembering past formulations to make more complex decisions based on previous results. However, this can be avoided by using prompt engineering software.[16] Specifically, prompt engineering techniques, such as incorporating key information from previous interactions into subsequent prompts or using a structured template to ensure consistency in the input provided to the LLM could have potentially improved the model's ability to retain the relevant information throughout the iterative process. However, due to the exploratory nature of this initial study, these techniques were not fully implemented.

It is important to note that the limitations discussed in this study, particularly the LLM's difficulty with memory retention, mathematical calculations, and contextual reasoning, reflect the specific performance of ChatGPT version 3.5, which was the model used during the research period. Since then, newer iterations have demonstrated substantial improvements in these areas, including better internal memory within sessions, greater numerical accuracy, and enhanced capacity for multi-step reasoning. While these advancements may mitigate some of the challenges observed here, the findings remain relevant for educators and researchers working with freely available or entry-level models. Future studies could evaluate whether newer LLMs offer more robust support for formulation training and how their capabilities shift the balance between student autonomy and model oversight.

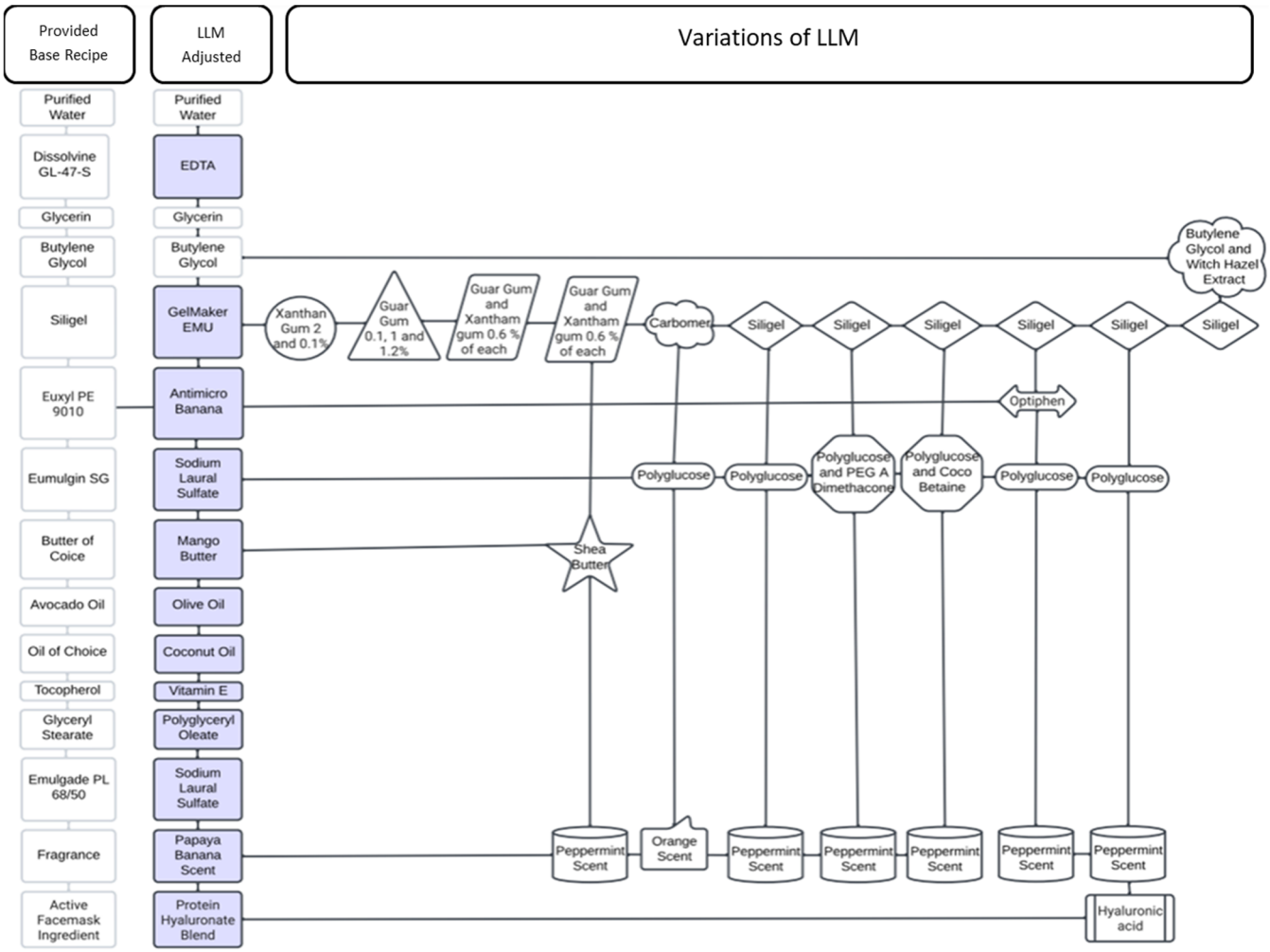


Figure 1: Flowchart illustrating ingredient modifications across multiple formulation iterations, guided by LLM suggestions. The "Provided Base Recipe" represents the original formulation supplied by a professional formulator. The "LLM Adjusted" column shows the initial modified formula created by ChatGPT using only available ingredients. Each subsequent column reflects a single iteration, highlighting only the ingredient changes from the previous version. These modifications aimed to improve specific product characteristics such as viscosity, stability, and cost-efficiency.

The improvements seen in later iterations may be partially attributed to the growing use of the structured prompt format outlined in **Table 1**, which helped reduce variability in LLM output despite its memory limitations. This led to the consensus that the LLM excelled at generating ideas for quick adjustments based on the immediate prompt but lacked the ability to contribute effectively to the iterative adjustments typically made during the formulation process, where precise, data-driven refinements are required. As such, it was concluded that the value the LLM brings to the process is enhanced by the level of formulation training of the user overseeing its responses. This also began to be apparent that as the level of training of the user increased the level of use of the LLM in the initiation of a product decreased. Future research should prioritize the integration of prompt engineering strategies to enhance the effectiveness of LLMs in formulation development in the hands of novice formulators to further test the use of an LLM for formulator training.

While this study primarily relied on qualitative and semi-quantitative assessments typical of educational formulation work, future research could benefit from incorporating more robust, standardized quantitative metrics. Many of these, such as rheological profiling over time, spectrophotometric clarity measurements, and microbial challenge testing with standardized colony counts, are already widely used by professional formulators to assess whether a product meets desired performance and stability criteria. In industry, these technical evaluations are often paired with structured customer feedback to refine products based on sensory attributes, user experience, and consumer satisfaction. Although implementing such detailed analyses and consumer testing may be impractical in most classroom settings with limited time and resources, introducing students to the principles behind these processes—even through informal peer reviews or simplified surveys—can help bridge the gap between educational experiences and industry expectations. Incorporating these more quantitative elements into extended student projects or choosing one or two in advanced formulation courses could enhance scientific rigor, deepen student engagement, and strengthen the applicability of LLM-guided formulation training.

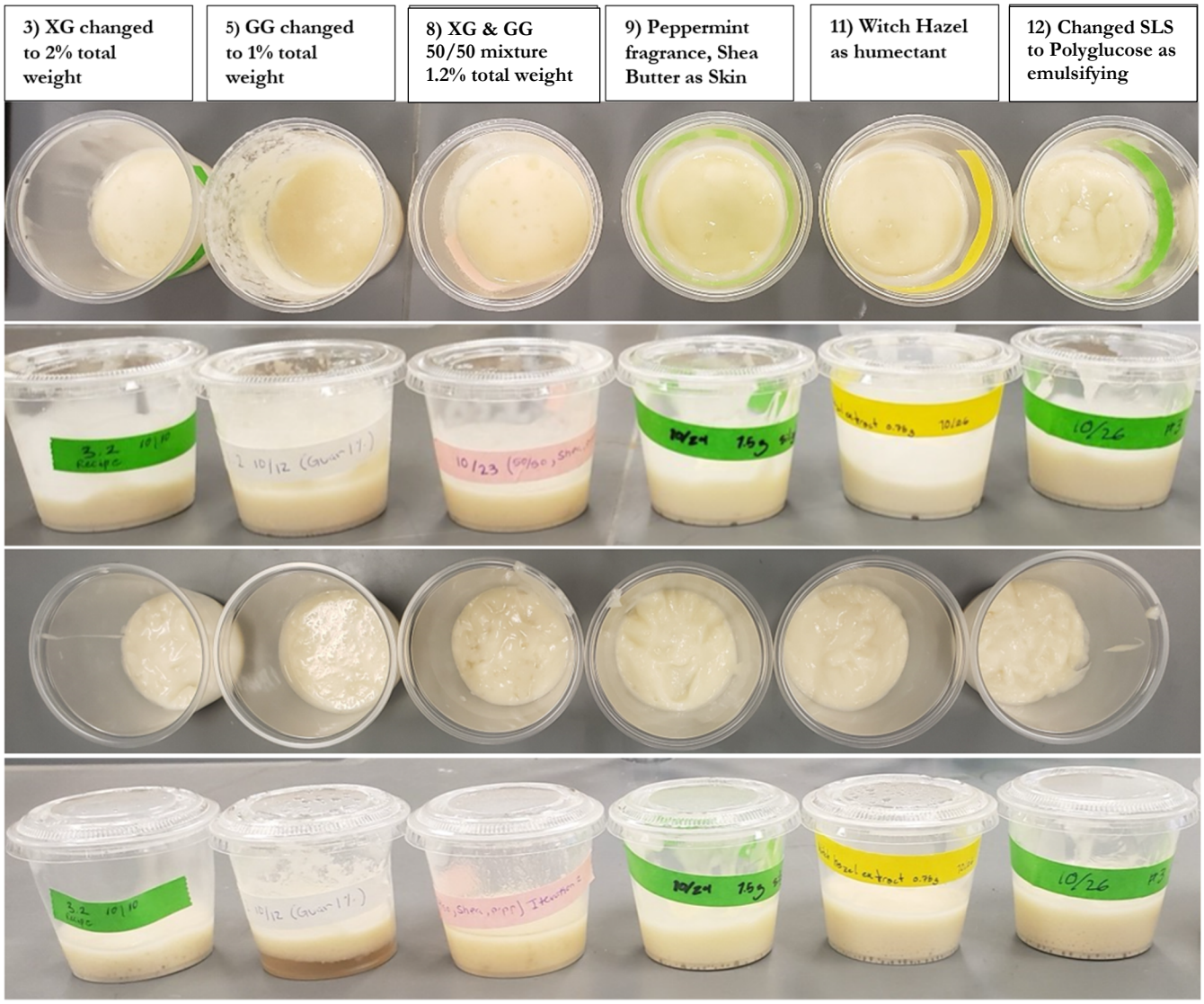


Figure 2. Photographic comparison of six sample formulations (iterations 3, 5, 8, 10, 11, and 12) before (top two rows) and after (bottom two rows) undergoing a one-week stability challenge at 50ºC. Each sample was visually assessed for phase separation, with sample five showing notable instability post-challenge. This separation indicates breakdown in the emulsion or thickener performance. Abbreviations: XG = xanthan gum, GG = guar gum, SLS = sodium lauryl sulfate. These visual results align with semi-quantitative stability data in **Table 2**.

In observations of the LLM, it was also noted that there existed potential issues with how these models handle longer prompts. While it is well-documented that LLMs have a maximum limit for processing input, the findings from this research suggest that ChatGPT may not always recognize when a prompt exceeds its limit, resulting in an unintended focus on the latter sections of the input. This could imply that the model's limit might function in a manner that prioritizes the end of the prompt, effectively

applying the limit in reverse order from the conclusion of the input back toward the beginning. For instance, using the prompts from **Table 3**, when asked "What is a good replacement for xanthan gum in a face mask lotion?", ChatGPT suggested guar gum as a replacement. However, it did not specify how this substitution would be incorporated into the recipe, such as providing the required mass or percentage to maintain the desired consistency.

To further illustrate the potential impact of prompt length, ChatGPT's responses were compared between a shorter prompt and a longer prompt (see **Table 3**). The shorter prompt focused solely on identifying a replacement for xanthan gum. In contrast, the longer prompt included additional information about the other ingredients and their proportions in the recipe, as well as the desired total mass of the final product. It is important to note that this observation is based on empirical experiences and has not been previously documented in existing literature. Given the lack of comprehensive studies on this phenomenon, it would be beneficial for future research to investigate this behavior more thoroughly. If no additional research exists and these observations remain unique, providing further evidence or specific examples is recommended to substantiate this claim within the broader context of LLM limitations.

The time students took to complete each formulation iteration served as a key metric for assessing their learning and the effectiveness of LLM assistance. As students progressed through multiple iterations, a trend emerged where earlier formulations took significantly longer to complete, while later formulations were completed more efficiently. This decrease in time suggests that students became more adept at interpreting LLM-generated suggestions, refining their prompts, and identifying which formulation spaces were worth further exploration. Additionally, the time taken for each iteration reflects the iterative nature of formulation training, where trial and error are an essential part of the learning process. The students who quickly identified viable formulation paths were able to iterate more frequently, thereby improving their understanding of ingredient functionality and substitution strategies.

While the quality of the final product is inherently subjective, **Table 2** provides a structured, semi-quantitative framework to assess formulation performance relative to time and iteration count. The number of iterations a student required before achieving a product that met the desired specifications serves as an indicator of their increasing proficiency. These outcomes can be tracked across iterations in **Table 2**, which reveals how both qualitative and quantitative measures evolved alongside student skill development. Furthermore, even unsuccessful formulations contributed valuable insights, as they helped students identify which formulation pathways were unproductive and should not be further explored. This aspect of the study highlights how LLM-driven formulation training is not solely about achieving an ideal final product but also about developing strategic decision-making skills in the formulation process. By capturing the relationship between iteration count, time efficiency, and formulation refinement, this study offers a repeatable method for assessing student learning in any future application of LLM-assisted formulation training.

While the time required to complete each formulation iteration demonstrated a clear trend of improvement, it is important to acknowledge that several factors influenced the time required for each round of formulation. Variables such as familiarity with laboratory procedures, confidence in interpreting LLM-generated responses, and complexity of the formulation changes requested all played a role in determining how quickly students were able to complete their assigned tasks. Additionally, external factors, including group dynamics, laboratory setup, and availability of materials, could introduce variability in the time required for each round. The recorded times, therefore, should not be interpreted as rigid, highly controlled measurements but rather as general indicators of increasing efficiency in novice formulators as they became more proficient with LLM-assisted formulation.

Despite these inherent variations, a clear progression in efficiency was observed across the four rounds of formulation experiments. In the first round, where the entire class collaborated on a single formula, the process took over 120 minutes as students navigated the fundamental aspects of formulation and LLM-assisted ingredient substitution for the first time. In the second round, smaller groups of two to three students worked independently to replicate the same formulation, significantly reducing the required time to approximately 90 minutes per group. By the third effort, multiple groups worked on different formulations simultaneously, with no group requiring more than an hour to complete their assigned recipe. By the fourth round, students had become comfortable enough with the iterative process that they were able to complete two full formulations within an hour, meaning that the average time for a single formulation had dropped to approximately 30 minutes. However, even with increasing familiarity, the time required to complete a formulation never decreased significantly below the 25-minute range. This suggests that while experience and structured LLM guidance improved formulation efficiency, there remains an inherent time investment required to complete each product iteration due to the physical constraints of weighing, mixing, and evaluating ingredients.

While LLM provided quick formulation suggestions, it often produced inaccurate calculations, making its ingredient substitutions unreliable.[10] LLM particularly fell short in making decisions about necessary amounts for synthesizing cosmetic products. This

limitation is illustrated in **Table 4**, which presents a prompt-response exchange from the development of iteration 16. The LLM suggests two antimicrobial substitutions—Phenoxyethanol and Optiphen—but offers no specific dosing guidance, instead deferring to manufacturer recommendations. This reflects a key challenge in relying on LLMs for detailed formulation decisions, even when prompts are well-structured. When it provided estimations and suggestions, its calculations were frequently incorrect. LLM was not reliable when asked to determine the percent composition or the appropriate mass to add to reach a desired mass percent. This limitation impacted the ingredient replacement's ability to improve specifications. For example, when LLM suggested replacing xanthan gum with guar gum at a 1:1 ratio to produce iteration four, the resulting product decreased in texture quality compared to iteration three. Iteration four's product was unable to produce a stable viscosity and was heterogeneous. Multiple rounds were necessary to ascertain if the LLM-provided substitutions were an acceptable substitution, or if its concentration was incorrect to get the desired change. Therefore, future use of LLMs in formulations necessitates formulators to double-check and validate LLM calculations to ensure the accuracy of the final product.

Sample Number Brief Description	Odor (Ideal++)	Viscosity (Ideal +++)	Appearance (Ideal +)	Skin hydration (Ideal ++)	After effect (Ideal+)	Viscosity in Pa*s (% change)	% change price per gram (Base cost $0.034/g)
1) Initial LLM Formula	+	--	+	ND	ND	ND	0%
2) XG as emulsifier/ viscosity modifier 10% total weight	+++	+++++	++	++	+	ND	-20%
3) XG changed to 2% total weight	+++	++++	++	++	+	30.0 (3.2)	-50%
4) GG as emulsifier/viscosity modifier 0.1% total weight	+	-	+	+	-	ND	-51%
5) GG changed to 1% total weight	+	++	+	+	-	ND	-51%
6) GG changed to 1.2% total weight	+++	++++	+	+	++	ND	-50%
7) Silgel as emulsifier/ viscosity modifier 2% total weight	+	++	+	++	+	67.4 (76)	-12%
8) XG & GG 50/50 mixture 1.2% total weight	+	++++	+	+	+	35.9 (19)	-49%
9) Peppermint fragrance, Shea Butter as Skin Conditioning Agent	+++	++	+	+++	-	22.6 (17.9)	-54%
10) Silgel as emulsifier/viscosity modifier 3% total weight	+++	+++	+	-	++	ND	9%
11) Witch Hazel as humectant	+	++++	+	-	-	39.5 (4.5)	27%
12) Changed SLS to Polyglucose as emulsifying agent	+++	+++	+	-	-	62.7 (52)	7%
13) Peg A Dimethicone w/ Polyglucose	-	++	+	+	-	57.0 (15)	-1%
14) Coco Betain w/ Polyglucose	ND	ND	ND	ND	ND	66.0 (35)	41%
15) Replace Silgel w/ Carbomer	+	+++++	-	ND	ND	30.1 (37)	-57%
16) Optiphen as preservative	+	+++	+	ND	ND	23.0 (36)	-1%
17) Hyaluronic Acid as active	++	+++	+	++	+	40.5 (7.2)	10%
18) Aloe added as a humectant	ND	--	ND	ND	ND	ND	41%

Table 2. Qualitative and quantitative assessment of the initial formula and 17 LLM-generated face mask formulations compared to the initial adjusted base recipe. Specifications include odor, appearance, skin hydration, after-effect, and viscosity, rated against ideal values (shown in parentheses). Numerical viscosity was measured at baseline and after two minutes; percent change reflects non-Newtonian behavior. "ND" (Not Determined) indicates a formulation failure that prevented testing. Price per gram is reported as a percentage change from the base cost ($0.034/g). While pH was measured for each sample using indicator paper, it remained within an acceptable and consistent range across all formulations and is therefore not shown here. This table integrates sensory, physical, and cost-based metrics to evaluate LLM-driven formulation success. The visual differences for a subset of these samples are shown in **Figure 2**. Iteration pathways corresponding to these data are mapped in **Figure 1**. Trends across the table also reflect growing student proficiency in identifying effective substitutions and understanding the functional role of ingredients.

Adding an additional task of maintaining specifications while lowering a quantitative value (the cost) seemed to be too much for LLM in the hands of new formulators. LLM struggled to gather information on pricing while maintaining product quality, likely on account of the many variables that the LLM needed to account for in this task. Such inability is not only evident in performance metrics such as cost but also visibly apparent in **Figure 2**, where sample five demonstrates a clear phase separation after heat challenge testing on a sample with a suggested change. General information regarding the prices of materials provided by the LLM was generally reliable, but the specifics were not robust enough to provide reliable information compared to internet searches. Thus, LLM suggestions needed to be double-checked with internet and vendor catalog searches to verify pricing. The calculated cost of the iterations did fluctuate, but this fluctuation was close to the magnitude of different vendors' price fluctuation of the materials used, thus nullifying any real gains from LLM proposals (see **Table 2**).

LLM does have access to chemistry knowledge[7] and can piece much of the data together to provide simple answers and enough feedback to easily trick a novice into thinking it does have the ability to project more complex answers to questions with more variables. However, as mentioned above, the answers to complex problems provided were usually incorrect. Unfortunately, sometimes the inaccuracies in the answers require expertise and knowledge to identify. While it is challenging to quantify the depth of knowledge and abilities of LLM, providing a little context in the provided prompt, such as explaining the project at hand, enables LLM to make more connections with its database and thus offer more suggestions on how to proceed.

	Prompt Text	LLM Response Text
General Input Prompt	"What is a good replacement for xanthan gum in a face mask lotion?"	"Xanthan gum is often used in cosmetic and skincare formulations to provide a thickening and stabilizing effect. If you're looking for a substitute for xanthan gum that will give your face mask a creamy consistency, you can consider using Guar Gum as a replacement. Guar gum is a natural thickener and stabilizer commonly used in cosmetic products and can be substituted for xanthan gum."
Detailed Input Prompt	"From this face mask recipe, suggest a substitute for the Xanthan gum in the correct scaled mass amount that will give a creamy consistency. Purified water: 70.65% EDTA: 0.15% Glycerin: 6.00% Butylene Glycol: 3.00% Xanthan gum: 0.6% Guar Gum: 0.6% Antimicrobial Banana mixture: 0.90% Sodium Lauryl Sulfate: 1.00% Mango Butter: 3.00% Olive Oil: 5.00% Coconut Oil: 6.00% Vitamin E: 0.50% Polyglycerol Oleate: 2.00% Papaya Banana: 0.10% Protein-Hyaluronate blend: 0.50% The desired yield of this recipe is 50 grams."	"To determine the correct scaled mass amount of Guar Gum as a substitute for the 5 grams of Xanthan Gum, you can use a similar proportion. Here's the calculation: Original Xanthan Gum: 5 g Guar Gum Substitute: (5 g x Mass of Guar Gum) / Mass of Xanthan Gum To calculate the mass of Guar Gum, you can rearrange the equation: Mass of Guar Gum = (5 g x Mass of Guar Gum) / 5 g Mass of Guar Gum = Mass of Guar Gum So, you can substitute 5 grams of Guar Gum for the 5 grams of Xanthan Gum in your recipe to achieve a creamy consistency."

Table 3. Comparison of LLM responses to a general vs. detailed prompt for substituting xanthan gum in a face mask formulation. The general prompt lacks context and receives a vague, low-utility response. In contrast, the detailed prompt includes a full ingredient list and desired yield, prompting a more targeted suggestion (albeit with flawed math). This example illustrates the critical role of prompt structure in obtaining relevant LLM output. For a real-world example of how these types of prompts played out in iterative formulation, see **Table 4**.

The comparison in **Table 3** demonstrates a clear difference in the quality of LLM responses based on prompt structure. While the general prompt results in a vague suggestion lacking practical application, the detailed prompt elicits a more specific—though still mathematically flawed—response. This reinforces the importance of carefully crafted inputs when using LLMs for formulation work. However, not all LLM interactions were this cleanly contrasted. In actual formulation sessions, prompts were often part of a longer back-and-forth as students refined their questions and evaluated ChatGPT's responses in real time. These real interactions offer a more nuanced view of how LLM limitations manifest in practice, particularly when students ask for ingredient substitutions or specification adjustments. **Table 4** presents an excerpt from one such exchange during the development of iteration 16. It captures a typical scenario in which the LLM suggests valid alternative ingredients but fails to provide the specific dosing information needed for practical implementation. This example illustrates how even well-structured prompts can yield incomplete answers, requiring students to apply critical thinking and follow-up strategies to move forward.

CONCLUSION

This LLM proved highly effective in providing trainee formulators with quick ingredient substitution recommendations, particularly when selecting from an available ingredient library. It was also partially helpful in enhancing product characteristics to meet specifications. Based on these results, LLMs could be a valuable tool for teaching novice formulators how to adjust ingredients in a cosmetic formulation, especially when a detailed prompt template is used. However, determining the appropriate ratio for ingredient substitutions and optimizing costs would still require multiple iterations or the supervision of a trained formulator. It could even be suggested that an expert formulator may find LLM suggestions somewhat limited in offering truly novel insights.

Prompt Text	LLM Response Text
The following text is the recipe for a gel-based cosmetic face mask: • Purified water (base ingredient): 70.65% • EDTA (chelating agent): 0.15% • Glycerin (humectant): 6.00% • Butylene Glycol (skin conditioner): 3.00% • Siligel (thickening agent): 2.00% • AntiMicrobial Banana mixture (antimicrobial): 0.90% • Polyglucose (detergent/surfactant): 1.0% • Mango Butter (oil base): 3.00% • Olive Oil (oil base): 5.00% • Coconut Oil (oil base): 6.00% • Vitamin E (antioxidant): 0.50% • Polyglyceryl Oleate (emulsifier): 2.00% • Papaya Banana (fragrance): 0.10% • Protein-Hyaluronate blend (humectant): 0.50% The desired yield of this recipe is 50 grams. Suggest alterations to this recipe based on the following criteria: Change out the antiMicrobial banana with an alternative antiMicrobial ingredient that will not yellow following exposure to heat	To replace AntiMicrobial Banana with an alternative antimicrobial ingredient that won't yellow after exposure to heat, you can consider using a preservative like Phenoxyethanol. Phenoxyethanol is a widely used preservative that is effective against bacteria and fungi and is less likely to cause yellowing in formulations. Here's the altered recipe with Phenoxyethanol as the alternative antimicrobial ingredient: • Purified water (base ingredient): 70.65% • EDTA (chelating agent): 0.15% • Glycerin (humectant): 6.00% • Butylene Glycol (skin conditioner): 3.00% • Siligel (thickening agent): 2.00% • Phenoxyethanol (antimicrobial): 1.00% (You can adjust the amount based on the manufacturer's recommendations for effective preservation.) • Polyglucose (detergent/surfactant): 1.00% • Mango Butter (oil base): 3.00% • Olive Oil (oil base): 5.00% • Coconut Oil (oil base): 6.00% • Vitamin E (antioxidant): 0.50% • Polyglyceryl Oleate (emulsifier): 2.00% • Papaya Banana (fragrance): 0.10% • Protein-Hyaluronate blend (humectant): 0.50% With this alteration, Phenoxyethanol will provide effective antimicrobial protection without causing yellowing in your formulation. Adjust the Phenoxyethanol amount based on your specific product's needs and the manufacturer's recommendations to ensure proper preservation.
Suggest another alternative according to prompt above	If you prefer an alternative antimicrobial ingredient to replace AntiMicrobial Banana that won't yellow after exposure to heat, you can consider using Optiphen or Optiphen ND. These are effective preservatives suitable for cosmetic formulations that offer broad-spectrum protection against bacteria and fungi. Here's the altered recipe with Optiphen as the alternative antimicrobial ingredient: • Purified water (base ingredient): 70.65% • EDTA (chelating agent): 0.15% • Glycerin (humectant): 6.00% • Butylene Glycol (skin conditioner): 3.00% • Siligel (thickening agent): 2.00% • Optiphen or Optiphen ND (antimicrobial): 1.00% (Adjust the amount based on the manufacturer's recommendations for effective preservation.) • Polyglucose (detergent/surfactant): 1.00% • Mango Butter (oil base): 3.00% • Olive Oil (oil base): 5.00% • Coconut Oil (oil base): 6.00% • Vitamin E (antioxidant): 0.50% • Polyglyceryl Oleate (emulsifier): 2.00% • Papaya Banana (fragrance): 0.10% • Protein-Hyaluronate blend (humectant): 0.50% Optiphen and Optiphen ND are considered alternatives that should provide effective antimicrobial protection without causing yellowing in your formulation. Adjust the amount based on your specific product's needs and the manufacturer's recommendations to ensure proper preservation.

Table 4. Prompt-response exchange with the LLM during formulation of iteration 16. The initial prompt requests a non-yellowing alternative to Antimicrobial Banana, yielding Phenoxyethanol; a follow-up prompt results in Optiphen. While ingredient suggestions were relevant, the LLM offered no dosing guidance and deferred to manufacturer recommendations. This illustrates key limitations in LLM output even with well-structured prompts, emphasizing the need for user oversight. See **Table 3** for how prompt detail influences response quality.

Future studies could repeat these experiments with larger consumer groups, integrating consumer feedback into iteration loops and using a rubric to convert qualitative specifications into quantitative data. LLMs could also be explored as a tool for quantifying consumer feedback, despite potential challenges in interpreting and categorizing subjective input. In such cases, LLMs would effectively generate the "usable" data. Although potentially highly beneficial, this approach could also introduce significant risks, given the history of LLMs struggling with quantifiable parameters such as mathematical calculations and cost assessments.

A possible solution is to enhance LLMs with supplementary code tailored to specific recipes or laboratory settings, allowing LLM responses to be transformed into structured, usable data. Developing and integrating such external programs could improve the efficiency of LLMs in formulation workflows and expand their applicability to different areas of cosmetic development.
Given the typical lack of formal training among novice formulators, integrating LLMs into training programs could significantly accelerate the learning curve by exposing students to a wide variety of ingredient substitutions. Instructors looking to implement these findings can incorporate structured LLM-assisted formulation exercises into laboratory coursework. By providing students with a standardized prompt template, educators can lower the initial barrier to using LLMs and help students generate meaningful formulation modifications more quickly. These structured prompts guide students through iterative changes while encouraging them to think critically about ingredient functionality and formulation outcomes. Over time, students begin to associate prompt design with experimental planning—a valuable link between digital literacy and hands-on lab skills. This approach offers a practical, low-cost way for instructors to enhance formulation training while introducing students to tools increasingly used in the industry.

REFERENCES

1. Smith, J., Zhao, L., Patel, R., & Thompson, K. (2025) Survey on recent progress of AI for chemistry: Methods, applications, and opportunities. *J. Comp. Chem. AI*, 42(2), 123–145. *https://doi.org/10.48550/arXiv.2502.17456*
2. Bannigan, P., Aldeghi, M., Bao, Z., Häse, F., Aspuru-Guzik, A., & Allen, C. (2021) Machine learning-directed drug formulation development. *Adv. Drug Deliv. Rev.*, 175, 1–14. *https://doi.org/10.1016/j.addr.2021.05.016*
3. Webb, T., Holyoak, K., & Lu, H. *(2023)* Emergent analogical reasoning in large language models. *Nat. Hum. Behav.*, 7, 1526–1541. *https://doi.org/10.1038/s41562-023-01659-w*
4. Artrith, N., Butler, K., Coudert, F., Han, S., Isayev, O., Jain, A., & Walsh, A. (*2021)* Best practices in machine learning for chemistry. *Nat. Chem.,* 13, 505–508. *https://doi.org/10.1038/s41557-021-00716-z*
5. Murray, J., Lange, J., Bennett-Lenane, H., Holm, R., Kuentz, M., O'Dwyer, P., & Griffin, B. *(2023)* Advancing algorithmic drug product development: Recommendations for machine learning approaches in drug formulation. *Eur. J. Pharm. Sci.,* 191, 1–13. *https://doi.org/10.1016/j.ejps.2023.106562*
6. Xiao, P., Zhang, X., Pan, W., Ou, X., Bontozoglou, C., Chirikhina, E., & Chen, D. (2020) The development of a skin image analysis tool by using machine learning algorithms. *Cosmetics*, 7, 1–12. *https://doi.org/10.3390/cosmetics7030067*
7. Shi, Y., Yang, Z., Ma, S., Kang, P., Shang, C., Hu, P., & Liu, Z. (2023) Machine learning for chemistry: Basics and applications. *Engineering*, 27, 70–83. *https://doi.org/10.1016/j.eng.2023.04.013*
8. Paweł, K., Mazurek, K., & Kurasiński, A. (2023) Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrep. Bus. Econ. Rev.*, 11, 25–37. *https://doi.org/10.15678/eber.2023.110302*
9. Aljanabi, M., Yaseen, M., Ali, A., & Mohammed, M. (2023) Prompt engineering: Guiding the way to effective large language models. *Iraqi J. Comput. Sci. Math.*, 4, 151–155. *https://doi.org/10.52866/ijcsm.2023.04.04.012*
10. Tyson, J. (2023) Shortcomings of ChatGPT. *J. Chem. Educ.*, 100, 3098–3101. *https://doi.org/10.1021/acs.jchemed.3c00361*
11. Lu, H., Li, Y., Chen, M., Kim, H., & Serikawa, S. (2023) ChatGPT: A comprehensive review on background, applications, key challenges, and future directions. *J. Innov. Knowl.*, 8, 1–14. *https://doi.org/10.1016/j.iotcps.2023.04.003*
12. Vallurupalli, M., Shah, N. D., & Vyas, R. M. (2024) Validation of ChatGPT 3.5 as a tool to optimize readability of patient-facing craniofacial education materials. *Plast. Reconstr. Surg. Glob. Open*, 12(2), e5575. *https://doi.org/10.1097/GOX.0000000000005575*
13. Cong-Lem, N., Soyoof, A., Tsering, D. (2024) A systematic review of the limitations and associated opportunities of ChatGPT. *Int. J. Hum.-Comput. Interact.*, 40(2), 3851-3866. *https://doi.org/10.1080/10447318.2024.2344142*
14. Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023) ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv Preprint, arXiv:2304.05613. *https://doi.org/10.48550/arXiv.2304.05613*
15. Nascimento, C., & Pimentel, S. (2023) Do large language models understand chemistry? A conversation with ChatGPT. *J. Chem. Inf. Model.*, 63, 1649–1655. *https://doi.org/10.1021/acs.jcim.3c00285*
16. Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., & Resnik, P. (2024) The prompt report: A systematic survey of prompting techniques. arXiv Preprint, arXiv:2406.06608. *https://doi.org/10.48550/arXiv.2406.06608*

ABOUT THE STUDENT AUTHORS

McKinnley Bilbao is a dedicated undergraduate student at Utah Valley University, currently pursuing a Bachelor's degree in Psychology with a minor in Chemistry. Known for her exceptional academic achievements, McKinnley thrives as an Honors student, maintaining an impressive GPA. Her passion for research has led to contribute significantly to a project titled *Utilizing a Large Language Model for Training Students in Personal Care Product Formulation* poised for publication. With an ambition to pursue medical studies, McKinnley looks forward to leveraging her diverse academic background to make meaningful contributions in the field.

Caitlin West is a student at Utah Valley University studying in the Chemistry Program. She has a strong background in Culinary Arts and loves to find ways to utilize her baking skills in Chemistry. Following graduation, she hopes to continue her educational pursuits by attending graduate school to deepen her understanding of Chemistry and contribute to the scientific community.

Tomas Carmona, Morgan Covarrubia, Alex Goslin, and Garland Munn are UVU students in various years, busy trying to wrap up their education to continue to enhance their lives and their communities.

Katherine Judge is a biochemistry major with a curiosity for the intricate workings of chemistry. Katherine eagerly engages in research projects and laboratory experiments, fueled by a passion for understanding the fundamentals, and a desire to make meaningful contributions to the field of medicine. Committed to academic excellence, Katherine is driven by the pursuit of knowledge and discovery. She is eager to contribute to advancements in chemistry and inspire others through appreciation of the wonders of the scientific world.

Hazel Ticas is a researcher from El Salvador who graduated with a bachelor's degree in French from Brigham Young University.

Abe Tonioli, pursuing a biology degree from Utah Valley University, distinguishes himself through leadership, scientific research, and extensive community service. An Eagle Scout fluent in Tagalog, he excels in problem-solving and collaboration. He is passionate about rock climbing and DJing, blending analytical rigor with creative expression. Abe's commitment to science and personal growth, coupled with his unique blend of skills, positions him as a valuable asset to any team.

Collin Tuttle is a second-year professional chemistry undergraduate at Utah Valley University. Although originally from the northern California bay area, he moved to Utah to pursue higher education along with his older brother. While primarily interested in chemistry, he maintains an interest in a wide variety of scientific and academic topics, such as physics, astronomy, and computer science.

PRESS RELEASE

This research delves into the promising application of machine learning, specifically the ChatGPT 3.5 system, in training future product formulators. Through an experimental approach, the study examines the system's capability to aid students in developing a hydrating face mask recipe. Results indicate that while the model exhibits strong chemistry knowledge and offers useful suggestions for ingredient substitutions, it faces challenges with memory retention and mathematical computations. Nevertheless, it emerges as a valuable resource for guiding students in refining their formulations and achieving high-quality outcomes.

Print Edition ISSN 1536-4585
Online Edition ISSN 2375-8732