BACSA

The Biotechnology And Computer Science Association

BACSA RESEARCH MAGAZINE 2026

UNIVERSITY OF TORONTO

FROM THE PRESIDENTS

We are proud to present the 2026 BACSA Research Magazine, the result of the incredible talent, hard work, and creativity of the Research Project 2025-2026 participants and the entire BACSA team.

This year's collection showcases the power of interdisciplinary collaboration, where biotechnology meets computer science to solve real-world problems.

Enjoy!

Astrid Chavez Chavira
Co-President

Nathan Padua
Co-President

BACSA is not responsible for any academic misconduct

TABLE OF CONTENTS

RESEARCH COMPETITION AWARDS

FIRST PLACE

Team : VibraPatrol

Ioli Ntasi Vieira
Adaeze Egwuatuonwu
Winnie Shan Ying Gu

PAPER:
CNN-Based Vibroacoustic Detection of Triatomine Vectors for Early Chagas Disease Prevention (VibraPatrol)

SECOND PLACE

Team: MUTacj

Amir Kanbar
Calvin Joy Thomas
Jerin Varghese John

PAPER:
Cancer Mutation Pattern Recognition (CAMPR): An Interactive Simulation Platform for Teaching Cancer Genomics

CNN-Based Vibroacoustic Detection of Triatomine Vectors for Early Chagas Disease Prevention (VibraPatrol)

Ioli Ntasi Vieira, Adaeze Egwuatuonwu, Winnie Shan Ying Gu

Abstract:

Chagas disease is a parasitic disease resulting in over 10,000 deaths annually and progressing asymptomatically before severe cardiac complications, which remains a significant public health concern in Latin America. The disease, transmitted by the feces of triatomine ("kissing") bugs, is currently poorly addressed by public health agencies, resulting in late-stage diagnosis and reliance on manual vector detection, which in turn limits early intervention. That said, VibraPatrol proposes a machine-learning-based acoustic detection system designed to identify the specific vibroacoustic signals emitted by triatomine ("kissing") bugs upon disturbance. This system uses microphone-embedded sensors in public spaces, such as lamp posts, and trains a Convolutional Neural Network (CNN) on at least 50 validated triatomine acoustic patterns.

Audio signals are preprocessed using bandpass filtering and transformed into log-mel spectrogram representations for classification. The model performance would be evaluated based on accuracy, precision, recall, F1 score, and false positive rate under simulated low and high-noise environments. Based on comparable insect acoustic classification studies, VibraPatrol would allow detection of vector presence under varying noise and distance conditions, with a target accuracy of ~80-90% and a false-positive rate of ~15%.

Upon detection, automated risk alerts (high, medium, low) would be sent to nearby mobile devices, informing and encouraging early testing and treatment, while supporting the development of geospatial risk maps for disease regulation within the public health sector. Overall, VibraPatrol offers a scalable and data-driven approach to reduce Chagas disease risk in Latin American communities through earlier detection.

Introduction:

Chagas disease, or American trypanosomiasis, is a life-threatening parasitic infection caused by the protozoan parasite *Trypanosoma cruzi* and transmitted primarily by the feces and bites of infected triatomine insects known as "kissing bugs", affecting more than 7 million people around the world and resulting in more than 10,000 deaths every year (World, 2022). The asymptomatic nature of this disease results in the development of severe cardiac, digestive, and neurological medical conditions associated with chronic Chagas disease, resulting in high mortality rates, especially in Latin America (Carmo et al., 2024). While there is yet to be a vaccine, and testing methods require innovative solutions to boost productivity and improve diagnostics, research funding for this disease remains very low (Heukelbach et al., 2021). Moreover, few studies have applied Convolutional Neural Networks (CNNs) to detect triatomine bugs in urban settings. Therefore, earlier detection of Chagas disease is significant to reducing disease transmission and enabling timely testing and treatment, and the reduction of total annual healthcare burden and associated economic losses (Lee et al., 2013). Considering that early detection requires identifying triatomine bugs before they transmit infection, more effective detection methods leveraging technology to identify specific biological traits of the vector are required to reduce incidence rates of Chagas disease.

Triatomine bugs are known for the biological vibroacoustic signals they emit as stridulation frequencies when communicating with members of the species (Quiroga et al., 2019). Similar to existing acoustic technologies in agriculture, VibraPatrol aims to utilize this feature of *T. cruzi* by analyzing these signals through CNN software that draws comparisons between detected vibracoustic signals in the environment, with existing insect frequency databases to identify the vector. Subsequently, this alerts Latin American Populations of their potential exposure to Chagas disease, thereby promoting early diagnosis and disease prevention (Mankin et al., 2021). The CNN training model was chosen due to its effectiveness in the

analysis of bioacoustic data, and its specialty in detecting short recurring rhythms reflected through the analyzed spectrograms of the audio. The spectrograms will convert time-frequency sound data into images, allowing deep learning of the CNN, which creates models of high accuracy and attention to the recurring frequencies of the Triatomine bugs (Balingbing et al., 2024). Additionally, the CNN-based acoustic model shows high classification accuracy in varying noise environments, including conditions characterized by elevated noise levels (Stowell, 2022).

This study aims to develop and evaluate a CNN-based acoustic detection system trained to identify triatomine vibroacoustic signals with ≥80% accuracy under varying environmental conditions. We hypothesize that this detection system is capable of simultaneously alerting nearby mobile devices to triatomine presence, while providing information on Chagas disease test centers and treatment, which allows earlier treatment and reduced transmission risk of Chagas disease.

Method/Approach:

Data Acquisition

The method for the development of VibraPatrol commences with the collection of triatomine ("kissing") bug audio and/or vibration recordings from public entomology and bioacoustic repositories on published supplementary datasets. To include a recording within our data set, the recording should comprise a verified triatomine label with a sufficient duration (>5 seconds) for segmentation. At the end of this section of the method, a dataset should be produced with ~50 different triatomine recordings.

Preprocessing and signal standardization

In order to properly analyze all of the different recordings gathered in the "Data Acquisition" section of this method, all of the recordings should be standardized to a consistent sampling rate of 22,050 Hz and for normalized amplitude consistency. Background noise reduction was applied using a fourth-order Butterworth bandpass filter to isolate relevant frequency ranges, using the SciPy library. Audio clips were then trimmed into fixed-length

windows of 2-3 seconds to ensure input dimensions. All preprocessing was performed using Python (v3.11).

Feature extraction

Two feature representations were generated:

1. Log-mel spectrograms, created using the librosa library, capture time-frequency structure.
2. Mel-frequency Cepstral Coefficients (MFCCs) to quantify spectral texture and tone characteristics

A distinct buzzing pattern can then be cross-identified in the different audios. Spectrograms were then resized into standardized images for CNN input.

Model development:

The data would then be divided into 70% training (the data the model learns from), 15% validation (used during development to improve the model), and 15% unseen-test used to evaluate performance levels. subsets using stratified sampling to preserve class balance. These different sections of data are then fed into a CNN (Convolutional Neural Network), which specializes in detecting textures and recognizing frequency patterns which resemble insect acoustics. For model development and training, TensorFlow (v2.x) was used, combined with Keras API.

The structure of the CNN model consisted of the following:

1. Convolutional layer: detects simple patterns
2. ReLU activation layer: (converts negative values to zero and keeps positive values; allows more stable, faster, and better detection of complex patterns)
3. Max-pooling layers: reduce the images and analyze a small section
4. Dropout (0.3) for Overfitting Prevention: turns off 30% of machine-learning neurons to prevent performing poorly on new sounds
5. Fully Connected Dense Layer: combines all detected features and weighs their importance
6. Sigmoid Output for Binary Classification: 2 possible outcomes (triatomine present or absent); allows triggering of alerts based on probability

Performance Evaluation:

To match real-world conditions, the trained CNN product would be tested in different conditions. These would include high, medium, and low background noises from urban, public, and rural areas. Additionally, we would test the trained product under different distances (5m, 10m, 15m) of the triatomine bug frequencies. The results would be analyzed on the basis of:

1. Accuracy
2. Precision
3. Recall (sensitivity)
4. F1-score (measure of model accuracy while balancing missed detection and false alarms)
5. False positive rate

For performance evaluation metrics, scikit-learn was used.

Alert Prototype:

Simulate the alert output of the system under the same conditions in the "testing" section of the method. Different risk notification tags (low, medium, high) would be designed and based on the model probability outputs (>0.90 high, 0.70-0.90 medium, <0.70 low).

Market Analysis:

VibraPatrol's audience is directed towards municipal governments and public agencies in Chagas-endemic regions of Latin America. Secondary markets would include smart-city infrastructure initiatives and non-governmental organizations involved with Chagas disease. Current prevention methods in place within these areas are highly reliant on manual home inspections, reactive sprays, and passive reporting systems, which not only require significant amounts of time and money but are also commonly incapable of detecting the parasite before exposure has occurred. That said, current methods do not minimize the cost involved in medically treating Chagas disease, which in Brazil alone, has been approximated to $11.44 billion US dollars annually (Andrade et al., 2025). In this way, by providing continuous, auditory monitoring through a software-based system, VibraPatrol offers an innovative alternative to early detection treatment, which solely costs around $200, compared to the $45,034 lifetime cost per patient (Andrade et al., 2025). The system would run on a subscription model, where cities pay a

yearly fee for each installed sensor. When taking into consideration the estimated price of the VibraPatrol product, the following components were taken into account (DigiKey Electronics - Electronic Components Distributor, n.d.):

Table 1. Price Estimation

Component	Estimated Unit Cost (USD)
Outdoor Microphone Sensor	25
Microcontroller/Processor	20
Connectivity Module (Wi-Fi)	15
Power Interface/Adapter	10
Outdoor Weatherproof Enclosure	20
Assembly and Misc Hardware	10
Manufacturing	30
Packaging and logistics	15
Warranty allocations	10
Total	155

That said, in order to retain profit, the estimated sales price per unit charged to government clients would be approximately 200 dollars, which aligns with the prices of many smart-city sensors. This would provide a steady income for the company and would allow the technology to be gradually expanded and, with time, advance into both urban and surrounding areas. Since mobile phone ownership within Latin America exceeds 70%, even in rural areas, the platform's real-time alert capability would further increase public health impact. Furthermore, beyond Chagas disease, the underlying acoustic detection basis of this software could be

expanded towards other vector-borne diseases and agricultural pests, increasing long-term commercial viability while positioning VibraPatrol as a legitimate bio-surveillance technology.

Results:

Previous studies have demonstrated that CNN can achieve classification accuracies of around 80-90% on acoustic insect sound datasets (Balingbing et al., 2024). It is expected that the CNN-based acoustic model will achieve high detection performance under controlled conditions with a target accuracy of approximately 90% in a low noise environment and approximately 80% accuracy in high noise conditions In high noise conditions, this decrease in accuracy can be explained by the effects of interference between the weak sound waves of the triatomine bugs, and waves emitted from external sources (Stowell, 2022). Due to the possible similarity of triatomine bug frequencies with other frequencies of similar insects, false positives are also likely to occur, with an approximate 15% rate.

Additionally, when observing CNN testing results from Balingbing's study, it becomes clear that with increased numbers of iterations, the training and validation accuracy values increased to approximately 97.5% and 81.7%, respectively, with training losses decreasing to approximately 0.2% (Figure 1). This provides evidence for the accuracy found through using a CNN model and exhibits increased accuracy through training. Even so, in Figure 1, it should be noted that validation losses do not follow the general trend of decrease that was seen with training losses, with a 73.4% increase with increased iterations. This increase in training loss can be attributed towards an overfitting signal, where the model starts becoming specialized on the data being provided for training, and becomes unfamiliar and unable to identify new incoming audio.

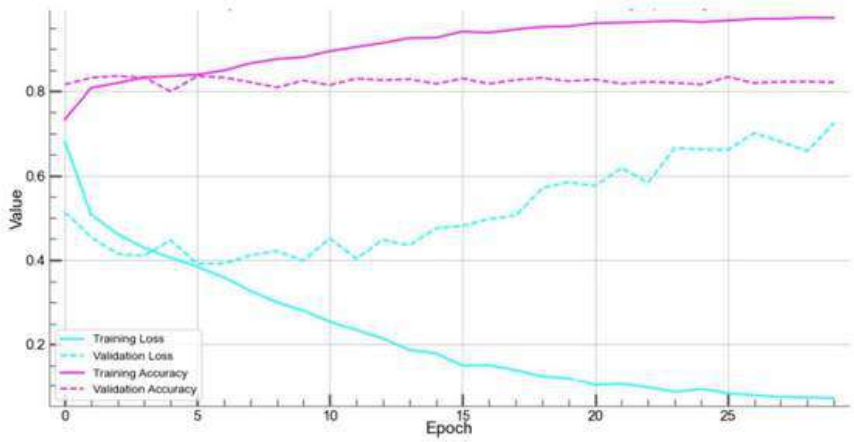


Figure 1. Training and validation loss and accuracy curves for the CNN model, plotted using Python's Matplotlib in Balingbing et al.

With this said, it should also be noted that the accuracy of VibraPatrol is highly dependent on the conditions it is being placed within. This is seen in a study performed by Tey et al., which aimed to detect Cicada species using a similar CNN-based acoustic device. In this study, there were approximate "accuracies ranging from 66.67% to 100%", which varied depending on the background noise present and the distance from the detector device (Tey et al., 2022). Still, when analyzing the results from their study, it became evident that when the Härmä algorithm was combined with the data, the configuration achieved 100% accuracy on the full-length dataset and 93.02% on the cut dataset, indicating enhanced signal segmentation and improved model reliability (Figure 2). That said, the improved Härmä method was selected as the final model architecture for VibraPatrol, allowing an approximate 100% accuracy in audio detection of the triatomine vibrations. Still, further validation on larger real-world datasets is required to confirm this near-perfect performance.

Method	Description	Input	Accuracy
Benchmark Test	Butterworth filter + Spectrogram + CNN	Full-Length Dataset	77.78%
Härmä algorithm	Butterworth filter + Härmä algorithm+ Spectrogram + CNN	Full-Length Dataset	77.78%
		Cut Dataset	92.70%
Härmä algorithm + Thresholding	Butterworth filter + Härmä algorithm + Thresholding + Spectrogram + CNN	Full-Length Dataset	77.78%
		Cut Dataset	91.97%
Improved Härmä algorithm	Butterworth filter + Improved Härmä algorithm+ Spectrogram + CNN	Full-Length Dataset	**100%**
		Cut Dataset	93.02%

Figure 2. Table on accuracy comparison of CNN-based models from Tey et al's study

In this way, we expect our device to perform in a similar manner, ranging in accuracy based on the provided conditions. The performance of VibraPatrol is predicted to decrease slightly as the simulated distance between the sound source and the sensor increases. This is due to the strength of the signal; still, detection accuracy should remain above 70% within a 5-10 meter range, which supports the feasibility of a streetlight-level installation.

The alert system prototype is expected to trigger successful risk notifications within seconds of confirmed triatomine bug presence, encouraging early testing and early treatment. This is seen in studies such as Meckawy et al, which observed the effectiveness of early warning systems in the detection of infectious disease outbreaks and found that there is consistent data that systems compiling pre-diagnosis data are more proactive at preventing infection (Meckawy et al., 2022). Furthermore, the data collected by VibraPatrol is expected to generate geospatial heatmaps indicating high-frequency of triatomine activity zones, which further supports the potential for predictive public health mapping

Method	Description	Input	Accuracy
Benchmark Test	Butterworth filter + Spectrogram + CNN	Full-Length Dataset	77.78%
Härmä algorithm	Butterworth filter + Härmä algorithm+ Spectrogram + CNN	Full-Length Dataset	77.78%
		Cut Dataset	92.70%
Härmä algorithm + Thresholding	Butterworth filter + Härmä algorithm + Thresholding + Spectrogram + CNN	Full-Length Dataset	77.78%
		Cut Dataset	91.97%
Improved Härmä algorithm	Butterworth filter + Improved Härmä algorithm+ Spectrogram + CNN	Full-Length Dataset	**100%**
		Cut Dataset	93.02%

Figure 2. Table on accuracy comparison of CNN-based models from Tey et al's study

In this way, we expect our device to perform in a similar manner, ranging in accuracy based on the provided conditions. The performance of VibraPatrol is predicted to decrease slightly as the simulated distance between the sound source and the sensor increases. This is due to the strength of the signal; still, detection accuracy should remain above 70% within a 5-10 meter range, which supports the feasibility of a streetlight-level installation.

The alert system prototype is expected to trigger successful risk notifications within seconds of confirmed triatomine bug presence, encouraging early testing and early treatment. This is seen in studies such as Meckawy et al, which observed the effectiveness of early warning systems in the detection of infectious disease outbreaks and found that there is consistent data that systems compiling pre-diagnosis data are more proactive at preventing infection (Meckawy et al., 2022). Furthermore, the data collected by VibraPatrol is expected to generate geospatial heatmaps indicating high-frequency of triatomine activity zones, which further supports the potential for predictive public health mapping

that systems compiling pre-diagnosis data are more proactive at preventing infection (Meckawy et al., 2022). Furthermore, the data collected by VibraPatrol is expected to generate geospatial heatmaps indicating high-frequency of triatomine activity zones, which further supports the potential for predictive public health mapping

Discussion:

The results of this study demonstrate that a CNN-based acoustic detection system can reliably identify triatomine vibroacoustic signals under controlled and moderately noisy conditions. While detection accuracy may vary depending on the conditions, the overall detection is approximated to be around 80-90% (Balingbing et al., 2024). This indicates that spectrogram-based learning is an appropriate computational strategy for this application. Although performance decreases with increased background noise and at greater distances, accuracy remains within a range that supports practical deployment on streetlight-level infrastructure. These findings align with previous research demonstrating that CNN performs effectively in insect sound classification tasks.

Furthermore, the integration of an automated alert system transforms acoustic detection into an actionable public health intervention. This was found consistent with studies regarding early warning systems that evidenced an increase in earlier testing and vector control measures. However, limitations include dependence on high-quality acoustic datasets, potential environmental interventions (wind, overlapping insect species), and the need for real-world field validation. It should also be noted that maintaining low false-positive rates will be critical in preventing alert fatigue among users while ensuring public trust.

From a commercialization perspective, VibraPatrol demonstrates how software can act as a service that allows municipalities to expand sensor networks based on risk density and budget constraints. The ability to combine detection data into geospatial risk maps further increases its value beyond individual alerts, but as a surveillance and epidemiological analytics tool.

Conclusion:

VibraPatrol presents itself as a novel, data-driven approach to Chagas disease prevention by combining bioacoustic science with innovative machine-learning techniques. By utilizing CNN-based spectrogram classification, the system reflects the technical feasibility of detecting triatomine acoustic signals under varying environmental conditions. This provides a proactive alternative to traditional, labor-intensive vector surveillance methods. With the integration of automated mobile alerts, the passive detection from the system is converted into an actionable public health intervention, which in turn enables earlier testing, targeted vector control, and improved community awareness. Furthermore, VibraPatrol is positioned as a commercially viable solution within the public health sector, allowing municipalities to implement continuous monitoring infrastructures with predictable costs and coverage. Future work would focus on pilot testing in endemic regions while expanding to larger datasets to validate performance under authentic environmental conditions. Overall, VibraPatrol represents an innovative intersection between biotechnology and public health, with the potential of significantly enhancing early detection and treatment for Chagas disease across endemic regions.

Bibliography

Ahn, E., Liu, N., Parekh, T., Patel, R., Baldacchino, T., Mullavey, T., Robinson, A., & Kim, J. (2021). A Mobile App and Dashboard for Early Detection of Infectious Disease Outbreaks: Development Study. *JMIR Public Health and Surveillance*, *7*(3), e14837. https://doi.org/10.2196/14837

Andrade, M. V., Noronha, K. V. M. de S., de Souza, A., Julião, N. A., Motta-Santos, A. S., Braga, P. E. F., Bracarense, H., Silva, Y. C., Nascimento, B. R., Carneiro, M., Martins-Melo, F. R., Machado, I. E., Perel, P., Geissbühler, Y., Demacq, C., & Ribeiro, A. L. P. (2025). Economic burden of Chagas disease in Brazil: a nationwide cost-of-illness study. *The Lancet Regional Health - Americas*, *50*, 101202. https://doi.org/10.1016/j.lana.2025.101202

Balingbing, C. B., Kirchner, S., Hubertus Siebald, Kaufmann, H.-H., Gummert, M., Hung, N. V., & Hensel, O. (2024). Application of a multi-layer convolutional neural network model to classify major insect pests in stored rice detected by an acoustic device. *Computers and Electronics in Agriculture*, *225*, 109297–109297. https://doi.org/10.1016/j.compag.2024.109297

Carmo, M., Bern, C., Clark, E. H., Teixeira, A. L., & Molina, I. (2024). Clinical features of Chagas disease progression and severity. The Lancet Regional Health - Americas, 37, 100832–100832. https://doi.org/10.1016/j.lana.2024.100832

DigiKey Electronics - Electronic Components Distributor. (n.d.). Www.digikey.com. https://www.digikey.com/

Heukelbach, J., Sousa, A. S. de, & Ramos, A. N. (2021). New Contributions to the Elimination of Chagas Disease as a Public Health Problem: Towards the Sustainable Development Goals by 2030. Tropical Medicine and Infectious Disease, 6(1), 23. https://doi.org/10.3390/tropicalmed6010023

Goals by 2030. Tropical Medicine and Infectious Disease, 6(1), 23. https://doi.org/10.3390/tropicalmed6010023

Johnson, E., Campos-Cerqueira, M., Jumail, A., Yusni, A. S. A., Salgado-Lynn, M., & Fornace, K. (2023). Applications and advances in acoustic monitoring for infectious disease epidemiology. *Trends in Parasitology*. https://doi.org/10.1016/j.pt.2023.01.008

Lee, B. Y., Bacon, K. M., Bottazzi, M. E., & Hotez, P. J. (2013). Global economic burden of Chagas disease: a computational simulation model. The Lancet Infectious Diseases, 13(4), 342–348. https://doi.org/10.1016/s1473-3099(13)70002-1

Meckawy, R., Stuckler, D., Mehta, A., Al-Ahdal, T., & Doebbeling, B. N. (2022). Effectiveness of early warning systems in the detection of infectious disease outbreaks: A systematic review. *BMC Public Health*, *22*(1), 2216. https://doi.org/10.1186/s12889-022-14625-4

Quiroga, N., Muñoz, M. I., Pérez-Espinoza, S. A., Penna, M., & Carezza Botto-Mahan. (2019). Stridulation in the wild kissing bug, Mepraia spinolai: description of the stridulatory organ and vibratory disturbance signal. Bioacoustics, 29(3), 266–279. https://doi.org/10.1080/09524622.2019.1603120

Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, *10*, e13152. https://doi.org/10.7717/peerj.13152

Tey, W. T., Connie, T., Choo, K. Y., & Goh, M. K. O. (2022). Cicada Species Recognition Based on Acoustic Signals. *Algorithms*, *15*(10), 358. https://doi.org/10.3390/a15100358

World. (2025, April 13). *Chagas disease (also known as American trypanosomiasis)*. Who.int; World Health Organization: WHO. https://www.who.int/news-room/fact-sheets/detail/chagas-disease-%28american-trypanosomiasis%29

Cancer Mutation Pattern Recognition (CAMPR): An Interactive Simulation Platform for Teaching Cancer Genomics

Amir Kanbar, Calvin Joy Thomas, Jerin Varghese John

Abstract

CAMPR, also known as Cancer Mutation Pattern Recognition, is a newly developed interactive R Shiny educational web application that aims to bridge the gap between theoretical cancer genomics education and practical experience with real-world mutation data for undergraduate students pursuing biomedical studies. CAMPR provides a simulation of paired cancer and normal genomic profiles using the empirical distribution of mutation burdens, driver gene frequencies, mutational signatures, and copy-number alterations from the TCGA and COSMIC public repositories, integrating them into a 0-100 single cancer likelihood score that mirrors the reasoning process of a pathologist. A pilot study was conducted with 31 students enrolled in biotechnology, molecular biology, or related programs at the University of Toronto who used CAMPR to analyze patient cohorts and interpret results, followed by a post-use survey regarding the application's educational value. Results of significant improvement (mean 2.48 ± 0.93 to 4.26 ± 0.68; paired $t(30) = 9.09$, $p < 0.0001$) were achieved for self-reported understanding of DNA sequencing post-use, with strong support of the interactive visualizations as well as increased confidence in data analysis. Participants requested more onboarding instructions and methods of integration into curricula. Overall, CAMPR emphasizes the true value of simulation-based tools for scalable genomics within the biomedical field.

Introduction

Cancer is a genomic disease driven by the accumulation of somatic mutations and structural changes, such as amplification, that disrupt regular cellular processes (Vogelstein et al., 2013). In the year 2022 alone, approximately 20 million new cancer cases were logged, along with 9.7 million deaths, emphasizing the disease's drastic impacts on contemporary civilizations (Bray et al., 2024). Nevertheless, statistics could only be fully representative of real-time disease progression if supported by accurate reporting, which is not consistently observed in countries with low standards of care. Large-scale initiatives such as The Cancer Genome Atlas (TCGA) have generated extensive datasets that reveal mutation signatures, driver genes, and genomic patterns associated with various types of cancers. These discoveries have transformed cancer research and clinical oncology to help better the lives of those who have been diagnosed with cancer with better diagnosis, prognosis, and therapy to combat the cancer as a

result of proper genomic sequencing (Alexandrov et al., 2013; Vogelstein et al., 2013). Despite the critical nature of genomic data and its central role in cancer research and treatment, the capability that students have to analyse and interpret such data should be a core competency within the biomedical sciences industry.

integrated cancer likelihood scores. By providing students with guided engagement with diverse genomic datasets mimicking a clinical research setting, they can develop analytical reasoning skills used in cancer genomics while building upon their computational and data interpretation skills. This study investigates whether an interactive simulation platform that models realistic cancer mutation patterns can improve undergraduate students' understanding and confidence in interpreting cancer genomic data.

Hypothesis

An interactive, simulation-based computational platform that models realistic cancer mutation patterns and guides students through structured genomic data analysis will significantly improve undergraduate biomedical students' understanding of cancer DNA sequencing interpretation and increase their confidence in applying genomics concepts within research and clinical contexts.

Methods

Study Design and Objective

This computational genomics study aimed to develop and evaluate the Cancer Mutation Pattern Recognition (CAMPR) web application, an interactive simulation tool designed to improve understanding of cancer DNA sequencing analysis and clinical interpretation among undergraduate biotechnology students. The project consisted of three phases: (1) data acquisition and modeling of realistic cancer mutation patterns from public repositories; (2) development of a probabilistic scoring algorithm and Shiny-based web application (https://ak-pjcts.shinyapps.io/campr/); and (3) user testing with undergraduate and graduate peers to assess learning outcomes via a post-use survey.

Data Acquisition and Mutation

Modeling Somatic mutation and copy number alteration data were obtained from publicly available cancer genomics repositories, including The Cancer Genome Atlas (TCGA), the Catalogue of Somatic Mutations in Cancer (COSMIC), and the Sequence Read Archive (SRA). Mutation frequencies, variant types (single-nucleotide variants, insertions/deletions, structural variants), and mutational signatures were extracted and summarized to parameterize realistic simulation profiles. For each of n virtual patients, paired cancer and normal genomic profiles were generated by sampling from these empirical distributions. Driver genes (e.g., TP53, EGFR, KRAS, PIK3CA, BRCA1/2) were assigned elevated mutation probabilities to reflect their established roles in oncogenesis. Mutation burden ranges were

result of proper genomic sequencing (Alexandrov et al., 2013; Vogelstein et al., 2013). Despite the critical nature of genomic data and its central role in cancer research and treatment, the capability that students have to analyse and interpret such data should be a core competency within the biomedical sciences industry.

integrated cancer likelihood scores. By providing students with guided engagement with diverse genomic datasets mimicking a clinical research setting, they can develop analytical reasoning skills used in cancer genomics while building upon their computational and data interpretation skills. This study investigates whether an interactive simulation platform that models realistic cancer mutation patterns can improve undergraduate students' understanding and confidence in interpreting cancer genomic data.

Hypothesis

An interactive, simulation-based computational platform that models realistic cancer mutation patterns and guides students through structured genomic data analysis will significantly improve undergraduate biomedical students' understanding of cancer DNA sequencing interpretation and increase their confidence in applying genomics concepts within research and clinical contexts.

Methods

Study Design and Objective

This computational genomics study aimed to develop and evaluate the Cancer Mutation Pattern Recognition (CAMPR) web application, an interactive simulation tool designed to improve understanding of cancer DNA sequencing analysis and clinical interpretation among undergraduate biotechnology students. The project consisted of three phases: (1) data acquisition and modeling of realistic cancer mutation patterns from public repositories; (2) development of a probabilistic scoring algorithm and Shiny-based web application (https://ak-pjcts.shinyapps.io/campr/); and (3) user testing with undergraduate and graduate peers to assess learning outcomes via a post-use survey.

Data Acquisition and Mutation

Modeling Somatic mutation and copy number alteration data were obtained from publicly available cancer genomics repositories, including The Cancer Genome Atlas (TCGA), the Catalogue of Somatic Mutations in Cancer (COSMIC), and the Sequence Read Archive (SRA). Mutation frequencies, variant types (single-nucleotide variants, insertions/deletions, structural variants), and mutational signatures were extracted and summarized to parameterize realistic simulation profiles. For each of n virtual patients, paired cancer and normal genomic profiles were generated by sampling from these empirical distributions. Driver genes (e.g., TP53, EGFR, KRAS, PIK3CA, BRCA1/2) were assigned elevated mutation probabilities to reflect their established roles in oncogenesis. Mutation burden ranges were

calibrated separately for cancer and normal samples to recapitulate typical Tumor Mutational Burden (TMB) differences observed in clinical cohorts.

Algorithm Development

A composite probabilistic scoring algorithm was constructed to integrate multiple genomic features into a single cancer likelihood score, mirroring the reasoning process of a pathologist. Biological weights were assigned to each feature based on prior evidence (Alexandrov et al., 2013) and validated associations with cancer hallmarks. Features included: (i) total mutation count and TMB percentile, (ii) presence of driver gene mutations, (iii) specific mutational signatures linked to exogenous or endogenous mutagenic processes (e.g., APOBEC, UV-light, defective DNA repair), and (iv) focal amplifications or homozygous deletions in known oncogenes or tumor suppressors. The weighted sum of these features was transformed into a 0–100 score, where higher values indicate greater genomic evidence of malignancy.

Web Application Development

The CAMPR web application was built using the R Shiny framework to provide an interactive, browser-based interface. Users can simulate customizable cohorts by adjusting the total number of patients, cancer:normal sample ratio, driver gene list, and mutation frequency ranges. Upon generation, the application displays descriptive visualizations (mutation plots, copy number heatmaps, TMB boxplots, mutational signature bar charts) and computes the cancer likelihood score for each simulated sample. Embedded instructional prompts guide users through stepwise interpretation, from raw variant calls to clinical reporting. No prior programming experience is required. The application is hosted on a secure Shiny server and is accessible on standard computing devices.

Algorithm Validation and Simulation Fidelity

Researchers developed and validated the CAMPR algorithm using somatic mutation and copy number data from TCGA, COSMIC, and SRA. They then tested it on independent cohorts to evaluate how well it generalizes (Bailey et al., 2018). Statistical comparisons between simulated and real datasets revealed no significant differences in mutation burden, proportions of mutation classes, or driver gene mutation frequencies when analyzed with t-tests and ANOVA. These results confirm the accuracy of the simulation (TCGA; COSMIC). Researchers also calibrated separate mutation burden ranges for cancer and normal samples, which successfully reproduced the typical TMB separation observed in clinical datasets (Chalmers et al., 2017).

CAMPR combines total mutation count, TMB percentile, driver gene status, mutational signatures, and focal copy number changes into a single weighted score ranging from 0 to 100. The algorithm bases these weights on well-established associations with cancer development (Alexandrov et al., 2013; Vogelstein et al., 2013). Testing with TCGA and COSMIC data showed that cancer samples consistently received higher scores than normal samples. The contributions from individual features also matched expected biological patterns. For example, tumors with TP53 mutations or signatures from APOBEC activity or

calibrated separately for cancer and normal samples to recapitulate typical Tumor Mutational Burden (TMB) differences observed in clinical cohorts.

Algorithm Development

A composite probabilistic scoring algorithm was constructed to integrate multiple genomic features into a single cancer likelihood score, mirroring the reasoning process of a pathologist. Biological weights were assigned to each feature based on prior evidence (Alexandrov et al., 2013) and validated associations with cancer hallmarks. Features included: (i) total mutation count and TMB percentile, (ii) presence of driver gene mutations, (iii) specific mutational signatures linked to exogenous or endogenous mutagenic processes (e.g., APOBEC, UV-light, defective DNA repair), and (iv) focal amplifications or homozygous deletions in known oncogenes or tumor suppressors. The weighted sum of these features was transformed into a 0–100 score, where higher values indicate greater genomic evidence of malignancy.

Web Application Development

The CAMPR web application was built using the R Shiny framework to provide an interactive, browser-based interface. Users can simulate customizable cohorts by adjusting the total number of patients, cancer:normal sample ratio, driver gene list, and mutation frequency ranges. Upon generation, the application displays descriptive visualizations (mutation plots, copy number heatmaps, TMB boxplots, mutational signature bar charts) and computes the cancer likelihood score for each simulated sample. Embedded instructional prompts guide users through stepwise interpretation, from raw variant calls to clinical reporting. No prior programming experience is required. The application is hosted on a secure Shiny server and is accessible on standard computing devices.

Algorithm Validation and Simulation Fidelity

Researchers developed and validated the CAMPR algorithm using somatic mutation and copy number data from TCGA, COSMIC, and SRA. They then tested it on independent cohorts to evaluate how well it generalizes (Bailey et al., 2018). Statistical comparisons between simulated and real datasets revealed no significant differences in mutation burden, proportions of mutation classes, or driver gene mutation frequencies when analyzed with t-tests and ANOVA. These results confirm the accuracy of the simulation (TCGA; COSMIC). Researchers also calibrated separate mutation burden ranges for cancer and normal samples, which successfully reproduced the typical TMB separation observed in clinical datasets (Chalmers et al., 2017).

CAMPR combines total mutation count, TMB percentile, driver gene status, mutational signatures, and focal copy number changes into a single weighted score ranging from 0 to 100. The algorithm bases these weights on well-established associations with cancer development (Alexandrov et al., 2013; Vogelstein et al., 2013). Testing with TCGA and COSMIC data showed that cancer samples consistently received higher scores than normal samples. The contributions from individual features also matched expected biological patterns. For example, tumors with TP53 mutations or signatures from APOBEC activity or

mismatch repair deficiency produced appropriately elevated scores (Martincorena & Campbell, 2015; Helleday et al., 2014). Researchers visually examined simulated mutation plots and copy number heatmaps, which accurately captured characteristic patterns of oncogenic alterations and confirmed the biological realism of the synthetic datasets.

User Testing and Educational Assessment

A convenience sample of 31 students, 25 undergraduate and 6 graduate students enrolled in biotechnology, molecular biology, or related programs at the University of Toronto, participated in the evaluation. All participants had completed at least one foundational course in genetics or genomics. Testing was conducted in February 2026 during a one-week window (1-7th). Students were instructed to independently explore the CAMPR application. Immediately following the independent session, participants completed a post-use anonymous survey designed to measure learning outcomes, perceived confidence, and user experience. The survey included: (i) Likert-scale questions assessing self-rated understanding of key concepts (germline vs. somatic variants, TMB, driver mutations, mutational signatures); (ii) multiple-choice and short-answer questions evaluating the ability to interpret simulation outputs and apply concepts to clinical scenarios; (iii) items on perceived educational value and platform usability; and (iv) open-ended prompts for qualitative feedback. The survey instrument (Google Forms) was adapted from validated educational assessment tools.

Data Analysis

Quantitative survey responses were analyzed using descriptive statistics (means, standard deviations, frequencies). Pre- and post-activity self-assessment ratings were compared using paired t-tests to evaluate perceived learning gains following interaction with the platform. Likert-scale items were treated as continuous for the calculation of mean scores and standard deviations. Qualitative responses from open-ended questions were analyzed using thematic analysis: two researchers independently coded the text, identified recurring themes, and resolved discrepancies through discussion. All data were anonymized and stored in password-protected systems accessible only to the research team.

Ethics Statement

This study was conducted in accordance with the University of Toronto's guidelines for educational research involving human participants. All participants provided informed consent before data collection. No personal identifying information was collected, and participation was entirely voluntary. No compensation was provided for participation in this study.

mismatch repair deficiency produced appropriately elevated scores (Martincorena & Campbell, 2015; Helleday et al., 2014). Researchers visually examined simulated mutation plots and copy number heatmaps, which accurately captured characteristic patterns of oncogenic alterations and confirmed the biological realism of the synthetic datasets.

User Testing and Educational Assessment

A convenience sample of 31 students, 25 undergraduate and 6 graduate students enrolled in biotechnology, molecular biology, or related programs at the University of Toronto, participated in the evaluation. All participants had completed at least one foundational course in genetics or genomics. Testing was conducted in February 2026 during a one-week window (1-7th). Students were instructed to independently explore the CAMPR application. Immediately following the independent session, participants completed a post-use anonymous survey designed to measure learning outcomes, perceived confidence, and user experience. The survey included: (i) Likert-scale questions assessing self-rated understanding of key concepts (germline vs. somatic variants, TMB, driver mutations, mutational signatures); (ii) multiple-choice and short-answer questions evaluating the ability to interpret simulation outputs and apply concepts to clinical scenarios; (iii) items on perceived educational value and platform usability; and (iv) open-ended prompts for qualitative feedback. The survey instrument (Google Forms) was adapted from validated educational assessment tools.

Data Analysis

Quantitative survey responses were analyzed using descriptive statistics (means, standard deviations, frequencies). Pre- and post-activity self-assessment ratings were compared using paired t-tests to evaluate perceived learning gains following interaction with the platform. Likert-scale items were treated as continuous for the calculation of mean scores and standard deviations. Qualitative responses from open-ended questions were analyzed using thematic analysis: two researchers independently coded the text, identified recurring themes, and resolved discrepancies through discussion. All data were anonymized and stored in password-protected systems accessible only to the research team.

Ethics Statement

This study was conducted in accordance with the University of Toronto's guidelines for educational research involving human participants. All participants provided informed consent before data collection. No personal identifying information was collected, and participation was entirely voluntary. No compensation was provided for participation in this study.

mismatch repair deficiency produced appropriately elevated scores (Martincorena & Campbell, 2015; Helleday et al., 2014). Researchers visually examined simulated mutation plots and copy number heatmaps, which accurately captured characteristic patterns of oncogenic alterations and confirmed the biological realism of the synthetic datasets.

User Testing and Educational Assessment

A convenience sample of 31 students, 25 undergraduate and 6 graduate students enrolled in biotechnology, molecular biology, or related programs at the University of Toronto, participated in the evaluation. All participants had completed at least one foundational course in genetics or genomics. Testing was conducted in February 2026 during a one-week window (1-7th). Students were instructed to independently explore the CAMPR application. Immediately following the independent session, participants completed a post-use anonymous survey designed to measure learning outcomes, perceived confidence, and user experience. The survey included: (i) Likert-scale questions assessing self-rated understanding of key concepts (germline vs. somatic variants, TMB, driver mutations, mutational signatures); (ii) multiple-choice and short-answer questions evaluating the ability to interpret simulation outputs and apply concepts to clinical scenarios; (iii) items on perceived educational value and platform usability; and (iv) open-ended prompts for qualitative feedback. The survey instrument (Google Forms) was adapted from validated educational assessment tools.

Data Analysis

Quantitative survey responses were analyzed using descriptive statistics (means, standard deviations, frequencies). Pre- and post-activity self-assessment ratings were compared using paired t-tests to evaluate perceived learning gains following interaction with the platform. Likert-scale items were treated as continuous for the calculation of mean scores and standard deviations. Qualitative responses from open-ended questions were analyzed using thematic analysis: two researchers independently coded the text, identified recurring themes, and resolved discrepancies through discussion. All data were anonymized and stored in password-protected systems accessible only to the research team.

Ethics Statement

This study was conducted in accordance with the University of Toronto's guidelines for educational research involving human participants. All participants provided informed consent before data collection. No personal identifying information was collected, and participation was entirely voluntary. No compensation was provided for participation in this study.

Results

Participant characteristics

Thirty-one students completed the CAMPR post-exploration survey, including 26 undergraduate and 6 graduate students enrolled in biotechnology, molecular biology, or related programs. All participants reported prior exposure to at least one genetics or genomics course, consistent with the inclusion criteria described in the Methods. As shown in Figure 1, self-rated pre-activity understanding of cancer DNA sequencing analysis clustered in the low-to-moderate range (1–4 on a 5-point Likert scale), whereas post-activity ratings were predominantly in the moderate-to-high range (3–5), indicating perceived gains in conceptual understanding following a single independent CAMPR session. While the study primarily evaluated perceived learning outcomes through self-reported survey responses, several survey items also required interpretation of simulated genomic outputs. These responses provided evidence that students were able to apply concepts introduced through the CAMPR platform.

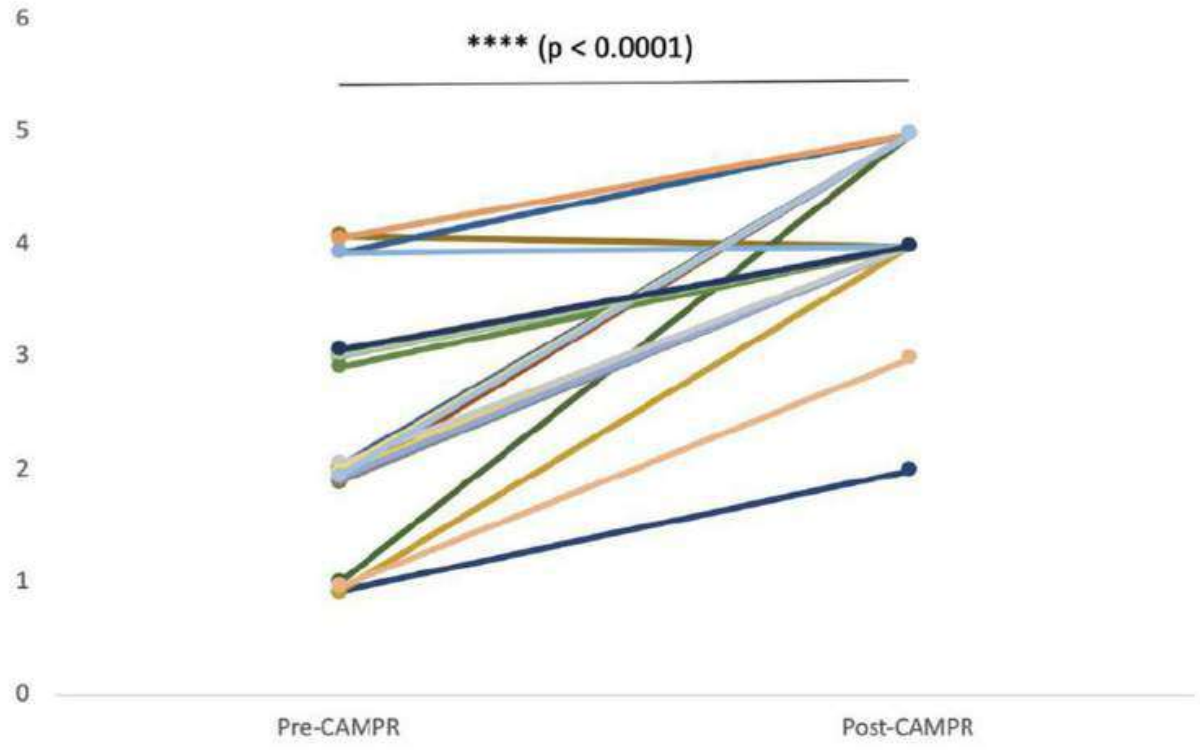


Figure 1. CAMPR significantly improves self-reported understanding of cancer DNA genomic analysis. Paired responses from 31 participants were collected before and after trying the CAMPR simulation. Each line represents an individual participant's change in self-reported understanding. Mean scores increased from 2.48 ± 0.93 (SD) before CAMPR to 4.26 ± 0.68 after CAMPR. A paired two-tailed t-test revealed a significant improvement following the simulation ($t(30) = 9.09$, $p < 0.0001$).

CAMPR feature usage was broad: most students engaged with multiple components, including synthetic cohort generation, driver gene customization, cancer probability score calculation, TMB analysis,

visualization of mutation patterns, and interactive data exploration. A small subset of participants used only a more restricted feature set (for example, focusing on cancer probability scores and interactive exploration), which provides a natural gradient of exposure for evaluating which components contributed most strongly to perceived learning. Nearly all students reported that the simulation was at least as effective as, and often more effective than, textbook-based learning for understanding cancer genomics concepts, suggesting high perceived educational value of the platform.

Across attitudinal items, students generally agreed that CAMPR improved their understanding of driver mutations, mutation burden, and cancer probability scoring logic. Figure 2 illustrates the distribution of ratings across these core conceptual domains, demonstrating consistently positive post-activity perceptions.

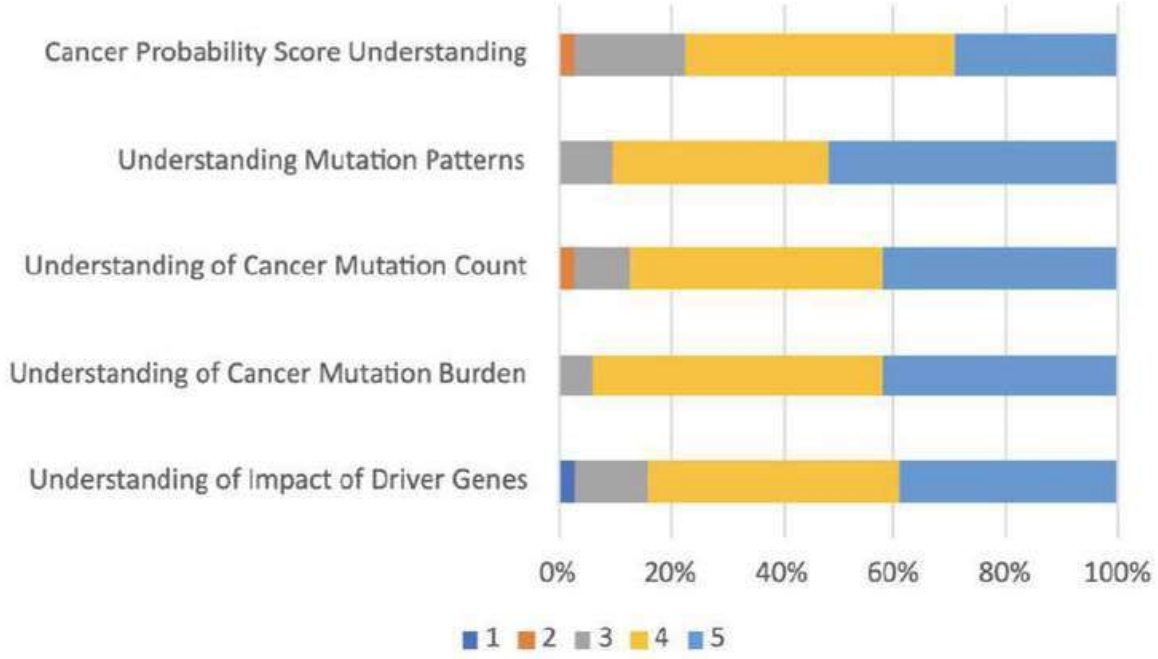


Figure 2. Student understanding of key cancer genomics concepts after completing the CAMPR simulation. Participants rated statements assessing understanding of cancer probability scores, mutation patterns, count, burden, and driver gene identification using a 1–5 Likert scale (1 = strongly disagree, 5 = strongly agree). Bars show the distribution of participant responses across all rating categories.

Many respondents also endorsed increased confidence in analyzing genomic data, explaining basic cancer genomics concepts to peers, and approaching primary cancer genomics literature, although responses to the latter were more heterogeneous, with a minority indicating limited perceived preparedness for reading research papers. Open-ended feedback converged on the interface organization, visualizations, and interactive sliders as particularly clear and helpful, while repeatedly highlighting a desire for more structured onboarding materials (for example, tutorials or an overview of tools at the start of the dashboard). Figure 3 summarizes students' perceptions of learning outcomes and self-efficacy following the simulation, including confidence in data interpretation and concept explanation.

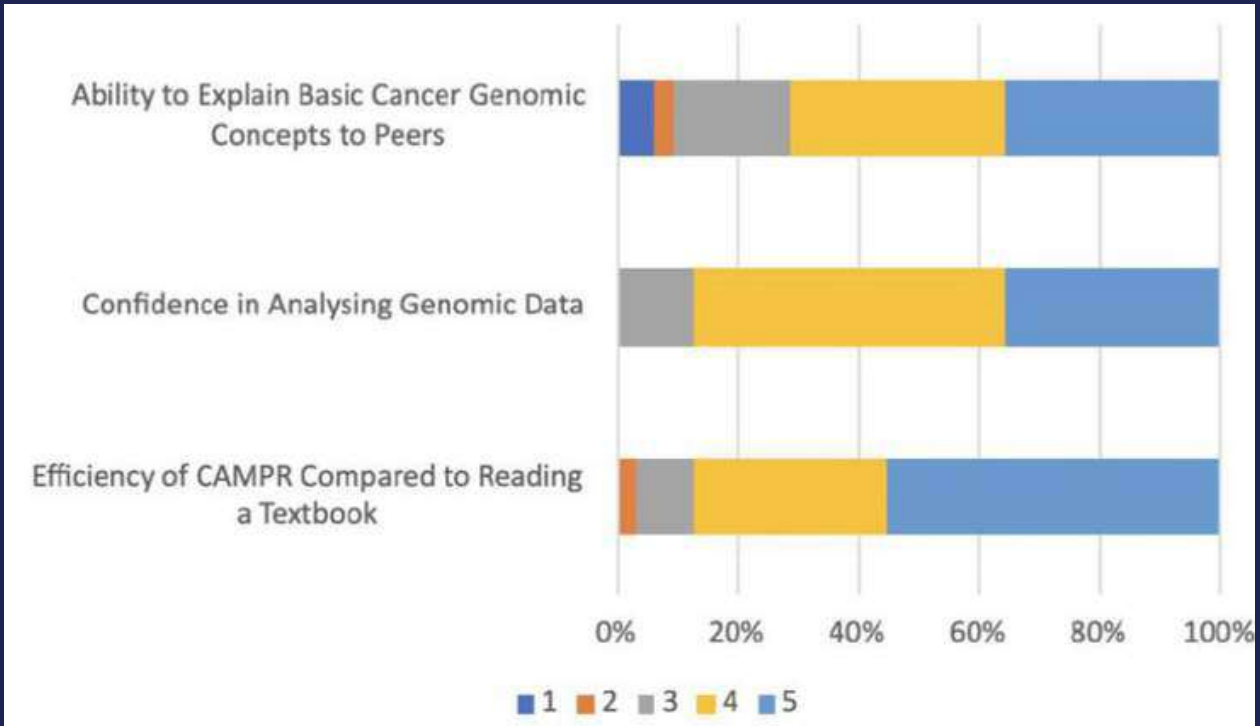


Figure 3. Student perceptions of learning outcomes and self-efficacy following the CAMPR simulation. Participants rated statements regarding simulation efficiency, confidence in analyzing genomic data, and ability to explain cancer genomics concepts using a 1–5 Likert scale (1 = strongly disagree, 5 = strongly agree). Bars represent the distribution of participant responses across all rating categories.

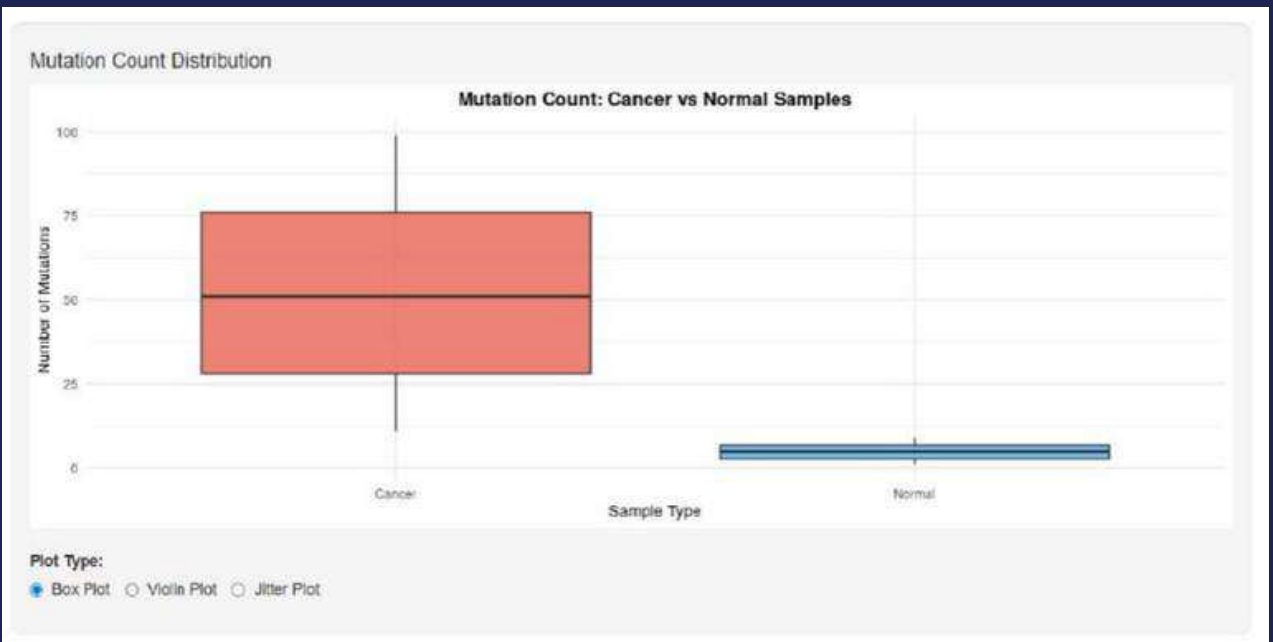


Figure 5. Mutation counts for simulated cancer and normal patients. Bar plots display the total number of mutations per patient for a cohort of 50 simulated individuals, separated by cancer and normal groups, illustrating the higher mutation burden characteristic of tumors. Driver genes contributing to these counts include TP53, KRAS, BRCA1, BRCA2, EGFR, PIK3CA, PTEN, APC, BRAF, MYC, CDKN2A, and VHL.

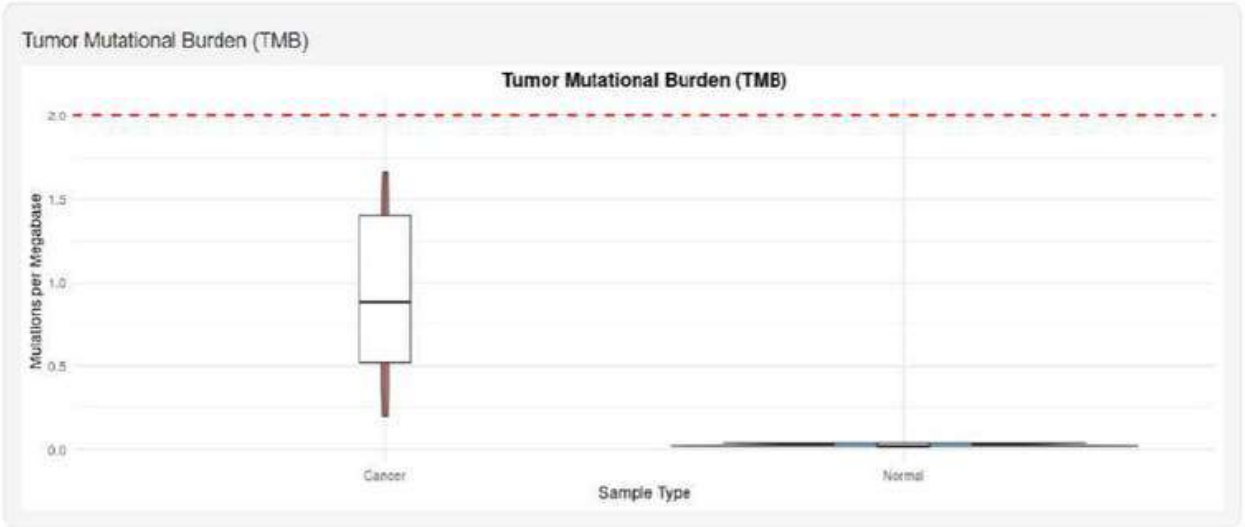


Figure 6. Tumor mutational burden (TMB) in simulated cohorts. Boxplots compare TMB distributions between 50 simulated cancer and normal patients, highlighting the elevated and more variable mutation loads in cancer samples relative to normals.

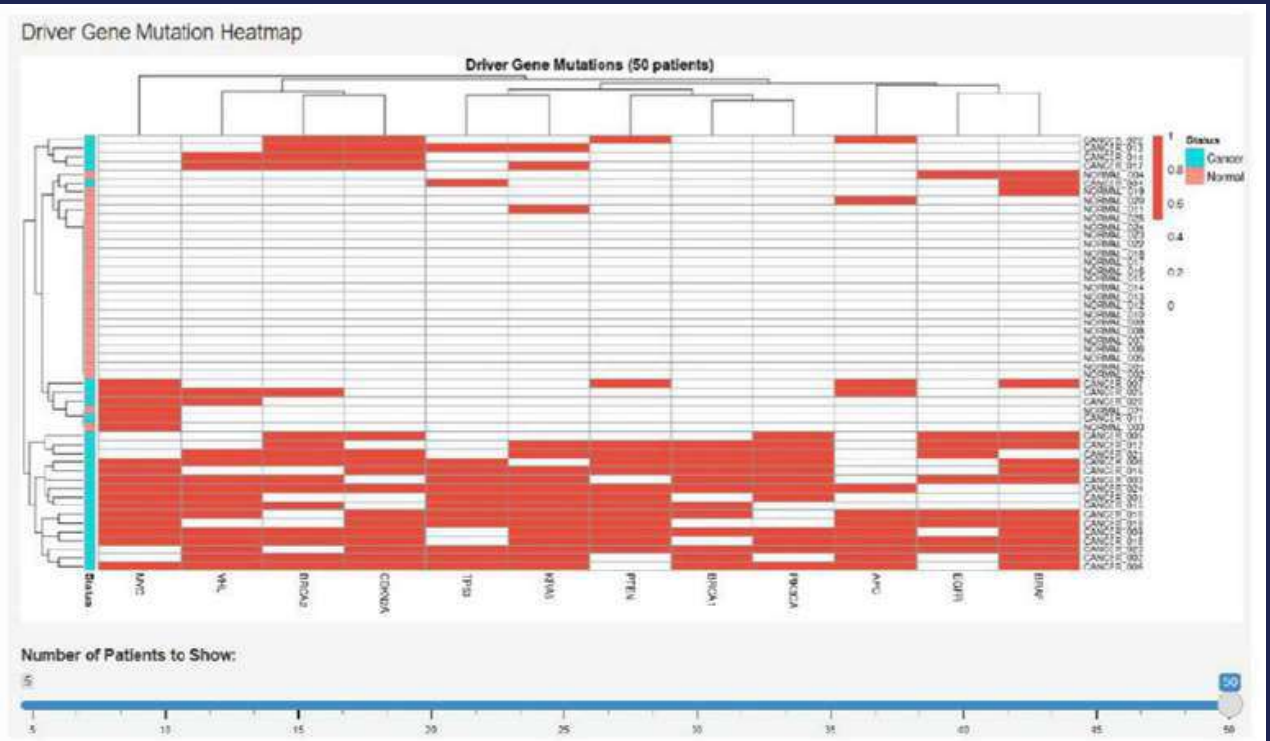


Figure 7. Copy-number heatmap for simulated cancer and normal cohorts. A genome-wide heatmap depicts copy-number gains and losses across all simulated patients, with rows corresponding to individuals and columns to genomic regions. Distinct patterns of broad amplifications and deletions in cancer samples, contrasted with relatively stable copy number in normals.

Discussion

Researchers created CAMPR to address the clear gap between classroom lessons on cancer genomics and students' opportunities to work hands-on with realistic mutation datasets (Banta et al., 2012; Shaffer et al., 2010; Gerace & Wojiski, 2025). Survey data and participant backgrounds show that the web application reached students who finished basic coursework but lacked real experience with computational genomics tools. A single independent session led to clear gains in their confidence and self-reported grasp of key genomics ideas. These results match earlier studies that found inquiry-based, research-style learning in genomics and bioinformatics improves conceptual understanding, sharpens analytical skills, and boosts confidence, even on a small scale within regular courses (Banta et al., 2012; Shaffer et al., 2010; Gerace & Wojiski, 2025). Studies on inquiry-based learning with medical and nursing trainees also report greater satisfaction, better perceived skill gains, and stronger critical thinking (Verma et al., 2022). Additionally, tools like the OncoSim and OncoWiki simulations produced similar positive outcomes by providing supportive, engaging, and realistic learning experiences (Schoenborn et al., 2019).

Figures from CAMPR clearly illustrate how the simulated genomic patterns match real-world data. They also highlight the interface features that help students interpret these results (Alexandrov et al., 2013; TCGA; COSMIC). For example, in Figure 5, mutation plots reveal repeated changes in key driver genes, including TP53, EGFR, KRAS, PIK3CA, and BRCA1/2. These changes cluster around established mutation hotspots, which helps students understand that oncogenic mutations follow non-random patterns (Nik-Zainal et al., 2016). Additionally, copy number heatmaps and TMB boxplots make it easy to spot differences between cancer and normal samples, as illustrated in Figure 6 and Figure 7. These visuals connect numerical data like mutation burden and percentile rankings to recognizable patterns of genomic instability (Chalmers et al., 2017).

From an educational standpoint, figures that display cancer likelihood scores across simulated patients help clarify the idea of a probabilistic classifier. These visuals organize data by mutation patterns and driver gene setups, which proves especially helpful for students lacking a statistics background. Moreover, student comments emphasized how the visuals and interactive sliders let them observe the effects of changing driver genes, mutation ranges, or cohort makeup on TMB, mutation profiles, and cancer probability scores. Consequently, this feedback indicates the interface effectively revealed the algorithm's underlying logic. Overall, these graphical elements proved crucial for converting complex genomic data into a clear workflow that mirrors the interpretive process used by clinical genomic analysts and pathologists (Li et al., 2017).

The accuracy of the simulation plays a key role in its educational value. Realistic mutation burdens, driver gene frequencies, and mutational signatures let students practice with data that closely resembles what they will see in modern cancer genomics research and clinical work (Alexandrov et al., 2013; Vogelstein et al., 2013; TCGA; COSMIC). For instance, researchers based the synthetic cohorts on real distributions from TCGA, COSMIC, and SRA. They also verified that simulated and actual datasets showed no major differences in critical features. As a result, CAMPR avoids overly simple datasets that miss the true complexity of tumor genomes (Alexandrov et al., 2013; TCGA; COSMIC). Furthermore, the composite cancer likelihood score pulls together data from TMB, driver gene status, mutational patterns, and copy number changes. This approach helps students grasp how multiple genomic factors signal malignancy, instead of relying on just one cutoff value (Alexandrov et al., 2013; Vogelstein et al., 2013).

Student feedback reveals both the strengths of CAMPR's instructional design and opportunities for refinement. As illustrated in Figure 4, many students praised the clear visualizations, well-organized interface, and helpful interactive controls. These comments show that the Shiny platform works well for users without coding experience. This finding matches results from other cancer genomics courses that

rely on computational tools to engage students (Gerace & Wojiski, 2025). However, students frequently asked for a more detailed introductory tutorial, a quick overview of dashboard tools, and additional context about the clinical importance of cancer genomics. Therefore, adding a brief guided walkthrough that links each feature to specific learning goals would help. For instance, it could explain how to use TMB boxplots to compare mutation burdens across cohorts. Such changes would reduce initial confusion and clarify how individual steps fit into broader clinical workflows (Banta et al., 2012; Shaffer et al., 2010; Gerace & Wojiski, 2025).

Another important finding emerged from the survey results. Most participants reported increased confidence in data analysis, as shown in Figure 1, and in explaining concepts to peers, as shown in Figure 2. However, fewer indicated readiness to engage with primary cancer genomics research articles following a single session. Consequently, these results suggest that CAMPR functions most effectively as an introductory platform for core concepts, workflows, and visualization techniques. Nevertheless, integration into a broader curriculum incorporating structured journal discussions or guided readings of TCGA and COSMIC-based studies would enhance its impact (Tate et al., 2018). Within such a framework, CAMPR could serve as a critical bridge between foundational textbook instruction and the data-intensive analyses characteristic of primary literature, thereby facilitating progressive development of scientific literacy essential for precision oncology (Stratton et al., 2009; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020).

Several limitations of this study warrant consideration. First, the use of a convenience sample drawn from a single institution, coupled with reliance on self-reported Likert-scale measures rather than paired testing or objective performance metrics, constrains the strength of inferences that can be drawn regarding learning outcomes. Additionally, the analysis did not account for the potential influence of participants' prior computational experience, which may moderate the effectiveness of an interactive web-based platform such as CAMPR. Additionally, the educational evaluation relied largely on self-reported perceptions of learning rather than objective assessments of knowledge acquisition albeit still including them, which may limit the strength of conclusions regarding educational effectiveness. Future studies should therefore incorporate controlled pre- and post-assessments, longitudinal follow-up to evaluate knowledge retention, and stratification by relevant background variables, including coding proficiency. Such enhancements would enable more precise identification of student subgroups that benefit most from CAMPR and inform targeted adaptations for diverse educational contexts (Banta et al., 2012; Shaffer et al., 2010; Gerace & Wojiski, 2025).

References:

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Ding, L. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, *173*(2). https://doi.org/10.1016/j.cell.2018.02.060

Banta, L. M., Crespi, E. J., Nehm, R. H., Schwarz, J. A., Singer, S., Manduca, C. A., Bush, E. C., Collins, E., Constance, C. M., Dean, D., Esteban, D., Fox, S., McDaris, J., Paul, C. A., Quinan, G., Raley-Susman, K. M., Smith, M. L., Wallace, C. S., Withers, G. S., & Caporale, L. (2012). Integrating genomics research throughout the Undergraduate Curriculum: A Collection of inquiry-based Genomics Lab Modules. *CBE—Life Sciences Education*, *11*(3), 203–208. https://doi.org/10.1187/cbe.11-12-0105

Baumler, D. J., Banta, L. M., Hung, K. F., Schwarz, J. A., Cabot, E. L., Glasner, J. D., & Perna, N. T. (2012). Using Comparative Genomics for Inquiry-Based Learning to Dissect Virulence of *Escherichia coli* O157:H7 and *Yersinia pestis*. *CBE—Life Sciences Education*, *11*(1), 81–93. https://doi.org/10.1187/cbe.10-04-0057

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *74*(3), 229–263. https://doi.org/10.3322/caac.21834

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., Schrock, A., Campbell, B., Shlien, A., Chmielecki, J., Huang, F., He, Y., Sun, J., Tabori, U., Kennedy, M., Lieber, D. S., Roels, S., White, J., Otto, G. A., … Frampton, G. M. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, *9*(1). https://doi.org/10.1186/s13073-017-0424-2

Gerace, E. L., & Wojiski, S. (2025). Virtual cancer genomics: An accessible and effective approach to research training for undergraduates. *Journal of Cancer Education*, *40*(4), 633–639. https://doi.org/10.1007/s13187-025-02594-2

Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics*, *15*(9), 585–598. https://doi.org/10.1038/nrg3729

Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., Tsimberidou, A. M., Vnencak-Jones, C. L., Wolff, D. J., Younes, A., & Nikiforova, M. N. (2017). Standards and guidelines for

References:

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Ding, L. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, *173*(2). https://doi.org/10.1016/j.cell.2018.02.060

Banta, L. M., Crespi, E. J., Nehm, R. H., Schwarz, J. A., Singer, S., Manduca, C. A., Bush, E. C., Collins, E., Constance, C. M., Dean, D., Esteban, D., Fox, S., McDaris, J., Paul, C. A., Quinan, G., Raley-Susman, K. M., Smith, M. L., Wallace, C. S., Withers, G. S., & Caporale, L. (2012). Integrating genomics research throughout the Undergraduate Curriculum: A Collection of inquiry-based Genomics Lab Modules. *CBE—Life Sciences Education*, *11*(3), 203–208. https://doi.org/10.1187/cbe.11-12-0105

Baumler, D. J., Banta, L. M., Hung, K. F., Schwarz, J. A., Cabot, E. L., Glasner, J. D., & Perna, N. T. (2012). Using Comparative Genomics for Inquiry-Based Learning to Dissect Virulence of *Escherichia coli* O157:H7 and *Yersinia pestis*. *CBE—Life Sciences Education*, *11*(1), 81–93. https://doi.org/10.1187/cbe.10-04-0057

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *74*(3), 229–263. https://doi.org/10.3322/caac.21834

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., Schrock, A., Campbell, B., Shlien, A., Chmielecki, J., Huang, F., He, Y., Sun, J., Tabori, U., Kennedy, M., Lieber, D. S., Roels, S., White, J., Otto, G. A., … Frampton, G. M. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, *9*(1). https://doi.org/10.1186/s13073-017-0424-2

Gerace, E. L., & Wojiski, S. (2025). Virtual cancer genomics: An accessible and effective approach to research training for undergraduates. *Journal of Cancer Education*, *40*(4), 633–639. https://doi.org/10.1007/s13187-025-02594-2

Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics*, *15*(9), 585–598. https://doi.org/10.1038/nrg3729

Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., Tsimberidou, A. M., Vnencak-Jones, C. L., Wolff, D. J., Younes, A., & Nikiforova, M. N. (2017). Standards and guidelines for the interpretation and reporting of sequence variants in cancer. *The Journal of Molecular Diagnostics*, *19*(1), 4–23. https://doi.org/10.1016/j.jmoldx.2016.10.002

Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, *349*(6255), 1483–1489. https://doi.org/10.1126/science.aab4082

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjærde, O. C., Langerød, A., Ringnér, M., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54. https://doi.org/10.1038/nature17676

Shaffer, C. D., Alvarez, C., Bailey, C., Barnard, D., Bhalla, S., Chandrasekaran, C., Chandrasekaran, V., Chung, H.-M., Dorer, D. R., Du, C., Eckdahl, T. T., Poet, J. L., Frohlich, D., Goodman, A. L., Gosser, Y., Hauser, C., Hoopes, L. L. M., Johnson, D., Jones, C. J., … Elgin, S. C. R. (2010). The Genomics Education Partnership: Successful integration of research into laboratory classes at a diverse group of undergraduate institutions. *CBE—Life Sciences Education*, *9*(1), 55–69. https://doi.org/10.1187/09-11-0087

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2018). Cosmic: The catalogue of somatic mutations in cancer. *Nucleic Acids Research*, *47*(D1). https://doi.org/10.1093/nar/gky1015

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature* 578, 82–93 (2020). https://doi.org/10.1038/s41586-020-1969-6

Verma, S., Yacob, M. S., & Kirpalani, A. (2022). Outcomes of inquiry-based learning in health professions education: a scoping review. *Canadian Medical Education Journal*, *14*(2). https://doi.org/10.36834/cmej.75144

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, *339*(6127), 1546–1558. https://doi.org/10.1126/science.1235122

Wang, L.-J., Ning, M., Nayak, T., Kasper, M. J., Monga, S. P., Huang, Y., Chen, Y., & Chiu, Y.-C. (2024). shinyDeepDR: A user-friendly R Shiny app for predicting anti-cancer drug response using deep learning. *Patterns*, *5*(2), 100894. https://doi.org/10.1016/j.patter.2023.100894

A Patient-Focused Interface for Optimizing Care During Long Emergency Department Wait Times

Tisya Pramanik, Seevitha Totapalli

Abstract

Overcrowding in emergency departments remains a chronic issue, where long wait times often leave patients feeling anxious and neglected. Without readily available staff to provide updates, many patients struggle to monitor their symptoms effectively, creating a clear gap in patient-centred care. This project explored whether a digital interface offering structured, non-diagnostic guidance could improve psychological reassurance and clarity for those waiting, all without encouraging risky self-diagnosis.

To improve this, mobile and computer-friendly data interfaces were created using Wix, following a design process that focused on conceptual modelling and iterative refinement. The interface featured symptom-category menus, evidence-based coping tips, and doctor-approved AI-driven descriptions, alongside automated prompts to notify hospital staff if symptoms worsened. We ran usability tests with 1000 volunteers in non-clinical settings, using 7-point Likert scales to measure user efficiency and satisfaction.

The results showed mean scores of 5/7 for navigation and clarity, while utility and trust both averaged 4/7. Qualitative feedback was that the participants felt less stressed and confused, and found that the AI-driven de-escalation prompts were highly reassuring. The overall findings indicate that thoughtfully designed digital tools can bridge the support gap during delays in the emergency department, offering patients comfort while ensuring clinical oversight remains the priority.

Emergency departments serve as critical entry points into healthcare systems by managing urgent and unpredictable medical needs. Regardless of its essential nature, Emergency department overcrowding has become a persistent national issue. Patients frequently experience prolonged waiting periods before assessment or treatment. (Richert & Jacobs, 2018) This creates stress, and potential health risks that can be avoided by implementing health care policies and strategies to address prolonged waiting periods. Scenario-based studies demonstrate that patients waiting without communication experience increased anxiety and reduced trust, especially as symptoms progress. (Kim et al, 2023) In Canada, the demand for emergency services is rapidly rising and is paired with workforce shortages, intensifying the crisis. Increasing patient volumes and limited inpatient capacity contribute to extended wait periods, reducing system effectiveness. Studies indicate that delays potentially lead to patients leaving before receiving care, or misunderstanding severity, worsening outcomes. (Darraj et al., 2023) Approximately 14% of emergency visits result in hospitalization. (Barrett et al., 2022) This highlights the clinical need for timely assessment. National healthcare analyses continue to show a consistent increase in Emergency Department wait times, with overcrowding linked to poor patient satisfaction and delayed treatment initiation. Alongside, the worsening symptom severity increases operational costs and staff burden. (Savioli et al., 2022)

Several approaches have been attempted to address wait times in hospitals, including lean workflow optimization, triage restructuring, and digital queue management. These management models streamline patient flow; however, they primarily target institutional processing rather than patient experience. While operational improvements play a key role in shortening wait times, patients often remain unsupported during the delays. Educational brochures and informational displays are often present in waiting areas. However, due to their static and

generalized nature, they present limitations as the diagnostic suggestions increase anxiety and encourage inappropriate self-assessment. Current systems largely overlook patient needs, creating a gap between operational efficiency solutions and patient-centred tools. Our proposed model addresses this gap by providing structured, non-diagnostic support, emphasizing reassurance, safety awareness, and clear escalation guidance. By avoiding diagnostic claims and implementing proper ethical safeguards, this interface aims to enhance patient autonomy, while remaining clinically responsible.

Hypothesis

It was hypothesized that implementation of a patient-centered digital interface providing structured, non-diagnostic symptom guidance would significantly improve users' perceived clarity and psychological reassurance during emergency department wait times, without increasing reliance on self-diagnosis or reducing perceived necessity for professional medical evaluation.

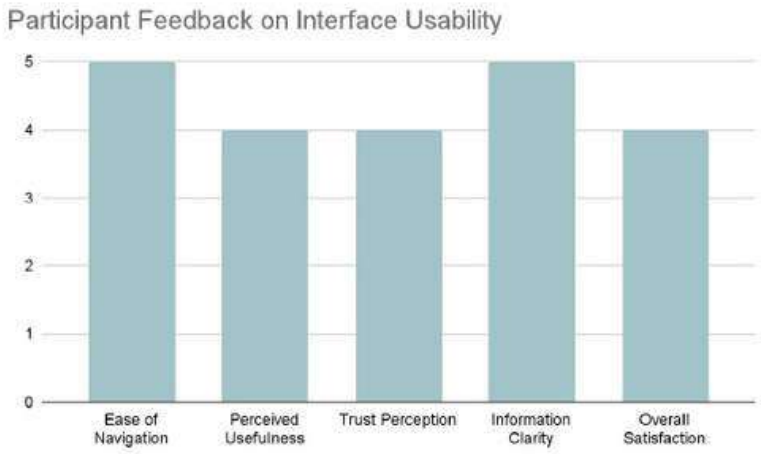
Methods

The prototype application and accompanying website were developed using Wix Sites, and Wix App Builder. The methodology used to develop this app centered around users, emphasizing accessibility, simplicity, and risk mitigation, especially during high-stress environments. The development followed three stages: conceptual modeling, structural implementation, and refinement. During conceptual modelling, a literature review was conducted to understand the root causes of patient anxiety during emergency wait times. The app and website prototype will address anxiety related to worsening symptoms, and accurate symptom management. During structural implementation, the interface design was created using Wix. Core elements of the app

diagnose. For improved usability purposes, language preferences, patient status summary and visuals were also embedded. In the refinement stage, we conducted pilot usability testing with 1000 volunteers in non-clinical settings to navigate the app, and provide insight on its ease of navigation, information clarity, perceived usefulness, trust perception and overall satisfaction. The Likert Scale from 1 to 7 was used to assess each section. Qualitative written feedback was collected to identify strengths and limitations.

Figure 1: Prototype of proposed website questionnaire consisting of patient data summary, symptom change surveys, language preferences, physician-approved AI chatbot, and emergency button, for life-threatening situations.

Results



Graph 1: Summary of feedback from 1000 participants, suggesting highest ease of navigation and information clarity in the digital platform. Overall, very positive feedback with the website.

Participants demonstrated rapid adaptation to interface navigation. This outlines that the usability is intuitive. A Likert rating of 5 indicates high clarity ratings, with minimal confusion when selecting symptoms and monitoring progression. Perceived usefulness scores averaged at 4. Participants indicated appreciation for escalation prompts, which were identified as "reassuring" and "not alarming." Qualitative feedback revealed recurring accounts of reassurance and organization in intense, waiting scenarios. Several volunteers reported structured steps that reduced uncertainty about appropriate actions while waiting. Minor challenges arose in participants requested a wait time estimation feature in the emergency department.

Discussion

The present study tested the usability and perceived impact of a digital interface offering support to patients during hospital wait times. Overall, the findings support the hypothesis that patient-centred digital tools meaningfully enhance patient support during prolonged emergency

patient-centred digital tools meaningfully enhance patient support during prolonged emergency department wait periods. The results highlight the role of ethically designed digital health tools. Strengths of this project lie in its system-level positioning. Prior interventions focused on workflow redesign. This interface addresses experiential differences, specifically informational gaps during delays. By targeting patient cognition and perception, the intervention complements existing strategies. Additionally, the design approach enhanced accessibility and minimized cognitive load. The approach steps towards combatting decision fatigue and misinterpretation risk. Persistent disclaimers controlled the informational environment. Several limitations can be acknowledged. Firstly, the pilot sample size was small and limited demographically, restricting its generalizability. The absence of a control group prevented definitive conclusions on interface usage and impact. Usability testing occurred outside the emergency environment. The real-world setting would involve noise, heightened emotions, physical discomfort, and time pressure. These contextual variables can influence navigation behaviour. Lastly, long-term behavioural effects were not assessed. It remains unknown whether repeated exposure would change trust dynamics and reliance patterns over time. This interface contributes to the evolving field of patient engagement technologies. This app extends beyond the emergency department and can be used in any setting requiring medical emergency. Future research should include controlled trials measuring objective outcomes. Longitudinal assessment is also warranted to examine sustained behavioural effects and unintended consequences, such as delayed help-seeking.

Conclusion

This study demonstrates the usability of patient centered digital interfaces designed to support individuals during prolonged wait times. The prototype improved perceived clarity, reassurance, and usability without compromising safety or encouraging diagnostic behavior. As healthcare systems increasingly implement digital engagement tools, patient support interfaces may stand as an important complement to traditional care. Future scaling efforts could expand accessibility, integrate multilingual support, and connect with institutional systems to enhance real-time public impact.

References

Reichert, A., & Jacobs, R. (2018). The impact of waiting time on patient outcomes: Evidence from early intervention in psychosis services in England. *Health Economics*, *27*(11), 1772–1787. https://doi.org/10.1002/hec.3800

Kim, S., Chang, H., Kim, T., & Cha, W. C. (2023). Patient Anxiety and Communication experience in the Emergency Department: A Mobile, Web-Based, Mixed-Methods study on patient isolation during the COVID-19 pandemic. *Journal of Korean Medical Science*, *38*(39), e303. https://doi.org/10.3346/jkms.2023.38.e303

Darraj, A., Hudays, A., Hazazi, A., Hobani, A., & Alghamdi, A. (2023). The Association between Emergency Department Overcrowding and Delay in Treatment: A Systematic Review. *Healthcare*, *11*(3), 385. https://doi.org/10.3390/healthcare11030385

Barrett, M. L., Owens, P. L., & Roemer, M. (2022, October 18). *Changes in emergency department visits in the initial period of the COVID-19 pandemic (April–December 2020), 29 states*. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK586682/

Savioli, G., Ceresa, I. F., Gri, N., Piccini, G. B., Longhitano, Y., Zanza, C., Piccioni, A., Esposito, C., Ricevuti, G., & Bressan, M. A. (2022). Emergency department overcrowding: Understanding the factors to find corresponding solutions. *Journal of Personalized Medicine*, *12*(2), 279. https://doi.org/10.3390/jpm12020279

Ethical Integrity Statement

We confirm that this submission is our original work completed without the use of AI tools , that all sources are properly cited in APA 7 format, and that we understand BACSA is not responsible for any plagiarism or academic misconduct

Cold-Weather Ammonia Removal and Downstream Ecological Risk in Ontario Wastewater Systems: A BioCord-Informed Assessment Framework

Yi Luo, Jincheng Yang, Shiqi Lan, Ziang Chen

Introduction

The health and sustainability of Ontario's freshwater ecosystems are fundamentally linked to the efficacy of municipal wastewater treatment infrastructure. A long-term study in the Grand River watershed has demonstrated that ammonia in municipal wastewater effluent poses a direct threat to aquatic ecosystem health. Nikel et al. (2023) reported that elevated ammonia concentrations (25–26 mg/L) downstream of the Waterloo wastewater treatment plant were associated with significant physiological disruption in fish populations, while reductions in ammonia concentrations corresponded with recovery of normal biological function. These findings highlight the importance of effective ammonia removal in protecting downstream ecosystems.

Ammonia removal during cold weather presents a particular challenge due to the temperature sensitivity of nitrifying microorganisms. Alawi et al. (2007) identified cold-adapted nitrite-oxidizing bacteria capable of maintaining activity at temperatures as low as 4°C, suggesting that microbial community composition plays a critical role in winter treatment performance. This is supported by Skoyles (2019), who observed a seasonal shift toward cold-adapted nitrifying bacteria in BioCord-based lagoon systems in Ontario, indicating that attached-growth biofilm systems may help sustain nitrification under low-temperature conditions.

BioCord technology, consisting of high-surface-area polymer media that support biofilm growth, has demonstrated strong ammonia removal performance under controlled conditions. Tian et al. (2019) reported removal efficiencies of 92–97% under moderate loading rates, while Yuan et al. (2012) showed that BioCord biofilms support diverse microbial communities capable of simultaneous nitrification and denitrification. These

findings suggest that BioCord systems have the potential to enhance treatment performance, particularly in cold climates.

Despite these advances, important gaps remain in understanding how BioCord systems perform under real operational conditions. Existing studies are largely experimental or descriptive and do not provide quantitative models linking environmental variables to ammonia concentrations in full-scale systems. Additionally, operational monitoring data collected from wastewater facilities are often underutilized for statistical analysis and predictive assessment.

This study addresses these gaps by integrating operational monitoring data and literature-derived datasets to evaluate ammonia removal performance in BioCord wastewater treatment systems. Specifically, this study aims to (1) evaluate ammonia removal efficiency under varying loading conditions using bench-scale data, and (2) develop a regression-based model relating ammonia concentration to environmental variables including temperature, pH, and biochemical oxygen demand in a full-scale lagoon system. By linking environmental conditions to treatment performance, this study provides a quantitative framework for understanding ammonia removal dynamics in cold-weather wastewater systems.

Hypothesis and Objective

The objective of this study is to evaluate ammonia removal performance in BioCord wastewater treatment systems and to identify key environmental and operational factors influencing treatment efficiency under cold-weather conditions.

We hypothesize that:

- Lower water temperatures are associated with reduced ammonia removal efficiency.
- Environmental variables including temperature, pH, and biochemical oxygen demand (BOD) influence ammonia concentrations in BioCord-enhanced lagoon systems.
- Increasing surface area loading rate (SALR) reduces ammonia removal efficiency while increasing total ammonia removal capacity.

Method

Data Base:

Water quality monitoring data were retrieved from the Dundalk Wastewater Lagoon BioCord Performance Dataset as reported in Skoyles (2019), a technical thesis evaluating the performance of BioCord media in the Dundalk, Ontario lagoon treatment system.

Bench-scale ammonia removal performance data were extracted from Skoyles (2019), Microbial Community Dynamics of Attached Biofilm BioCord Technology in Wastewater Treatment (University of Windsor, Canada). Specifically, influent and effluent ammonia concentrations were digitized from Figure 2.4 (Page 25), and ammonia removal efficiency values were obtained from Page 19. The experiments were conducted at 20°C with stable pH (7.8–8.4) under increasing influent ammonia concentrations (0.07, 4.01, 10.3, and 28.2 mg NH_3-N/L).

Surface area loading rate (SALR), surface area removal rate (SARR), and removal efficiency data were obtained from Tian et al. (2017), Nitrifying bio-cord reactor: performance optimization and effects of substratum and air scouring, Environmental Technology. Data were extracted from Table 1, which reports performance metrics under varying surface area loading rates.

Data inclusion criteria consisted of all steady-state measurements reported under the specified loading conditions. No filtering or transformation was applied beyond unit standardization. A total of n = 12 observations were used in regression analysis of SALR versus removal efficiency.

All sources were accessed February 2026 via institutional library databases.

Software & Tools:

All statistical analyses were conducted using Python (version 3.11). Data processing and regression modeling were performed using the following libraries:

- NumPy (numerical computation)
- pandas (data organization and manipulation)
- statsmodels (ordinary least squares regression)
- matplotlib (data visualization)

All analyses were executed in a local virtual environment (.venv) using Python's OLS implementation in statsmodels.

Analytical Procedures

Raw data were manually extracted from published tables and figures and entered into structured data frames using pandas. Units were verified for consistency (mg NH_3-N/L and g NH_4^+-N/m^2·d). No imputation or outlier removal was performed.

For bench-scale influent–effluent analysis, removal efficiency was calculated as:

$$\text{Removal}\ (\%) = \frac{C_{in} - C_{out}}{C_{in}} \times 100$$

Ammonia removal capacity was calculated as:

$$\text{Removed}(\text{mg}/L) = \frac{C_{in} - C_{out}}{}$$

For loading-based analysis, quadratic regression modeling was performed to evaluate the non-linear relationship between surface area loading rate (SALR) and ammonia removal efficiency:

$$\text{Removal} = \beta_0 + \beta_1 \cdot SALR + \beta_2 \cdot SALR^2$$

Model parameters were estimated using ordinary least squares (OLS). Model significance was assessed using F-tests and t-tests. Statistical significance was defined as $p<0.05p < 0.05p<0.05$.

No cross-validation was performed due to limited sample size (n = 12).

Result

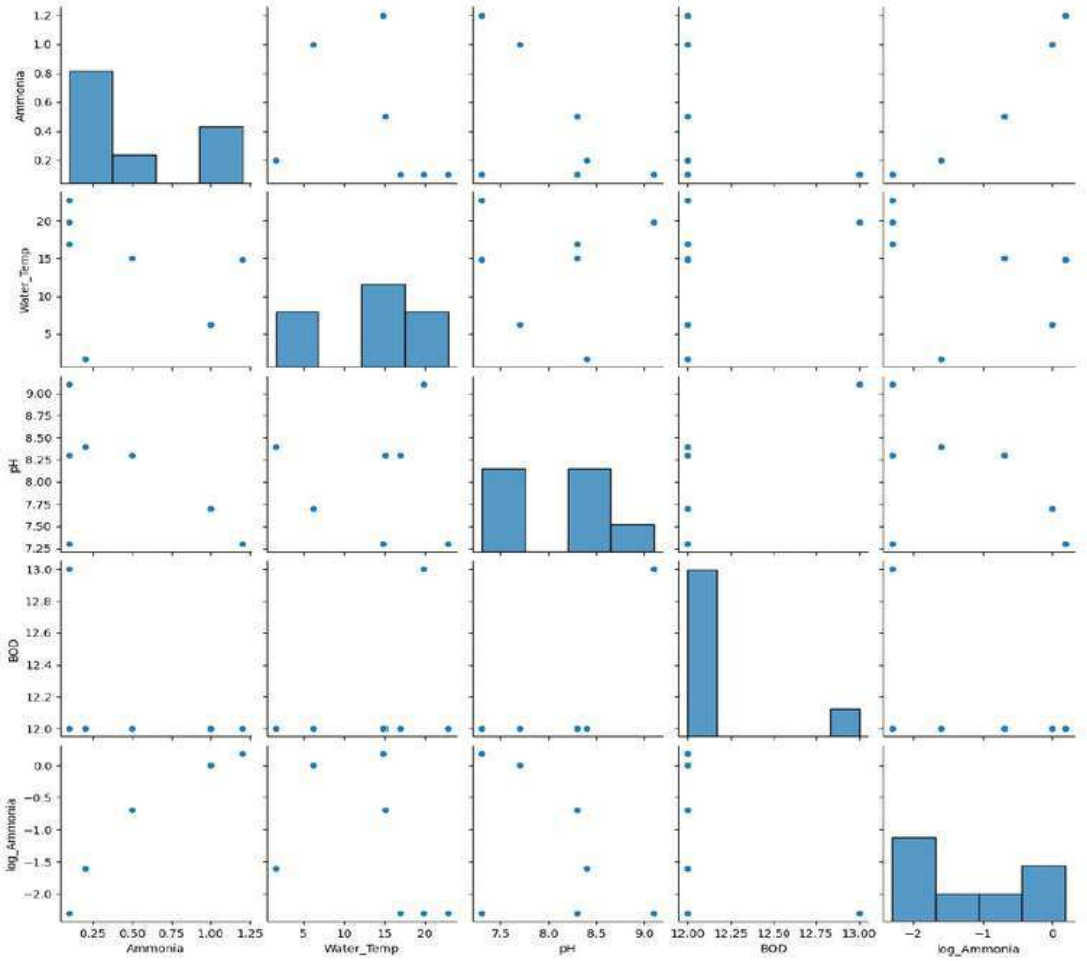


Figure 1. Pairwise Scatter Plot Matrix of Water Quality Variables Under BioCord-Enhanced Lagoon Conditions

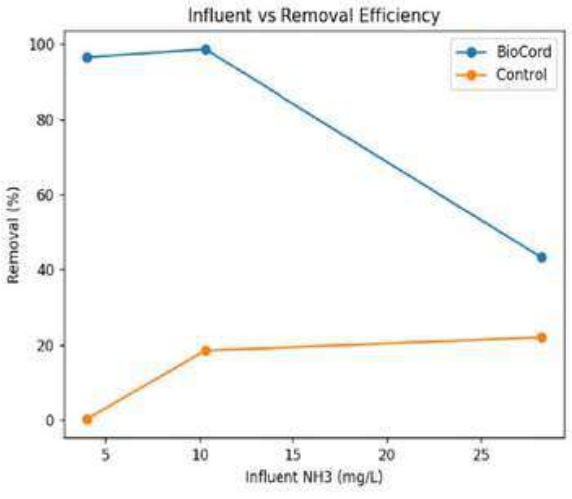


Figure 2

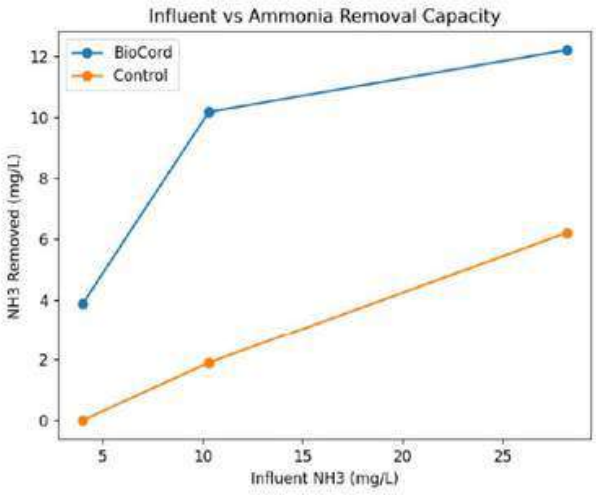


Figure 3

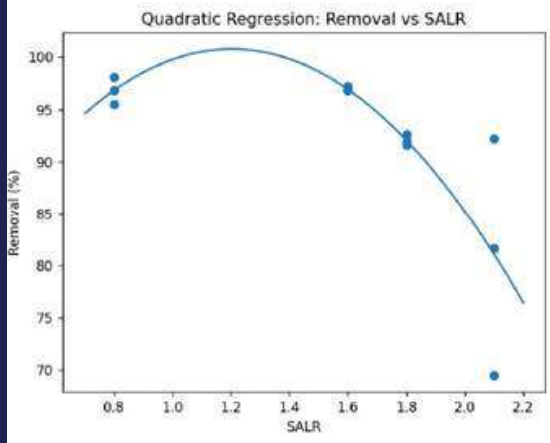


Figure 4

OLS Regression Results

Dependent Variable: y
No. Observations: 12
R-squared: 0.655
Adjusted R-squared: 0.578
F-statistic: 8.545
Prob (F-statistic): 0.00832

		Std. Error	t	p-value	95% CI
Intercept	65.382	16.327	4.005	0.003	[28.449, 102.315]
x1	58.839	25.428	2.314	0.046	[1.317, 116.360]
x2	-24.456	8.955	-2.731	0.023	[-44.713, -4.198]

Model interpretation:
- x1 has a positive and statistically significant association with y (p = 0.046).
- x2 has a negative and statistically significant association with y (p = 0.023).
- The model explains 65.5% of the variance in y.

Figure 5

A log-linear ordinary least squares (OLS) regression model was specified as:

$$\textbf{Ammonia} = 0.274 \cdot e^{(-0.0878T)} \cdot e^{(-1.2234pH)} \cdot e^{(0.9124BOD)}.$$

To evaluate the non-linear relationship between surface area loading rate (SALR) and ammonia removal efficiency, a quadratic regression model of the form:

$$Removal = 65.38 + 58.84 \cdot SALR - 24.46 \cdot SALR^{12}$$

1. The linear regression model was used to evaluate the effects of water temperature, pH, and BOD on ammonia concentration. Data were obtained from the Dundalk Wastewater Lagoon BioCord performance dataset reported by Skoyles (2019).The results showed that: Water temperature had a negative coefficient ($\beta = -0.0878$). pH also had a negative coefficient ($\beta = -1.2234$), suggesting a decrease in ammonia concentration with increasing pH. BOD had a positive coefficient ($\beta = +0.9124$), indicating that ammonia concentration increased with higher BOD levels. The overall model had an R^2 value of 0.481, but none of the predictors were statistically significant ($p > 0.05$)(Skoyles, 2019).(figure 1)

2. Ammonia removal capacity: Ammonia removal capacity increased with influent concentration. BioCord removed approximately 3.9 mg/L at 4 mg/L influent; 10.2 mg/L at 10 mg/L influent; 12.2 mg/L at 28 mg/L influent. In comparison, the control removed significantly less ammonia 0 mg/L; 1.9 mg/L; 6.2 mg/L. This demonstrates the superior ammonia removal capacity of the BioCord system.(figure 2)

3. Surface area ammonia removal rate (SARR): Surface area ammonia removal rate increased with increasing surface area loading rate. SARR increased from approximately 0.79 g NH_4-N/m^2·d to 1.72 g NH_4-N/m^2·d.

However, the rate of increase slowed at higher loading levels, suggesting the system approached its maximum removal capacity.(figure 3)

4. Regression between SALR and removal efficiency: Regression analysis showed a negative relationship between surface area loading rate and ammonia removal efficiency. Removal efficiency decreased from approximately 98% to approximately 70–90% as SALR increased. The regression model showed:$R^2 = 0.369$, $p = 0.036$, indicating a statistically significant relationship. This suggests that increasing loading rates reduce removal efficiency.(figure 4)

Discussion

1. Interpretation of key findings

The results demonstrate that the BioCord system effectively enhances ammonia removal performance. Removal efficiency was highest at lower surface area loading rates (SALR), while efficiency declined as SALR increased. In contrast, total ammonia removal capacity and removal rate increased with increasing SALR. This pattern indicates that the system was able to process more ammonia at higher loading rates, but its efficiency declined due to biological and physical limitations. The quadratic regression model further confirmed this relationship, showing a statistically significant non-linear association between SALR and removal efficiency ($R^2 = 0.655$, $p = 0.008$). The negative quadratic coefficient suggests that removal efficiency reaches an optimal range and declines at higher loading levels. This indicates that the BioCord system has a finite biological processing capacity, beyond which efficiency decreases.

2. Literature comparison

The results of this study are consistent with previous research on BioCord technology. BioCord has been shown to achieve ammonia removal efficiency up to approximately 98%, significantly higher than control systems, due to the activity of attached nitrifying biofilm communities (Skoyles, 2019). Key nitrifying bacteria such as Nitrosomonas, Nitrospira, and Candidatus Nitrotoga have been identified as major contributors to ammonia oxidation in BioCord systems (Skoyles, 2019). The high removal efficiency observed in this study supports these findings and confirms the effectiveness of BioCord in enhancing ammonia removal. Additionally, previous studies have shown that reducing ammonia discharge from wastewater treatment systems can improve aquatic ecosystem health. (Nikel, Kirsten E., Gerald R. Tetreault, Patricija Marjan, Keegan A. Hicks, Meghan L. M. Fuzzen, Nivetha

Srikanthan, Emily K. McCann, et al. 2023). These findings support the potential application of BioCord technology in wastewater treatment.

3. Biological interpretation

Ammonia removal in the BioCord system is primarily driven by nitrifying bacteria attached to the biofilm surface. At lower loading rates, ammonia supply remains within the processing capacity of the microbial community, allowing efficient conversion. However, at higher loading rates, several limitations may occur, including oxygen diffusion limitation, substrate overload and biofilm saturation. These limitations reduce the proportion of ammonia that can be processed efficiently. As a result, removal efficiency declines even though the total amount of ammonia removed increases. This reflects the biological capacity limits of the system.

4. Environmental and practical implications

Ammonia pollution poses significant risks to aquatic ecosystems and wastewater treatment systems. The strong ammonia removal performance observed in the BioCord system demonstrates its potential as an effective treatment technology. The quadratic relationship between loading rate and removal efficiency suggests that system performance depends strongly on operational conditions. Operating the system within an optimal loading range may maximize efficiency while maintaining high treatment capacity. This finding has practical importance for system design and operation.

5. Model implications

The regression models provided useful insights into factors affecting ammonia removal performance. Environmental variables explained approximately 48% of ammonia concentration variation, indicating that environmental conditions play an important role. The quadratic SALR model explained approximately 65.5% of removal efficiency variation, suggesting that loading rate is a major determinant of system performance. However, some variation remains unexplained, indicating that additional factors such as microbial activity, oxygen availability, and biofilm structure may also influence performance. This highlights the complexity of biological treatment systems.

6. Limitations

This study has several limitations. First, the sample size was relatively small, particularly for the environmental regression model (n = 7), which reduced statistical power and limited the significance of some predictors. Second,

the quadratic regression model included only SALR as a predictor. Other environmental and biological variables were not included and may also affect removal efficiency. Third, the observed data covered a limited range of environmental and loading conditions. Therefore, model predictions outside this range may not be reliable. Fourth, regression analysis identifies statistical relationships but does not confirm causation. Finally, the use of secondary data sources may introduce variability due to differences in measurement conditions. These limitations should be considered when interpreting the results.

7. Future improvements

Future studies should expand the dataset to improve statistical reliability and model robustness. Increasing the number of observations across a wider range of surface area loading rates (SALR) would help better define the optimal loading range and reduce uncertainty in regression estimates. Controlled laboratory experiments should also be conducted to isolate the effects of individual variables such as temperature, pH, and dissolved oxygen. This would allow more precise identification of the environmental factors influencing ammonia removal performance. In addition, future studies should include additional environmental and operational variables, such as dissolved oxygen concentration, hydraulic retention time, and influent ammonia concentration, to improve the explanatory power of regression models.

Conclusion

This study evaluated ammonia removal performance in the BioCord wastewater treatment system using regression analysis and literature-derived performance data. The results showed that BioCord achieved substantially higher ammonia removal efficiency compared to the control lagoon across a range of influent concentrations. Removal efficiency was highest at lower loading levels and declined at higher loading rates, while overall removal capacity increased. Regression analysis indicated that environmental variables such as temperature, pH, and BOD influenced ammonia concentration, although statistical significance was limited by the small sample size. Quadratic regression analysis further demonstrated a significant non-linear relationship between loading rate and removal efficiency, suggesting the presence of an optimal loading range for maximum performance. These findings demonstrate that BioCord is an effective biological treatment technology for ammonia removal and highlight the importance of operational conditions in determining treatment efficiency. This study contributes to the understanding of BioCord performance and supports its potential application in wastewater treatment systems. Future research with larger datasets and controlled experiments will help improve predictive accuracy and optimize system design.

Reference

Alawi, M., A. Lipski, T. Sanders, and E. Spieck. 2007. "Cultivation of a Novel Cold-Adapted Nitrite-Oxidizing Betaproteobacterium from the Siberian Arctic." *The ISME Journal* 1: 256–264. https://doi.org/10.1038/ismej.2007.34.

Fuller, Megan, Emma Wells, Laith Furatianc, Ian Douglas, and Kaycie Lane. 2023. "Drinking Water Quality Management Progress in Ontario, Two Decades after Walkerton." *Journal of Water and Health* 21 (8): 1073–1090. https://doi.org/10.2166/wh.2023.099.

Nikel, Kirsten E., Gerald R. Tetreault, Patricija Marjan, Keegan A. Hicks, Meghan L. M. Fuzzen, Nivetha Srikanthan, Emily K. McCann, et al. 2023. "Wild Fish Responses to Wastewater Treatment Plant Upgrades in the Grand River, Ontario." *Aquatic Toxicology* 255: 106375. https://doi.org/10.1016/j.aquatox.2022.106375.

Skoyles, Adam. 2019. *Microbial Community Dynamics of Attached Biofilm BioCord Technology in Wastewater Treatment*. PhD diss., University of Windsor. https://www.proquest.com/docview/2273314277.

Tian, Xiaolin, Waheed Ahmed, and Reza Delatolla. 2017. "Nitrifying Bio-Cord Reactor: Performance Optimization and Effects of Substratum and Air Scouring." *Environmental Technology*.

Yuan, Xiaoyan, Xueling Qian, Rui Zhang, Rui Ye, and Wei Hu. 2012. "Performance and Microbial Community Analysis of a Novel Bio-Cord Carrier during Treatment of a Polluted River." *Bioresource Technology* 117: 33–39. https://doi.org/10.1016/j.biortech.2012.04.058.

Ethical Integrity Statement

We confirm that this submission is our original work completed without the use of AI tools, that all sources are properly cited in APA 7 format, and that we understand BACSA is not responsible for any plagiarism or academic misconduct.

Comparative Analysis of CRISPR-Cas9 Strategies in Wheat and Yeast for Gluten Detoxification

Joelle Weir, Dhritya Nair, Basim Mahmood Usmani, Asmaa Gaal

water-insoluble proteins are called glutenins (Mesta-Corral et al., 2024). Gliadin and glutenin, together, trap gas within the bread dough to enhance the viscosity and elasticity of the dough (Ting et al., 2025). However, they are also the gluten protein strands (peptides) that cause the inflammatory autoimmune reaction within gluten intolerant individuals (*Gliadin - an Overview | ScienceDirect Topics*, n.d.). Gliadin is rich in amino acids, glutamines and prolines, which compromise the digestibility of gluten in the intestines. This gives rise to indigestible peptides, mimicking harmful microorganisms, causing the autoimmune reaction. The biological mechanism allows for increased permeability in tight junctions, the barrier between cell layers; this is caused by two-alpha-gliadin motifs and zonulin, a molecule released by gliadin. After breaking through the tight junction complex, crossing the intestinal barrier, gluten stimulates the counteractive expression of the transferring receptor CD71. This leads to retrotranscytosis of IgA peptides, attached with gliadin peptides creating the IgA-gliadin complex. This complex protects gliadin particles from lysosomal degradation and allows entry into the intestinal lamina propria, causing inflammation in the intestines. The gluten immunogenic peptides continue further into the bloodstream, perpetuating inflammation until it is finally excreted in urine (Caio et al., 2019).

The gluten composition within daily consumed bread makes it difficult for gluten-free individuals to find alternatives and protect their immune systems from gluten particles. A method to remove gluten from bread includes using CRISPR-Cas9 to genetically engineer bread yeast to provide a healthy, reliable, gluten-free solution to individuals with celiac disease, gluten tolerance, irritable bowel syndrome, etc. This research paper focuses on *Saccharomyces cerevisiae*, an industry-wide standard for Baker's yeast, and *Triticum aestivum*, Hexaploid wheat naturally occurring offspring of macaroni wheat and wild goat grass (Aberkane et al., 2020; Liang et al., 2024). *S. cerevisiae* secretes neprosin, an enzyme that neutralizes the harmful, inflammatory components of gluten. Recombinant neprosin degrades gliadin particles in wheat flour, and further digest all immunogenic epitopes in alpha-gliadin. This was tested in-vitro and has shown promising results in creating gluten-free, safe food options. While this research provides the groundwork for a revolution of realizable, gluten-safe options for commercialization. Gluten detoxification with neprosin impacts the rheology, elasticity, and durability of the baked bread (Ting et al., 2025). This limitation suggests further research in utilizing CRISPR-Cas9 and neprosin in various approaches to explore the best method of producing uncompromised gluten-free bread. In pursuit of the best method, this research paper questions the optimal methodology utilizing CRISPR-Cas9 to cleave gene segments encoding for immunodominant epitopes, while maintaining viscoelasticity.

water-insoluble proteins are called glutenins (Mesta-Corral et al., 2024). Gliadin and glutenin, together, trap gas within the bread dough to enhance the viscosity and elasticity of the dough (Ting et al., 2025). However, they are also the gluten protein strands (peptides) that cause the inflammatory autoimmune reaction within gluten intolerant individuals (*Gliadin - an Overview | ScienceDirect Topics*, n.d.). Gliadin is rich in amino acids, glutamines and prolines, which compromise the digestibility of gluten in the intestines. This gives rise to indigestible peptides, mimicking harmful microorganisms, causing the autoimmune reaction. The biological mechanism allows for increased permeability in tight junctions, the barrier between cell layers; this is caused by two-alpha-gliadin motifs and zonulin, a molecule released by gliadin. After breaking through the tight junction complex, crossing the intestinal barrier, gluten stimulates the counteractive expression of the transferring receptor CD71. This leads to retrotranscytosis of IgA peptides, attached with gliadin peptides creating the IgA-gliadin complex. This complex protects gliadin particles from lysosomal degradation and allows entry into the intestinal lamina propria, causing inflammation in the intestines. The gluten immunogenic peptides continue further into the bloodstream, perpetuating inflammation until it is finally excreted in urine (Caio et al., 2019).

The gluten composition within daily consumed bread makes it difficult for gluten-free individuals to find alternatives and protect their immune systems from gluten particles. A method to remove gluten from bread includes using CRISPR-Cas9 to genetically engineer bread yeast to provide a healthy, reliable, gluten-free solution to individuals with celiac disease, gluten tolerance, irritable bowel syndrome, etc. This research paper focuses on *Saccharomyces cerevisiae*, an industry-wide standard for Baker's yeast, and *Triticum aestivum*, Hexaploid wheat naturally occurring offspring of macaroni wheat and wild goat grass (Aberkane et al., 2020; Liang et al., 2024). *S. cerevisiae* secretes neprosin, an enzyme that neutralizes the harmful, inflammatory components of gluten. Recombinant neprosin degrades gliadin particles in wheat flour, and further digest all immunogenic epitopes in alpha-gliadin. This was tested in-vitro and has shown promising results in creating gluten-free, safe food options. While this research provides the groundwork for a revolution of realizable, gluten-safe options for commercialization. Gluten detoxification with neprosin impacts the rheology, elasticity, and durability of the baked bread (Ting et al., 2025). This limitation suggests further research in utilizing CRISPR-Cas9 and neprosin in various approaches to explore the best method of producing uncompromised gluten-free bread. In pursuit of the best method, this research paper questions the optimal methodology utilizing CRISPR-Cas9 to cleave gene segments encoding for immunodominant epitopes, while maintaining viscoelasticity.

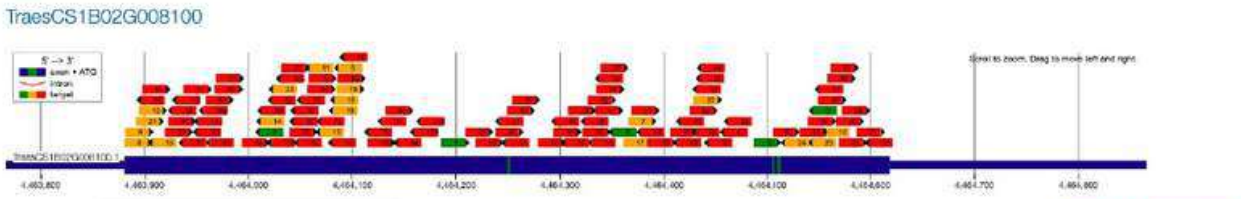
Hypothesis

The null hypothesis is that CRISPR-Cas 9 on yeast and CRISPR-Cas 9 on wheat would both effectively eliminate immunodominant epitopes through the tTG and PEP pathways while maintaining dough viscoelasticity. The alternative hypothesis CRISPR - Cas9 on yeast most effectively eliminates immunodominant epitopes through the tTG and PEP pathways while maintaining dough viscoelasticity.

Methods

The gene IDs for both *Triticum aestivum* and *Saccharomyces cerevisiae* can be found in *CHOPCHOP*. *CHOPCHOP* is a computational biology platform that matches coding sequence made of alphabetic nucleotides (A,T,C,G) to represent the chemical subunits that make up DNA with gene IDs known colloquially amongst researchers to efficiently inquire for selecting and designing target sites for CRISPR/Cas9.

The process begins with designing a specialized single-guide RNA (sgRNA) targeting the γ-gliadin genes in wheat, a family of genes responsible for immunoreactive gluten proteins. This sgRNA is designed using bioinformatics tools like CHOPCHOP to ensure specificity and minimize off-target effects (Sánchez-León, 2024).

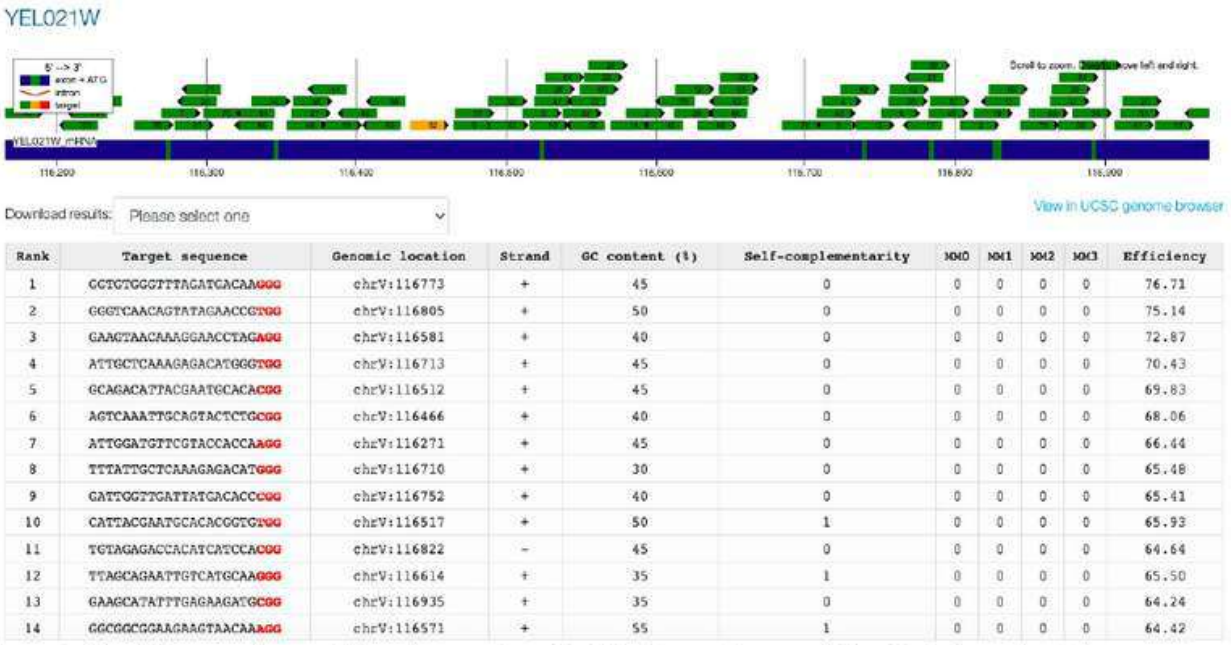


Download results: Please select one

View in UCSC genome browser

Rank	Target sequence	Genomic location	Strand	GC content (%)	Self-complementarity	MM0	MM1	MM2	MM3	Efficiency
1	GTACGTGTCTGAGATACTAGTGG	chr1B:4464187	+	45	0	0	0	0	0	68.58
2	TGGTTTGTGGGGATGCTCCCGGG	chr1B:4464012	-	60	1	0	0	0	0	63.57
3	GAAGTAGATCTGCGTCTCGATGG	chr1B:4464488	+	50	0	0	0	0	0	61.25
4	TGCGAGCGACGACCGTCCCCCGG	chr1B:4464352	-	75	3	0	0	0	0	62.39
5	AACTTCCTTGGTACCGGGTTTGG	chr1B:4464546	-	50	0	0	0	0	0	22.66
6	ATGTACGCGTGTGATTCCGACGG	chr1B:4463802	+	50	0	0	0	0	1	67.65
7	CGCTCGCACTGTCCTAGGCCAGG	chr1B:4464367	+	70	1	0	0	0	1	59.39
8	TGTACGCGTGTGATTCCGACGGG	chr1B:4463803	+	55	0	0	0	0	1	57.53
9	ACCCGTCTGCAGAGATGGCTCGG	chr1B:4464088	-	60	1	0	0	0	1	56.02
10	TTCCGACGGGCAGCCTCTGGTGG	chr1B:4463896	+	70	3	0	0	0	1	53.77
11	GCCCGTCAGCGGGAGCTTGCCGG	chr1B:4464060	-	75	0	0	0	0	1	47.00
12	AGCGAGGCCACGAACTTCCTTGG	chr1B:4464558	-	60	1	0	0	0	1	46.18
13	CTCGGTAGGGGCCCGTCAGCGGG	chr1B:4464070	-	75	2	0	0	0	1	41.35
14	CTGGTTTGTGGGGATGCTCCGGG	chr1B:4464013	-	60	1	0	0	0	1	39.77

Figure 1: Target Sequence to remove γ-Gliadin from Chromosome 1B *Triticum aestivum*.



Rank	Target sequence	Genomic location	Strand	GC content (%)	Self-complementarity	MM0	MM1	MM2	MM3	Efficiency
1	GGTGTGGGTTTAGATGACAAGGG	chrV:116773	+	45	0	0	0	0	0	76.71
2	GGGTCAACAGTATAGAACCGTGG	chrV:116805	+	50	0	0	0	0	0	75.14
3	GAAGTAACAAAGGAACCTAGAGG	chrV:116581	+	40	0	0	0	0	0	72.87
4	ATTGCTCAAAGAGACATGGGTGG	chrV:116713	+	45	0	0	0	0	0	70.43
5	GCAGACATTACGAATGCACACGG	chrV:116512	+	45	0	0	0	0	0	69.83
6	AGTCAAATTGCAGTACTCTGCGG	chrV:116466	+	40	0	0	0	0	0	68.06
7	ATTGGATGTTCGTACCACCAAGG	chrV:116271	+	45	0	0	0	0	0	66.44
8	TTTATTGCTCAAAGAGACATGGG	chrV:116710	+	30	0	0	0	0	0	65.48
9	GATTGGTTGATTATGACACCCGG	chrV:116752	+	40	0	0	0	0	0	65.41
10	CATTACGAATGCACACGGTGTGG	chrV:116517	+	50	1	0	0	0	0	65.93
11	TGTAGAGACCACATCATCCACGG	chrV:116822	-	45	0	0	0	0	0	64.64
12	TTAGCAGAATTGTCATGCAAGGG	chrV:116614	+	35	1	0	0	0	0	65.50
13	GAAGCATATTTGAGAAGATGCGG	chrV:116935	+	35	0	0	0	0	0	64.24
14	GGCGGCGGAAGAAGTAACAAAGG	chrV:116571	+	55	1	0	0	0	0	64.42

Figure 2: Target Sequence to insert Neprosin, a protease identified in carnivorous pitcher plants (source) , into Saccharomyces cerevisiae cleaving gluten peptides upon fermentation.

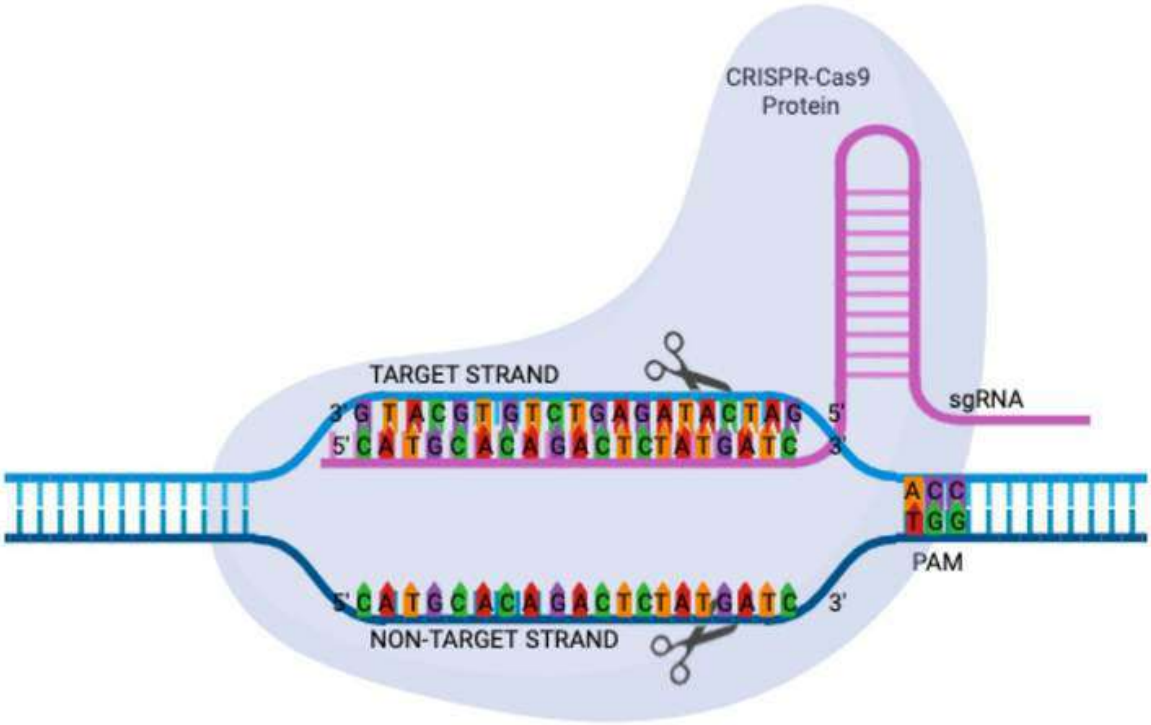


Figure 3: ' The Aligning Process': The Target sequence, in Triticum aestivum, from CHOPCHOP 'GTACGTGTCTGAGATACTAGTGG' has a 23 base pair including the PAM 3 base pair sequence 'TGG'. **The sgRNA binds Cas9 and directs it to the complementary DNA sequence adjacent to a PAM. Upon recognition, Cas9 undergoes a conformational change activating its HNH and RuvC nuclease domains, which cleave the target and non-target strands, respectively. The resulting double-strand break (Jiang & Doudna, 2017).**

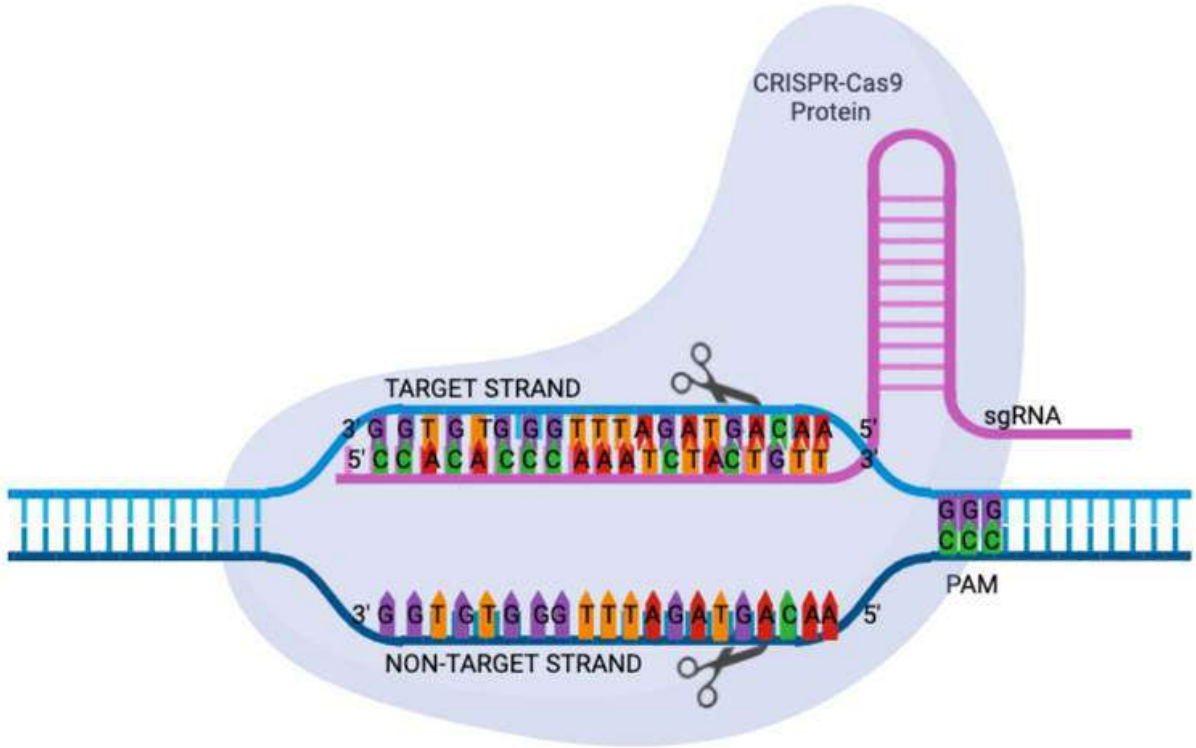


Figure 4: ' The Aligning Process': The Target sequence, in Saccharomyces cerevisiae, from CHOPCHOP 'CCACACCCAAATCTACTGTT' has a 23 base pair including the PAM 3 base pair sequence 'CCC'. **[See Figure 3 for Description of Double-strand break processes]**

For wheat, the plasmid can contain a Cas9 expression cassette with a strong constitutive promoter (maize ubiquitin, ZmUbi), a codon-optimized Cas9 coding sequence, and a nopaline synthase (Nos) terminator. An antibiotic resistance gene (*hpt* for hygromycin or *nptII* for kanamycin) can serve as a selectable marker during *Agrobacterium*-mediated transformation (Smedley et al., 2021). The guide RNA expression cassette can be driven by a plant U6 polymerase III promoter (TaU6) (Zhang et al., 2019). These components are assembled between the left and right borders of a binary vector for *Agrobacterium tumefaciens* delivery (Zhang et al., 2019).

For yeast, the plasmid can be driven by a yeast-specific promoter (constitutive TEF1 or inducible GAL1) (Elison et al., 2017). The Cas9 coding sequence can be then codon-optimized for S. cerevisiae and include a nuclear localization signal (NLS). A yeast terminator (CYC1) can complete the cassette (Goldberg et al., 2021). The guide RNA expression cassette can use a yeast U6 polymerase III promoter (SNR52) (Holland et al., 2025). An antibiotic resistance gene (kanMX) can confer G418 resistance for selection (Lorenz et al., 2015). The neprosin coding sequence, obtained from NCBI GenBank or UniProt and codon-optimized for yeast, can serve as the donor DNA for precise integration via HDR (Elison et al., 2017).

The sgRNA gene can be assembled alongside a Cas9 expression cassette and an antibiotic resistance marker into a single T-DNA construct. This construct is synthesized by a biotechnology company and introduced into *Agrobacterium tumefaciens* cells.

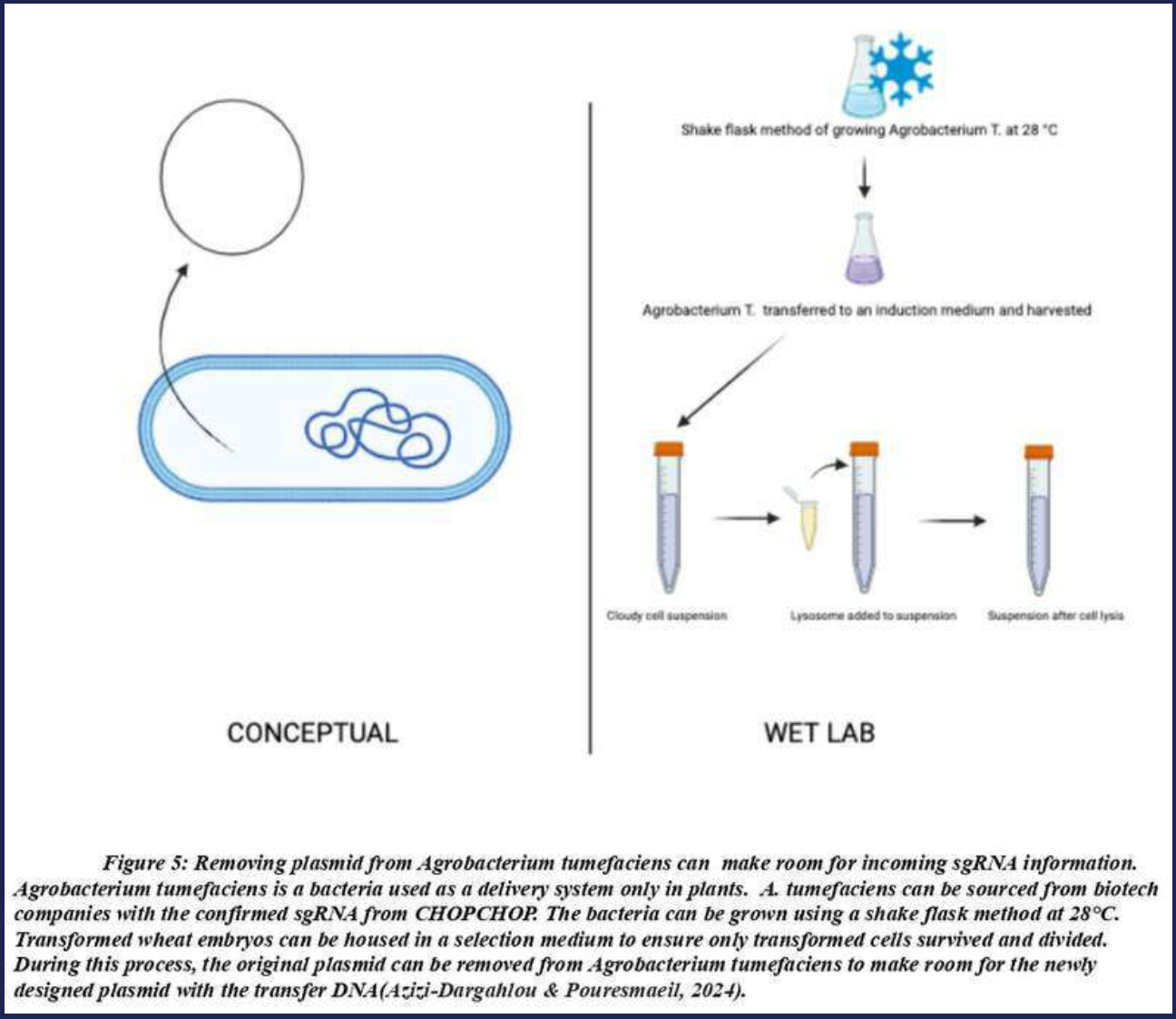


Figure 5: Removing plasmid from Agrobacterium tumefaciens can make room for incoming sgRNA information. Agrobacterium tumefaciens is a bacteria used as a delivery system only in plants. A. tumefaciens can be sourced from biotech companies with the confirmed sgRNA from CHOPCHOP. The bacteria can be grown using a shake flask method at 28°C. Transformed wheat embryos can be housed in a selection medium to ensure only transformed cells survived and divided. During this process, the original plasmid can be removed from Agrobacterium tumefaciens to make room for the newly designed plasmid with the transfer DNA(Azizi-Dargahlou & Pouresmaeil, 2024).

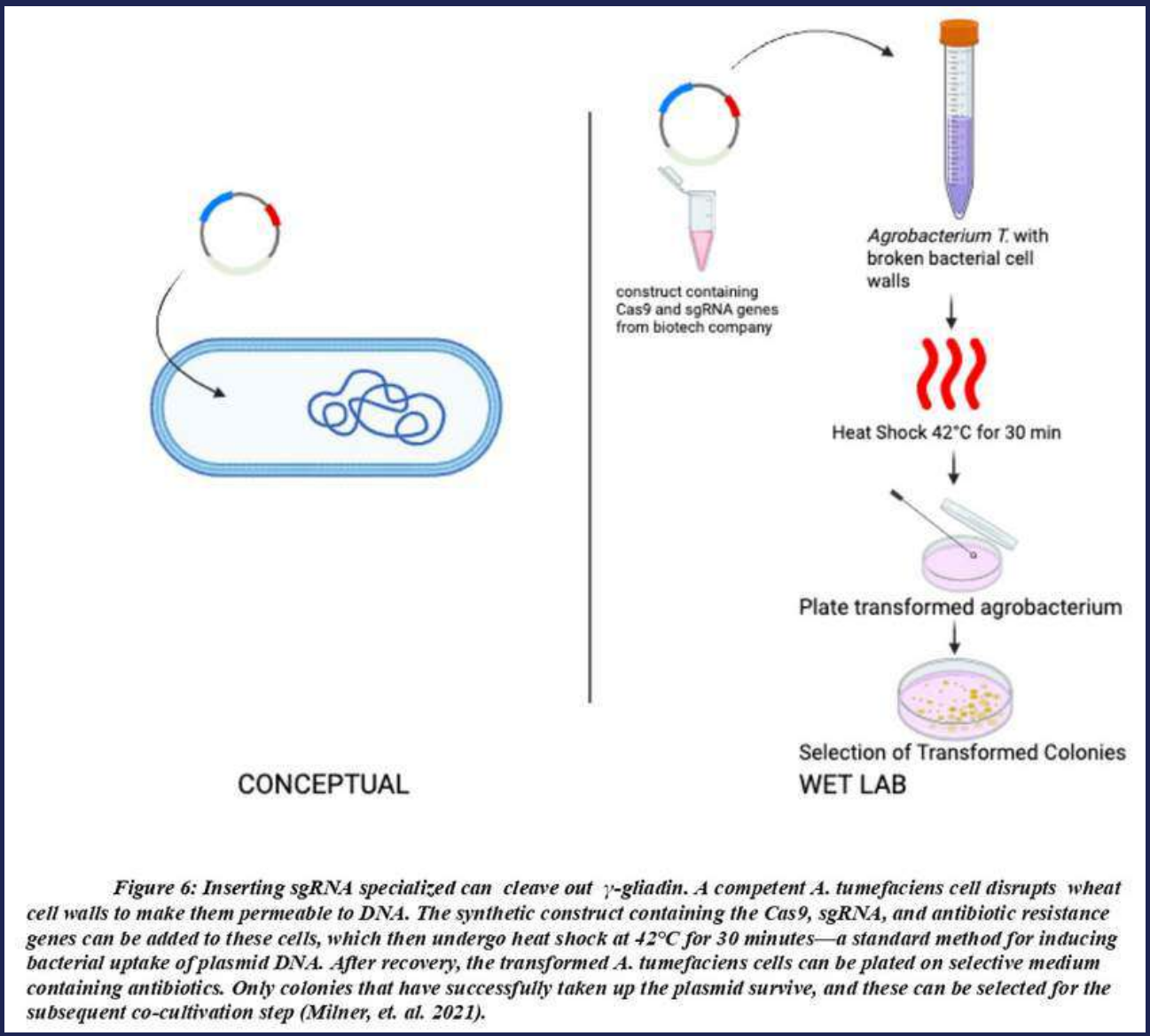


Figure 6: Inserting sgRNA specialized can cleave out γ-gliadin. A competent A. tumefaciens cell disrupts wheat cell walls to make them permeable to DNA. The synthetic construct containing the Cas9, sgRNA, and antibiotic resistance genes can be added to these cells, which then undergo heat shock at 42°C for 30 minutes—a standard method for inducing bacterial uptake of plasmid DNA. After recovery, the transformed A. tumefaciens cells can be plated on selective medium containing antibiotics. Only colonies that have successfully taken up the plasmid survive, and these can be selected for the subsequent co-cultivation step (Milner, et. al. 2021).

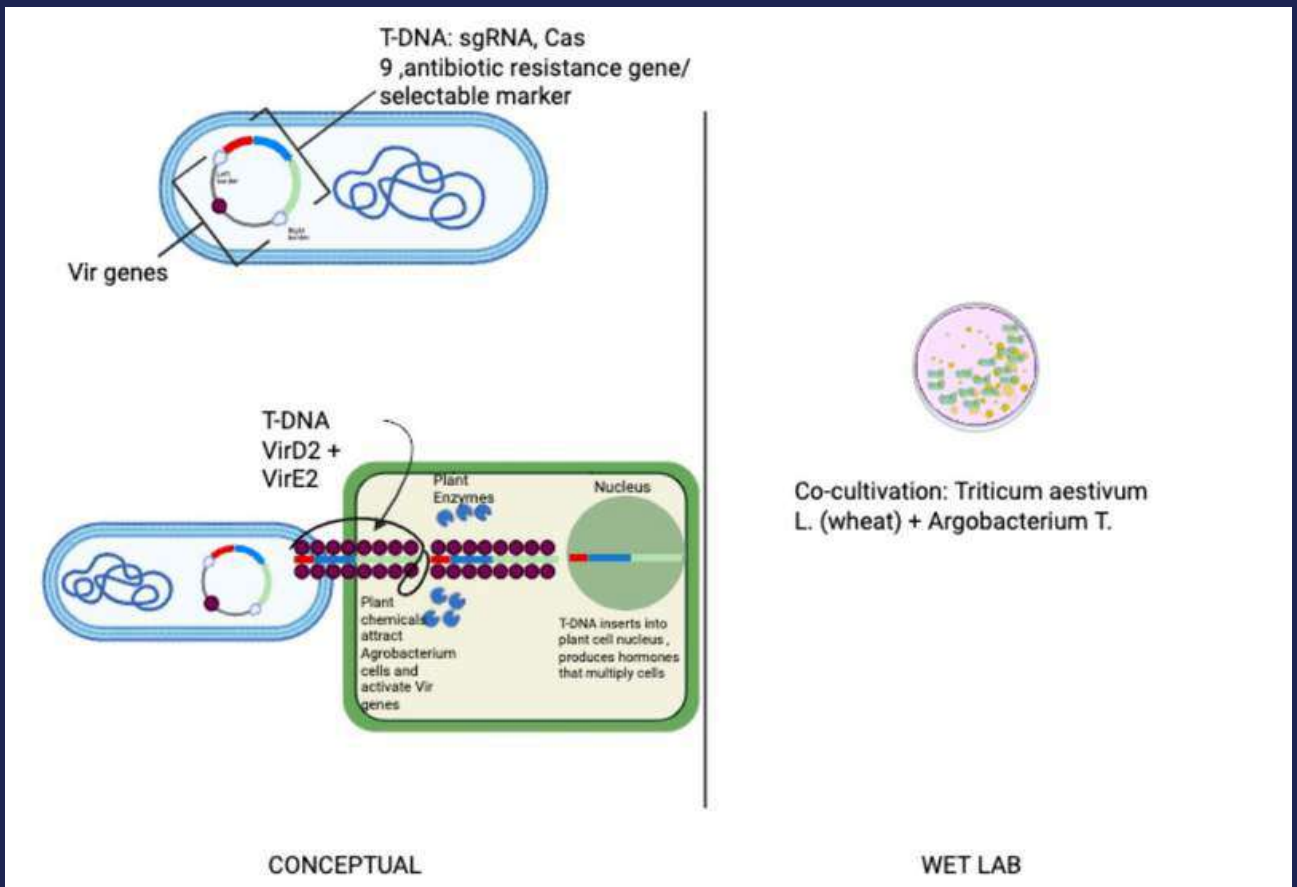


Figure 7: For wheat, Agrobacterium co-cultivation can be used to deliver T-DNA into the wheat cell nucleus.Once transformed A. tumefaciens colonies are obtained, they can be used to deliver the CRISPR machinery into wheat cells. A. tumefaciens can harbor a tumor-inducing (Ti) plasmid with virulence (vir) genes that become activated in the presence of plant-derived signals. These vir genes can process the T-DNA—which contains the sgRNA, Cas9, and antibiotic resistance cassette—and can produce VirD2 and VirE2 proteins that bind to the T-DNA, protecting it and guiding it into the plant cell nucleus. Upon integration into the wheat genome, the T-DNA can also carry genes that produce plant hormones, stimulating the transformed plant cells to multiply and form callus tissue (Zupan et. al , 1996).Immature wheat embryos can be co-cultivated with the transformed A. tumefaciens for a defined period, typically two to three days, under controlled conditions. During this time, A. tumefaciens can attach to wheat cells and transfer the T-DNA into them. After co-cultivation, the plant material can be washed with antibiotics to kill any remaining A. tumefaciens, ensuring that only plant cells that have integrated the T-DNA proceed to selection(Sparks et. al., 2013)

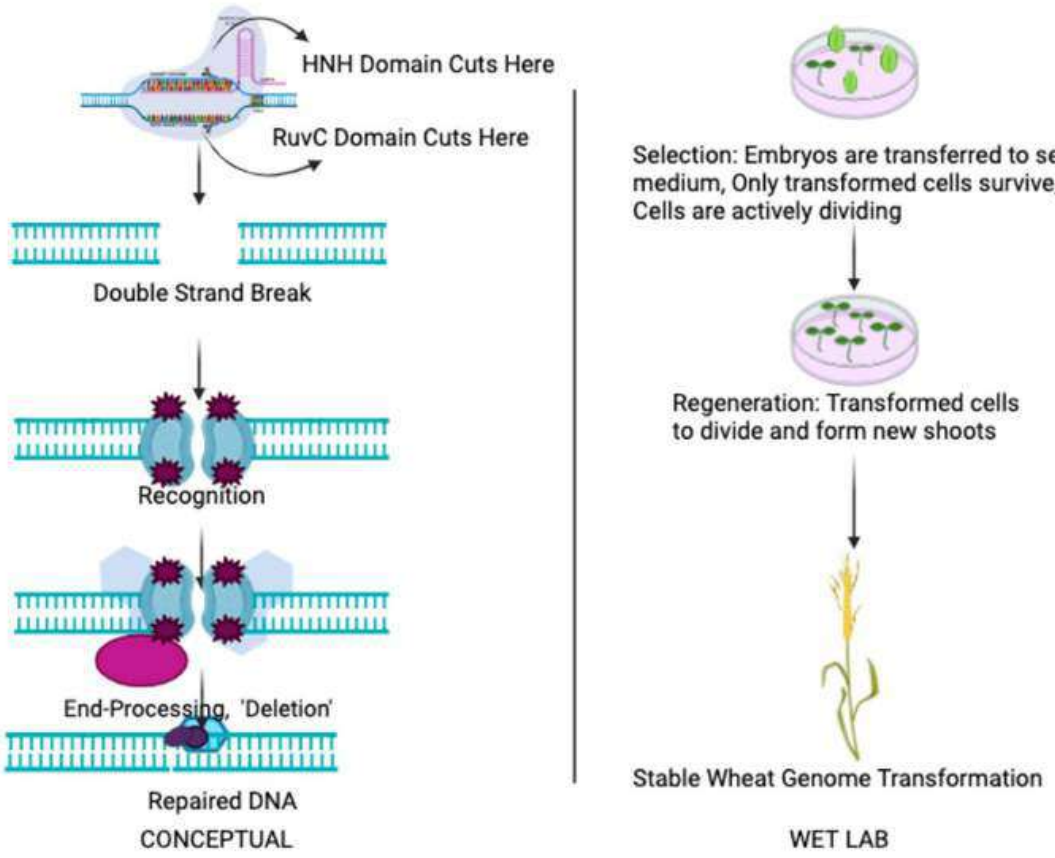


Figure 8: NHEJ can cleave γ-Gliadin in wheat. Upon recognition, Cas9 can introduce a double-strand break(Guzmán-López, 2021): the HNH domain cuts the target strand, while the RuvC domain cuts the non-target strand (Sun, 2017). The cell can repair this break through the error-prone non-homologous end joining (NHEJ) pathway, which can often introduce small insertions or deletions (indels) that disrupt the γ-gliadin gene, effectively knocking it out (Milner, 2021). Following co-cultivation, wheat embryos can be transferred to a selection medium containing the corresponding antibiotic (e.g., hygromycin or kanamycin). Only cells that have successfully integrated the T-DNA survive and can begin actively dividing(Sparks et. al., 2013). Surviving callus tissue can then be transferred to a regeneration medium containing plant growth regulators that induce shoot formation(Sparks et. al., 2013). These shoots can then root, acclimize, and grow to maturity. The resulting plants can be screened molecularly to confirm the presence of the desired edits in the γ-gliadin genes, yielding a stably transformed wheat line with reduced immunoreactive gluten content.

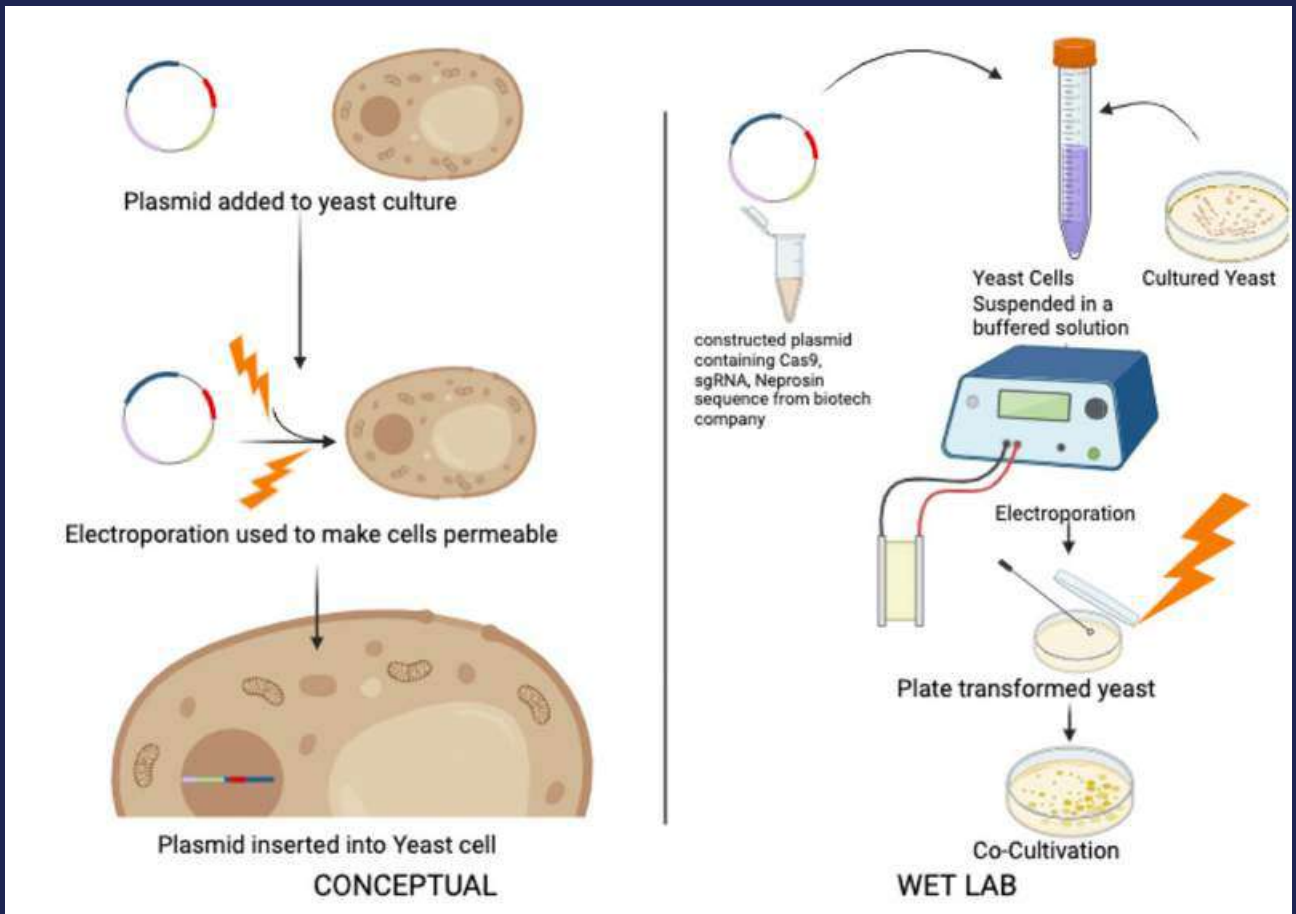


Figure 9: Yeast cells can be maintained in a buffered solution and cultured on plates following electroporation (Thompson, 1998).Transformed yeast cells can housed in a selection medium to ensure only transformed cells survived and divided. The yeast cell wall acts as an initial barrier to DNA uptake. It's thinner and more flexible structure compared to plant cell walls, thus ,allowing electroporation to create transient pores without prior wall removal (Zerbib & Donker, 2024; Ganeva et al., 2014).

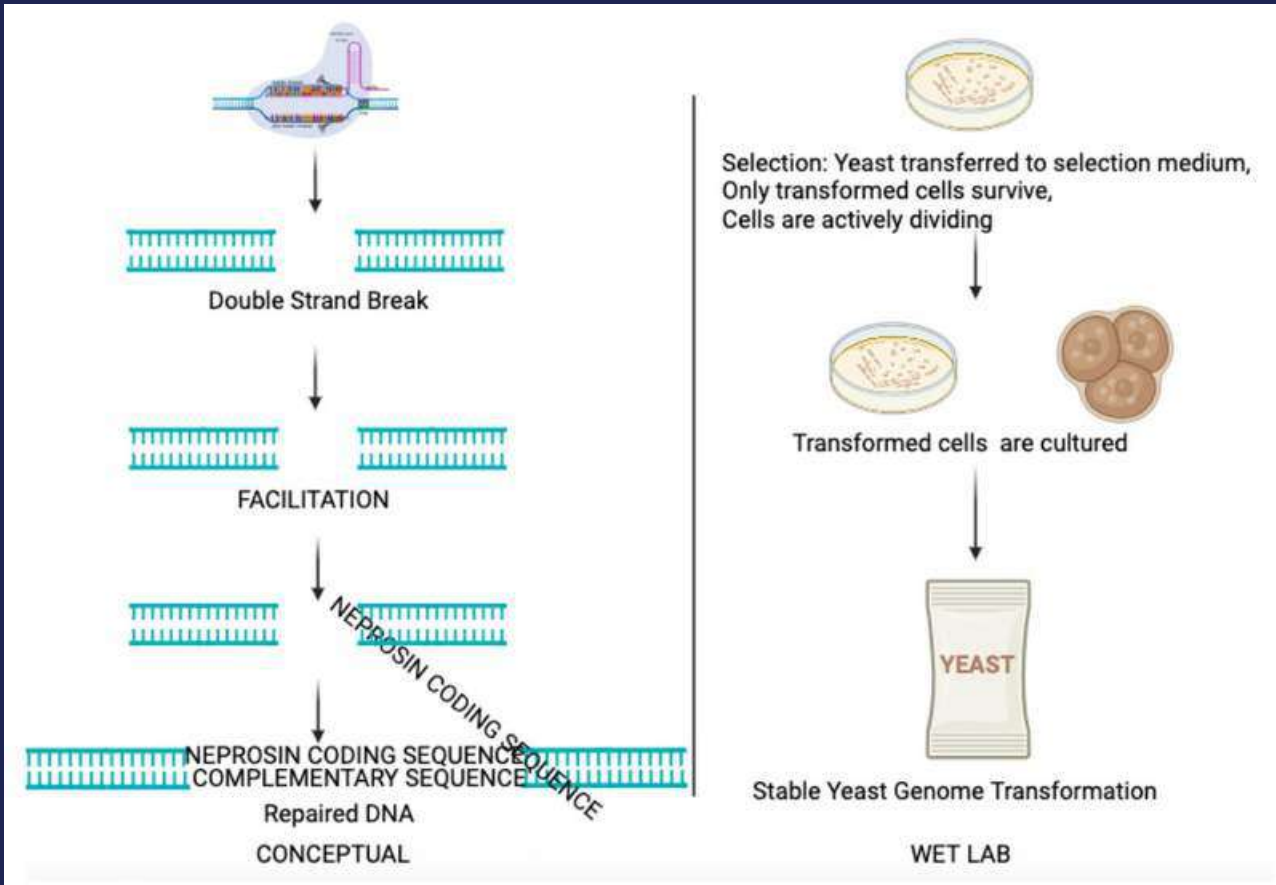


Figure 10: HDR resulting in Neprosin secretion to detoxify gluten in bread fermentation process. The new double-strand break serves as the insertion site for the neprosin coding sequence, a prolyl endopeptidase derived from Nepenthes species that degrades immunogenic gliadin peptides under acidic conditions. To enable secretion from yeast, the neprosin sequence can be fused to a yeast secretion signal (e.g., from SUC2) and codon-optimized for Saccharomyces cerevisiae(Schräder et. al., 2017).

Through homology-directed repair (HDR), transformed cells can then be transferred to selection medium containing G418; only those with successful integration survive and actively divide (Štafa et al., 2017). These selected transformants can be cultured to produce neprosin, which is secreted into the medium.

The final outcome can be stable genomic integration of the neprosin coding sequence, ensuring heritable and consistent enzyme production without continuous selection. In bread making, this engineered yeast can either be used directly as a leavening agent or serve as a source of secreted neprosin to degrade gliadin during dough fermentation, thereby reducing gluten immunogenicity.

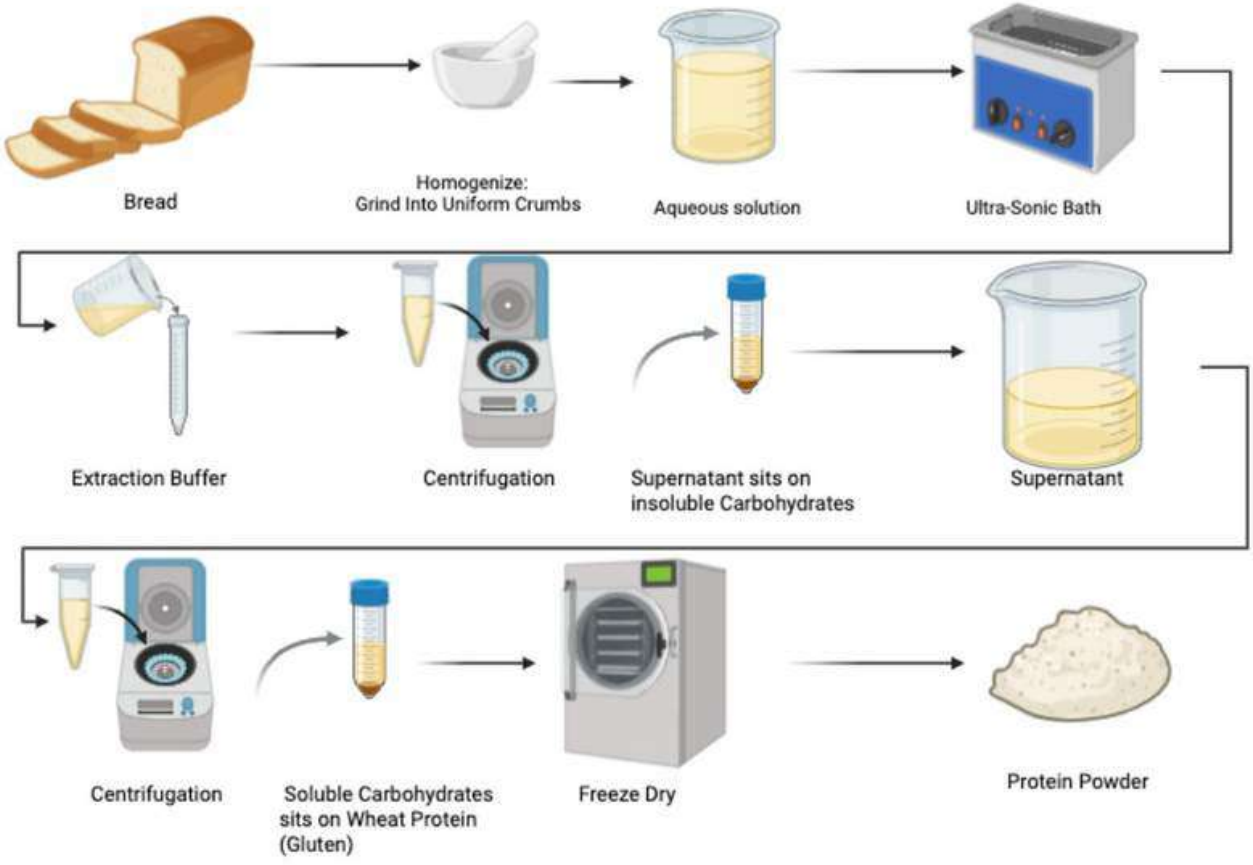


Figure 11: Extraction of gluten proteins with extraction buffer. Success can be validated (Lagrain, 2007).

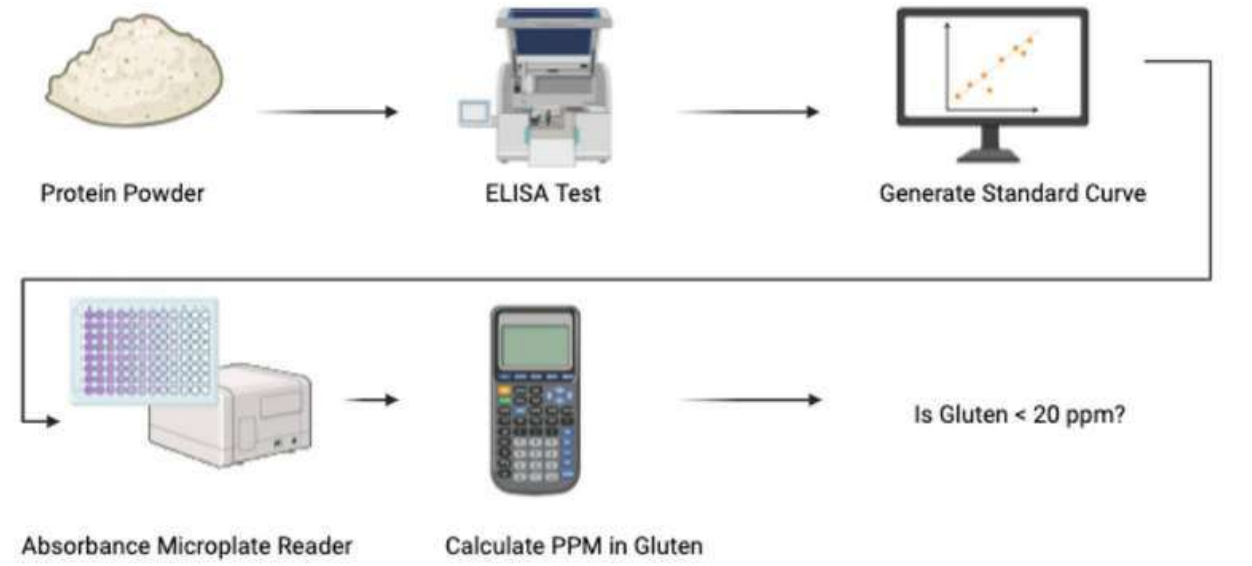


Figure 12: Validation. An ELISA test and an absorbance microplate reader can be used to generate a standard curve and calculate parts per million (PPM) of gluten (Sajic, 2017). The analytical goal is to determine if gluten levels were reduced to < 20 ppm (Thompson, 2015).

It was assumed that sgRNA design must prioritize high efficiency, low predicted off-targets, and target only exons/coding regions. (Stovicek et al., 2017) It was assumed that if efficiency is low in *CHOPCHOP*, there would be an allowance for the target sequence to be captured by a designed sgRNA sequence. It was assumed for the wheat genome is that *CHOPCHOP*'s alpha-gliadin method would work for γ-gliadin.

In yeast, it was assumed that the HDR (Homology-Directed Repair) pathway would dominate, allowing for predictable and precise gene insertions.

Results

The following results represent predicted outcomes generated using CHOPCHOP-based sgRNA design and supported by findings from existing literature. No experimental CRISPR editing was conducted in this study; instead, results reflect computational predictions and previously reported empirical data.

Case 1: CRISPR-Cas9 Editing in Hexaploid Wheat (α- and γ-Gliadin Silencing)

sgRNA Design and Predicted Editing Performance

Guide RNA design using CHOPCHOP for conserved α-gliadin target regions in *Triticum aestivum* identified multiple candidate SpCas9 sgRNAs distributed across the coding sequence. The top five ranked guides demonstrated predicted on-target efficiency scores ranging from 57.09 to 64.80 (mean = 61.04), with GC content between 40–55% and low self-complementarity (0–3). All five guides showed zero predicted off-target matches across MM0–MM3 categories within the selected genome index.

These parameters met the predefined design criteria of:

- High predicted efficiency
- Zero predicted off-targets
- Targeting within conserved exon regions
- Appropriate GC balance for stability

Rank	Target Sequence	Strand	GC(%)	Self Complementarity	MMO	Efficiency
1	GCGCGATTGTGCAATATGGAGGG	-	50	0	0	64.8
2	ACAACAACTGATTCCATGCATGG	+	40	1	0	63.34
3	TGGAACCTAACTGCAGTTGTGG	-	45	0	0	61.86
4	AACACCGTTTTCTCATGACGCGG	+	45	3		58.11

Impact on Immunodominant Epitopes

Saccharomyces cerevisiae does not produce gliadin. CRISPR editing in yeast therefore targeted insertion of genes encoding gluten-degrading enzymes rather than removal of wheat gluten genes (Liang et al., 2024). Engineered yeast expressing neprosin demonstrated effective degradation of the 33-mer α-gliadin peptide under physiological conditions (del Amo-Maestro et al., 2022; Ting et al., 2025).

Enzymatic cleavage of proline-rich gliadin peptides reduced the presence of intact immunogenic fragments during fermentation.

tTG Pathway Impact

Because enzymatic degradation occurred prior to intestinal exposure, the availability of intact gliadin peptides for tTG-mediated deamidation was reduced. This reduction occurred through substrate degradation before immune interaction.

PEP Pathway Impact

Yeast-expressed neprosin functioned similarly to prolyl endopeptidase therapy by cleaving proline-rich sequences resistant to endogenous digestive enzymes (Rey et al., 2016; Mitea et al., 2008).

Effect on Dough Viscoelasticity

Yeast-based editing did not alter the wheat genome, allowing native gluten gene expression to remain intact. Fermentation-dependent detoxification was less disruptive to rheological properties compared to upstream gliadin removal strategies (Pilolli et al., 2020; Zhang et al., 2024).

Feature	Hexaploid Wheat CRISPR	Yeast CRISPR
Target of editing	Gliadin genes	Yeast genome
Genome complexity	Extremely high	Low
Dominant repair pathway	NHEJ	HDR

controls (Jouanin et al., 2020; Zhang et al., 2024). Partial silencing approaches demonstrated less severe effects on viscoelastic properties.

Case 2: CRISPR-Cas9 Editing in Yeast (Neprosin Expression Strategy):

sgRNA Design and Predicted Editing Performance

For the yeast HO locus in *Saccharomyces cerevisiae*, CHOPCHOP identified top-ranked sgRNAs with predicted efficiency scores ranging from 67.40 to 74.49 (mean = 69.95). GC content ranged from 30–55%, and self-complementarity values were low (0–2). All five guides showed zero predicted off-target matches across MM0–MM3 categories.

Comparative analysis demonstrated higher predicted mean sgRNA efficiency in yeast relative to wheat under identical SpCas9 parameters.

Rank	Target Sequence	Strand	GC(%)	Self Complementarity	MMO	Efficiency
1	TATGGAAGATACAAATTTCAGCGG	-	30	0	0	74.49
2	CACAACTCTTATGAGGCCCCCGG	+	55	0	0	69.08
3	ATAGAAGTGAAATCATGTCGAGG	-	35	0	0	68.69
4	ATCATGTCGAGGCTGCTGTGTGG	-	55	2	0	70.12
5	CCTGTGTGACATTTATGACGCGG	+	45	0	0	67.4

Impact on Immunodominant Epitopes

Saccharomyces cerevisiae does not produce gliadin. CRISPR editing in yeast therefore targeted insertion of genes encoding gluten-degrading enzymes rather than removal of wheat gluten genes (Liang et al., 2024). Engineered yeast expressing neprosin demonstrated effective degradation of the 33-mer α-gliadin peptide under physiological conditions (del Amo-Maestro et al., 2022; Ting et al., 2025).

Enzymatic cleavage of proline-rich gliadin peptides reduced the presence of intact immunogenic fragments during fermentation.

tTG Pathway Impact

Because enzymatic degradation occurred prior to intestinal exposure, the availability of intact gliadin peptides for tTG-mediated deamidation was reduced. This reduction occurred through substrate degradation before immune interaction.

PEP Pathway Impact

Yeast-expressed neprosin functioned similarly to prolyl endopeptidase therapy by cleaving proline-rich sequences resistant to endogenous digestive enzymes (Rey et al., 2016; Mitea et al., 2008).

Effect on Dough Viscoelasticity

Yeast-based editing did not alter the wheat genome, allowing native gluten gene expression to remain intact. Fermentation-dependent detoxification was less disruptive to rheological properties compared to upstream gliadin removal strategies (Pilolli et al., 2020; Zhang et al., 2024).

Feature	Hexaploid Wheat CRISPR	Yeast CRISPR
Target of editing	Gliadin genes	Yeast genome
Genome complexity	Extremely high	Low
Dominant repair pathway	NHEJ	HDR
Precision	Moderate	High
Mosaicism risk	High	Low
Strategy	Upstream gene removal	Downstream processing modification
tTG pathway	Eliminates substrate	Degrades substrate
PEP pathway	Reduces need	Mimics PEP action
Dough viscoelasticity	Potentially altered	Preserved

Discussion

This research study compared CRISPR-Cas9 methods in yeast versus wheat to mediate the silencing of the α- and γ-gliadin gene families, evaluating which method most effectively eliminates immunodominant epitopes through the tTG and PEP pathways while preserving dough viscoelasticity. Both methods differed in how they reduced immunegenic gluten, where hexaploid wheat relied on upstream gene removal, and engineered yeast enabled downstream detoxification of gluten using neprosin (Pilolli et al., 2020; Zhang et al., 2024). In general, results support the alternative hypothesis that CRISPR-Cas9 in yeast most effectively eliminates immunodominant epitopes through the tTG and PEP pathways with higher precision whilst preserving dough viscoelasticity. However, one limitation that needs to be considered is that the results were drawn from different studies that conducted independent research under different uncontrolled experimental conditions.

The major impact on immunodominant epitopes in hexaploid wheat was that multiplex CRISPR-Cas9 targeting significantly reducing α- and γ-gliadin immunoreactive fractions and disrupted major immunodominant sequences encoding for 33-mer related regions, mainly through NHEJ-driven frameshifts that produced non-functional gliadin proteins (Jouanin et al., 2020; Verma et al., 2021; del Amo-Maestro et al., 2022; Ting et al., 2025). Nonetheless, all of α- and γ-gliadin epitopes could not be completely eliminated across the A,B, and D subgenomes with clusters at α- Gli-2 (chr 6) and γ- Gli-1 (chr 1) loci, and results were inconsistent due to factors such as mosaicism in polyploid systems with high gene redundancy and complex genomes (Verma et al., 2021). In multi-allelic crops such as wheat, α-gliadin genes are positioned in tandem arrays, leading to off-target genome targeting of bordering gliadin proteins, and increasing the difficulty in targeting conserved regions of the genome (Schaart et al., 2021). Conversely, CRISPR-engineered yeast affects gluten indirectly, as it does not produce gliadin, and instead enables neprosin-controlled expression as proposed by Shan et al (2002). In regards to the tTG and PEP pathways, hexaploid wheat eliminates substrate availability for tTG and reduces the need for the PEP pathway, whereas yeast degreades the substrate and mimics the PEP action by cleaving proline-rich peptides. Above all, the CRISPR-Cas9 editing effects on dough viscoelasticity provide the main tradeoff between the two systems because they determine baking performance. In wheat, gliadins mainly contribute to dough extensibility, while glutenins support elasticity, thus proving that extensive α- and γ-gliadin silencing changes rheological balance and decreases extensibility when compared to wild type controls (Jouanin et al., 2020; Zhang et al., 2024). In contrast, yeast's fermentation-dependent detoxification was less disruptive to rheological properties because its editing did not change the wheat genome and left gluten gene expression intact, rather than removing structural protein families as wheat, thereby preserving dough viscoelasticity (Piloli et al.2020; Zhang et al., 2024).

CRISPR-Cas9 efficiency strongly depends on sgRNA design, thus CHOPCHOP was used to identify target gene IDs that helped generate the target sequences, and the sgRNA design was compared between the two systems, wheat and yeast. Results showed that yeast had an overall higher mean efficiency of 69.95, as compared to wheat's mean efficiency of 61.04 and 0 predicted off-targets. However, this computational design had some limiting implications with the wheat genome as CHOPCHOP's database only contained chromosome 1 of *Triticum aestivum*. Moreover, editing performance depends on sgRNA design, in which the same target sequences adjacent to PAM must be conserved across wheat chromosomes, and if not conserved, then three different sgRNAs would need to be designed in one construct, making it less efficient than yeast's design. Comparatively, this study also contained some biological implications in terms of the repair pathway of both organisms, where wheat transformation relied on *Agrobacterium* delivering T-DNA into wheat cells, with the repair being mainly NHEJ-guided and thus more associated with indels and frameshift mutations, whereas yeast editing used more precise HDR-based insertion as neprosin expression enabling downstream detoxification, degradation of immunogenic peptides, and lowering intact epitopes before they pass through the intestine(Azizi-Dargahlou & Pouresmaeil, 2024). This difference in NHEJ and HDR mechanism of wheat and yeast supports why yeast was deemed to be the more precise system in comparison.

Future improvements are needed to better the research about the CRISPR-Cas9 strategies used in editing yeast and wheat genomes, which includes expanding multiplexing in wheat Gli-1 and Gli-2 targeting to cover a wider majority of gliadin copies to combat effects of mosaicism. Additionally, techniques as proteogenomic should be used to measure the presence of intact 33-mer immunogenic peptides instead of only relying on gene disruption. For yeast, testing and controlling fermentation conditions such as temperature, time and pH to increase efficiency of neprosin for breakdown of peptides, and quantifying detoxification of gluten by measuring the remaining immunogenic peptides after fermentation would provide a better comparison of results (Pilolli et al., 2020; Zhang et al., 2024). Moreover, future applications should take into account the biosafety protocols and ethical barriers, particularly for engineered yeast, as its implementation in the production of bread in the food industry would require consideration of several factors as industrial fermentation, food manufacturing, regulatory approvals, and biosafety law.

Conclusion

This study compared the effectiveness of two methods of CRISPR-Cas9 in yeast and wheat for reducing immunodominant epitopes of gluten whilst preserving dough viscoelasticity, including the upstream silencing of α- and γ-gliadin families in wheat and downstream processing modification using engineered neprosin-expressing yeast (Pilolli et al., 2020; Zhang et al., 2024). In hexaploid wheat, there are three distinct but closely related subgenomes A, B, and D with repeated clusters at α- Gli-2 (chr 6) and γ- Gli-1 (chr 1), which create a challenge in targeting all epitopes for removal due the complex nature of the wheat genome (Verma et al., 2021). Even though CHOPCHOP's sgRNA design provided SpCas9 targets with 0 predicted off-targets, editing still depends on the conserved 20-bp sequence adjacent to PAM sites across group 1 chromosomes (1AS, 1BS, 1DS) and group 6 chromosomes (6AS, 6BS, 6DS), with NHEJ-driven repair mainly producing indels and non-functional gliadin proteins. In contrast, yeast editing consists of a small genome with highly efficient homologous recombination, supporting easier transformation with CRISPR plasmids and more precise HDR gene insertions such as neprosin, while reducing immunogenic peptides reaching the intestines by limiting the tTG substrates and PEP pathways (Verma et al., 2021). This is done by modifying the environment instead of removing the gluten genes as in wheat. However, this comparison is limited as it draws their differences from several independent studies.Overall, results show more promising scalability and improvement potential for gluten detoxification using engineered yeast, whilst edited wheat strategies still require more improvement in key methods such as the range of multiplex coverage in its genome (Pilolli et al., 2020; Zhang et al., 2024).

Ethical Considerations: We confirm that this submission is our original work, that all sources are properly cited, and that we understand BACSA is not responsible for any plagiarism or academic misconduct.

Disclaimer: Lab protocols were used to compare the conceptual theory with practical lab work. Literature was written in compliance with BACSA's adherence to university-level biosafety rules. No lab work was performed under the guidelines of BACSA. All original images.

References

Aberkane, H., Payne, T., Kishi, M., Smale, M., Amri, A., & Jamora, N. (2020). Transferring diversity of goat grass to farmers' fields through the development of synthetic hexaploid wheat. Food Security, *12*(5), 1017-1033.

del Amo-Maestro, L., van den Broek, L. A. M., Cordfunke, R. A., et al. (2022). Molecular and in vivo studies of a glutamate-class prolyl-endopeptidase for coeliac disease therapy. *Nature Communications*, 13th ed, 4612. https://doi.org/10.1038/s41467-022-32215-1

Azizi-Dargahlou, S., & Pouresmaeil, M. (2024). *Agrobacterium tumefaciens*-mediated plant transformation: a review. *Molecular Biotechnology*, *66*(7), 1563-1580.

Caio, G., Volta, U., Sapone, A., Leffler, D. A., De Giorgio, R., Catassi, C., & Fasano, A. (2019). Celiac disease: a Comprehensive Current Review. *BMC Medicine*, *17*(1). https://doi.org/10.1186/s12916-019-1380-z

Elison, G. L., Song, Y., & Acar, M. (2017). A precise genome editing method reveals insights into the activity of eukaryotic promoters. *Cell Reports*, *18* (1), 275–286. https://doi.org/10.1016/j.celrep.2016.12.014

Ganeva, V., Galutzov, B., & Teissié, J. (2014). Electric field‐induced effects on yeast cell wall permeabilization. Bioelectromagnetics, *35*(2), 136–144.

Gliadin - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/gliadin

Goldberg, G. W., Spencer, J. M., Giganti, D. O., Camellato, B. R., Agmon, N., Ichikawa, D. M., Boeke, J. D., & Noyes, M. B. (2021). Engineered dual selection for directed evolution of SpCas9 PAM specificity. *Nature Communications*, *12* (1), 349. https://doi.org/10.1038/s41467-020-20650-x

Guzmán-López, M. H., Marín-Sanz, M., Sánchez-León, S., & Barro, F. (2021). A bioinformatic workflow for InDel analysis in the wheat multi-copy α-gliadin gene family engineered with CRISPR/Cas9. *International Journal of Molecular Sciences*, *22*(23), 13076. https://doi.org/10.3390/ijms222313076

Holland, K. L., et al. (2025). RNA polymerase III promoters compatible with CRISPR gene regulation in *Saccharomyces cerevisiae*. *ACS Synthetic Biology*. Advance online publication. https://doi.org/10.1021/acssynbio.5c00122

Jiang, F., & Doudna, J. A. (2017). CRISPR–Cas9 structures and mechanisms. *Annual review of biophysics*, 46th ed, 505-529.

Jouanin, A., Gilissen, L. J. W. J., Schäfer, B. W., Visser, R. G. F., & Smulders, M. J. M. (2020). CRISPR/Cas9 gene editing of gluten in wheat to reduce gluten content and exposure—Reviewing methods to screen for coeliac safety. *Frontiers in Nutrition*, 7th ed, 51. https://doi.org/10.3389/fnut.2020.00051

Lagrain, B., Thewissen, B. G., Brijs, K., & Delcour, J. A. (2007). Impact of redox agents on the extractability of gluten proteins during bread making. *Journal of Agricultural and food chemistry*, *55*(13), 5320-5325.

Liang, Y., Gao, S., Qi, X., Valentovich, L. N., & An, Y. (2024). Progress in Gene Editing and Metabolic Regulation of *Saccharomyces cerevisiae* with CRISPR/Cas9 Tools. *ACS Synthetic Biology*, *13* (2), 428–448. https://doi.org/10.1021/acssynbio.3c00685

Lorenz, D. R., Meyer, L. F., & Larin, Z. (2015). Single-step marker switching in *Schizosaccharomyces pombe* using a lithium acetate transformation protocol. *Bio-protocol*, *5* (22), e2075. https://bio-protocol.org/en/bpdetail?id=2075

Marshall, J. K. (2013). The Burden of Celiac Disease in Canada: More Work Needed to Lighten the Load. *Canadian Journal of Gastroenterology*, *27* (8), 448–448. https://doi.org/10.1155/2013/516498

Mesta-Corral, M., Gómez-García, R., Balagurusamy, N., Torres-León, C., & Hernández-Almanza, A. Y. (2024). Technological and Nutritional Aspects of Bread Production: An Overview of Current Status and Future Challenges. *Foods*, *13*(13), 2062. https://doi.org/10.3390/foods13132062

Milner, M. J., & Wallington, E. J. (2021). Genome editing and identification of targeted heritable mutations in wheat. In *Accelerated Breeding of Cereal Crops* (pp. 225-238). New York, NY: Springer US.

Mitea, C., Havenaar, R., Drijfhout, J. W., Edens, L., Dekking, L., & Koning, F. (2008). Efficient degradation of gluten by a prolyl endoprotease in a gastrointestinal model: Implications for coeliac disease. *Gut*, *57*(1), 25–32. https://doi.org/10.1136/gut.2006.111609

Molberg, Ø., McAdam, S. N., Körner, R., Quarsten, H., Kristiansen, C., Madsen, L., Fugger, L., Scott, H., Norén, O., Roepstorff, P., Lundin, K. E. A., & Sollid, L. M. (1998). Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease. *Nature Medicine*, *4*(6), 713–717. https://doi.org/10.1038/nm0698-713

Mudryj, A., Waugh, A., Slater, J., Duerksen, D. R., Bernstein, C. N., & Riediger, N. D. (2021). Dietary gluten avoidance in Canada: a cross-sectional study using survey data. *CMAJ Open*, *9*(2), E317–E323. https://doi.org/10.9778/cmajo.20200082

Pilolli, Rosa, et al. Prototype gluten-free breads from processed Durum wheat: use of monovarietal flours and implications for gluten detoxification strategies. *Nutrients*. *12* (2020): 3824.

Rey, M., Yang, M., Burns, K. M., Yu, Y., Lees-Miller, S. P., & Muench, D. G. (2016). Nepenthes digestive fluid contains a potent gluten-degrading enzyme. *Scientific Reports*, *6*, 30980. https://doi.org/10.1038/srep30980

Sajic, N., Oplatowska-Stachowiak, M., Streppel, L., Drijfhout, J. W., Salden, M., & Koning, F. (2017). Development and in-house validation of a competitive ELISA for the quantitative detection of gluten in food. *Food Control*, *80*, 401-410.

Sánchez-León, S., & Barro, F. (2024). CRISPR/Cas9-mediated multiplex gene editing of gamma and omega gliadins: paving the way for gliadin-free wheat. *Journal of Experimental Botany*, *75*(22), 7079–7095. https://doi.org/10.1093/jxb/erae376

Schaart, J. G., van de Wiel, C. C. M., & Smulders, M. J. M. (2021). Genome editing of polyploid crops: prospects, achievements and bottlenecks. *Transgenic Research*. https://doi.org/10.1007/s11248-021-00251-0

Schräder, C. U., Lee, L., Rey, M., Sarpe, V., Man, P., Sharma, S., Zabrouskov, V., Larsen, B., & Schriemer, D. C. (2017). Neprosin, a selective prolyl endoprotease for bottom-up proteomics and histone mapping. *Molecular & Cellular Proteomics*, *16*(6), 1162–1173. https://doi.org/10.1074/mcp.M116.066308

Shan, L., Molberg, Ø., Parrot, I., Hausch, F., Filiz, F., Gray, G. M., Sollid, L. M., & Khosla, C. (2002). Structural basis for gluten intolerance in celiac sprue. *Science*, *297*(5590), 2275–2279. https://doi.org/10.1126/science.1074129

Smedley, M. A., Hayta, S., Clarke, M., & Harwood, W. A. (2021). CRISPR-Cas9 based genome editing in wheat. *Current Protocols*, *1*(3), e65. https://doi.org/10.1002/cpz1.65

Sparks, C. A., Doherty, A., & Jones, H. D. (2013). Genetic transformation of wheat via *Agrobacterium*-mediated DNA delivery. In R. J. Henry & A. Furtado (Eds.), *Cereal genomics: Methods and protocols* (Methods in Molecular Biology, *1099*, pp. 235–250). Humana Press. https://doi.org/10.1007/978-1-62703-715-0_19

Štafa, A., Miklenić, M. S., Zandona, A., Žunar, B., Čadež, N., Petković, H., & Svetec, I. K. (2017). In *Saccharomyces cerevisiae* gene targeting fidelity depends on a transformation method and proportion of the overall length of the transforming and targeted DNA. *FEMS yeast research*, *17*(4), fox041.

Sun, Y. (2017). *Development and applications of CRISPR-Cas9 and RNAi for rice and wheat agronomic traits improvement* (Doctoral thesis). Université de Liège. http://hdl.handle.net/2268/211944

Thompson, J. R., Register, E., Curotto, J., Kurtz, M., & Kelly, R. (1998). An improved protocol for the preparation of yeast cells for transformation by electroporation. *Yeast*, *14*(6), 565-571.

Thompson, T., & Simpson, S. (2015). A comparison of gluten levels in labeled gluten-free and certified gluten-free foods sold in the United States. *European journal of clinical nutrition*, *69*(2), 143-146.

Ting, T.-Y., Lee, W.-J., Ramzi, A. B., & Goh, H.-H. (2025). Bioengineered *Saccharomyces cerevisiae* with neprosin for gluten detoxification. *Journal of Future Foods*. https://doi.org/10.1016/j.jfutfo.2025.01.008

Verma, S., Kumar, A., & Mishra, P. (2021). Recent advances in CRISPR/Cas-mediated genome editing of wheat for improved nutritional quality and reduced allergenicity. *Foods*, *10*(10), 2351. https://doi.org/10.3390/foods10102351

Wang, X., Anders, S., Jiang, Z., Bruce, M., Gidrewicz, D., Marcon, M., Turner, J. M., & Mager, D. R. (2024). Food insecurity impacts diet quality and adherence to the gluten-free diet in youth with celiac disease. *Journal of Pediatric Gastroenterology and Nutrition*, *79*(6). https://doi.org/10.1002/jpn3.12398

Zhang, Z., Hua, L., Gupta, A., Tricoli, D., Edwards, K. J., Yang, B., & Li, W. (2019). Development of an *Agrobacterium*-delivered CRISPR/Cas9 system for wheat genome editing. *Plant Biotechnology Journal*, *17*(8), 1623–1635. https://doi.org/10.1111/pbi.13088

Zhang, Y., Wu, H., & Fu, L. (2024). A review of gluten detoxification in wheat for food applications: approaches, mechanisms, and implications. Critical Reviews in Food Science and Nutrition, 1–17. https://doi.org/10.1080/10408398.2024.2326618

Zupan, J. R., Citovsky, V., & Zambryski, P. (1996). *Agrobacterium* VirE2 protein mediates nuclear uptake of single-stranded DNA in plant cells. *Proceedings of the National Academy of Sciences*, *93*(6), 2392–2397. https://doi.org/10.1073/pnas.93.6.2392

Early-Life Chronic Stress, Epigenetic Resilience, and Adult Stress-Related Outcomes: A Computational and Systematic Review Approach

Sosan Akram, Pol Reznichenko, Camilla de Solminihae

Abstract

Early-life stressors (ELS) and adverse childhood experiences (ACEs) trigger hypermethylation of the NR3C1gene (1Fexon), reducing glucocorticoid receptor (GR) production and disrupting HPA axis negative feedback, leading to prolonged cortisol elevation, poor sleep quality, emotional instability, and heightened suicide risk. This project systematically reviews public datasets via PubMed and related databases, applies linear regression and t-tests to correlate ELS with outcomes like sleep disruption and daily stress, and identifies resilient outliers despite high exposure. Findings highlight epigenetic plasticity, with positive childhood experiences (PCEs) showing reversal potential, informing interventions for stress resilience; limitations include reliance on existing data without new primary collection.

Introduction

Childhood adversity and chronic early-life stress (ELS), such as adverse childhood experiences (ACEs), are robust predictors of long-term mental and physical health vulnerabilities, including anxiety, depression, nervous system dysregulation, and impaired sleep quality. These outcomes emerge progressively due to sustained environmental adversity during sensitive developmental periods, often through epigenetic mechanisms like DNA hypermethylation, where methyl groups bind primarily to CpG sites on nucleotides, silencing gene expression and preventing protein production (Lopez et al., 2021).

Under normal acute stress, the Hypothalamic-Pituitary-Adrenal (HPA) axis, comprising the hypothalamus, pituitary gland, and adrenal glands, activates to release adrenocorticotropic hormone (ACTH), prompting cortisol production, a stress hormone, for environmental adaptation. Cortisol then binds to glucocorticoid receptors (GR), encoded by the NR3C1 gene (exon 1F), to provide negative feedback and halt further HPA activity (van der Knaap et al., 2014). However, ELS often induces hypermethylation at NR3C1 exon 1F, reducing GR expression, elevating the cortisol threshold, and resulting in prolonged hypercortisolemia, overactive stress responses, emotional instability, heightened suicide risk, and sleep disruptions via methylation of related genes like SLC6A4 (chromosome 17) or AVP (chromosome 5) (Bakusic et al., 2020; Dempster et al., 2007).

While epigenetic marks were once considered irreversible, recent evidence highlights substantial plasticity, including active demethylation pathways involving TET enzymes and base excision repair, as well as longitudinal data showing positive childhood experiences (PCEs) can mitigate adverse effects and foster natural resilience (Turecki et al., 2014). Not all ELS-exposed individuals develop maladaptive outcomes; some exhibit adaptive sleep patterns, lower daily stress, and preserved functioning, reflecting heterogeneity in vulnerability and resilience (Liu & Nusslock, 2018).

Despite these advances, few studies have simultaneously examined the relationship between ELS, sleep and stress outcomes, and NR3C1 methylation while explicitly identifying resilient individuals. Therefore, this study aims to investigate how early-life stress influences sleep quality, daily stress, and epigenetic regulation, while identifying individuals who demonstrate resilience despite high levels of exposure.

Rationale

Early-life stress has been widely studied in relation to adverse health outcomes and epigenetic mechanisms; however, limited research has integrated sleep and daily stress outcomes with epigenetic data while explicitly identifying resilient individuals. This gap restricts a more complete understanding of how biological mechanisms and behavioral outcomes interact in the context of resilience. This study addresses this limitation by examining the relationship between early-life stress, sleep quality, daily stress, and, where available, DNA methylation patterns, while identifying individuals who demonstrate better-than-expected outcomes. By doing so, it aims to clarify pathways of both vulnerability and resilience and contribute to more targeted approaches for mitigating the long-term effects of childhood adversity (DiMarzio et al., 2024).

Hypothesis

Individuals with high ELS/ACE exposure will exhibit significantly elevated mean daily stress scores ($p<0.05$) and diminished mean sleep quality/efficiency ($p<0.05$), and these differences will be partly associated with altered DNA methylation patterns in stress-related genes such as the glucocorticoid receptor gene. The data will be accompanied by greater between-subject standard deviations (SD highELS > SD lowELS), reflecting underlying heterogeneity in epigenetic resilience characterized by relatively favorable sleep and daily stress outcomes despite high early-life stress exposure.

Objective

The primary objectives are to: (1) systematically identify and curate public datasets linking ELS/ACEs to quantifiable sleep quality, daily stress, health outcomes, and NR3C1 methylation; (2) apply linear regression to estimate effect sizes of ELS on outcomes, adjusting for covariates like age/sex; (3) use t-tests and residual thresholds (>1.5 SD) to detect resilient outliers and compare their profiles; (4) synthesize findings with epigenetic plasticity literature (e.g., TET pathways, PCE effects) to model vulnerability-resilience and (5) propose intervention targets based on asset efficacy across exposure.

Methods

A systematic review protocol was conducted using databases including PubMed, ScienceDirect, the National Library of Medicine, and APA PsycINFO. Search terms included "adverse childhood experiences," "early-life stress," "NR3C1 methylation," "sleep quality," "resilience," and "HPA axis." Studies were included if they involved human participants, reported quantitative measures of early-life stress (e.g., 10-item ACE scores, ELS survey), and provided outcomes related to sleep, daily stress, or health, with accessible summary statistics or individual-level data suitable for analysis. While DNA methylation data were prioritized, studies without epigenetic measures were included if they met other criteria. Studies were also required to have accessible statistics and cohort data suitable for secondary analysis. Studies were excluded if they were non-human, qualitative-only, or lacked accessible data. preferred. Data extraction included means, standard deviations (SD), correlations, and relevant covariates such as age and sex.

In total, 12 primary datasets and supporting mechanistic studies were retained because each contributed a distinct link in the proposed pathway from early-life stress to adult outcomes. Large population surveys (*Client Challenge*, 2026; Wu et al., 2024) were selected for their quantitative measures of adverse childhood experiences (ACEs), health and sleep outcomes, and resilience or positive childhood experiences. These data directly informed the pooled estimates presented in Tables 1-2 and the regression analyses in Figures 1-3. Mechanistic and epigenetic studies (Bakusic et al., 2020; Cicchetti & Handley, 2017; Daskalakis et al., 2015; Liu & Nusslock, 2018; Forum et al., 2025) were included to explain underlying biological processes, demonstrating how early adversity is associated with differential methylation of NR3C1 and related stress-pathway genes, leading to dysregulation of hypothalamic-pituitary-adrenal (HPA) axis feedback. Additional methodological and contextual sources (Murata et al., 2019; Lopez et al., 2021; Turecki et al., 2014) supported the use of saliva-based methylation measures in large cohorts and reinforced evidence that epigenetic markers remain partially plastic and responsive to positive environmental influences.

Computational analyses were performed using Microsoft Excel. Linear regression models were applied to assess early-life stress (ELS) as a predictor of continuous outcomes, including daily stress and sleep efficiency. Independent t-tests were conducted to compare high- and low-ELS groups, as well as resilient and non-resilient individuals. Resilient outliers were identified using residual values exceeding ±1.5 standard deviations from the regression line, consistent with established approaches for detecting heterogeneity.

Effect sizes (Cohen's d and standardized regression coefficients, β) were calculated to quantify associations, and variability across groups was assessed using standard deviation comparisons as an indicator of heterogeneity. Statistical significance was set at $\alpha = 0.05$.

Analyses were conducted using de-identified secondary datasets. Ethical considerations included accurate representation of findings and sensitivity in reporting results related to trauma and early-life adversity,

Results

The review draws pooled data across 12 datasets, including DiMarzio et al. (2024), which linked early-life adversity (ELA) to reduced sleep efficiency and delayed sleep timing, and *Client Challenge* (2026), which examined relationships between ACE exposure and health outcomes. Studies were selected based on the following inclusion criteria: (1) subjects identified via the quantitative ACE/ELS surveys, (2) reported outcomes in relation to categories such as methylation, sleep, health, mental health as well as daily stress, (3) accessible data that could be used for secondary analysis.

For Table 1, data was compiled from *Client Challenge* (2026), a nationally representative British cohort (n = 2,452; ages 18-69), which utilized the standard 10-item ACE questionnaire alongside seven retrospectively reported childhood resilience assets, including feeling loved, access to trusted adults, opportunities for skill development, fair treatment, friendships, role models, and school enjoyment. As summarized in Table 1, increasing ACE/ELS exposure was associated with progressively poorer outcomes across domains. The prevalence of poor health rose from 17.2% in individuals with 0 ACEs to 37% in those with ≥4 ACEs. However, high levels of resilience assets reduced this risk by 38.5%, highlighting a substantial protective effect. Similar dose-response patterns were observed in sleep outcomes (DiMarzio et al., 2024; Wu et al., 2024), where individuals with high ACE/ELS exposure exhibited delayed sleep timing and reduced efficiency (SD = 2.95, n = 861), while low-exposure groups maintained typical sleep patterns. Notably, some individuals with high ACE/ELS exposure demonstrated preserved sleep function, indicating resilience despite adversity.

Epigenetic outcomes showed a parallel pattern, with accelerated aging observed in high ACE/ELS groups (PedBE SD = 0.85 above baseline), potentially moderated by positive childhood experiences (PCEs). Mechanistic studies further indicated that high ACE/ELS exposure was associated with increased NR3C1 methylation at exon 1F (Bakusic et al., 2020; Cicchetti & Handley, 2017), contributing to dysregulation of the HPA axis. In contrast, resilient individuals exhibited comparatively lower methylation levels, suggesting biological variability that may support adaptive functioning.

For Table 2, additional data were drawn from the Taiwan Youth Panel Survey, a national longitudinal cohort tracking 2,841 participants from early adolescence (age 14 in 2004) to emerging adulthood (ages 20–22 in 2011). This dataset included measures of depressive symptoms, insomnia, and positive childhood experiences. While insomnia symptoms were slightly higher in the high-ACE group—consistent with expected trends—depressive symptoms did not follow the same pattern. Specifically, individuals with ≤4 ACEs reported lower depressive symptoms compared to those with >4 ACEs, partially contradicting the hypothesis. This inconsistency highlights variability in individual responses to early-life stress and suggests that outcomes may differ across domains or be moderated by additional protective factors..

To further examine these relationships, linear regression models were applied (Figures 1-3). The resulting scatter plots demonstrated a clear graded relationship between ACE exposure and multiple outcomes, including high school absenteeism (Figure 1), asthma prevalence (Figure 2), and poor childhood health (Figure 3). In each case, increasing ACE count was associated with a higher prevalence of negative outcomes, reinforcing the overall pattern that greater exposure to early-life stress is linked to poorer health and behavioral outcomes.

Findings

Table 1. Pooled outcomes across high vs. low early-life stress (ELS)/ACE exposure groups from systematic review of sources, showing mean poor health prevalence, sleep efficiency metrics, epigenetic age acceleration, and NR3C1 methylation levels, alongside resilience asset effects (n > 7,000 total participants) (*Client Challenge*, 2026).

Outcome Metric	High ELS/ACEs (Means/SD)	Low ELS/ACEs (Means/SD)	Resilience Assets Effect
Poor Health %	37% (4+ ACEs)	17.2% (0 ACEs)	↓38.5% (high assets)
Sleep Efficiency	Reduced (SD=2.95; n=861)	Normal timing	Healthy despite EL
Epigenetic Age	Accelerated (PedBE SD=0.85)	Baseline	PCE-buffered
NR3C1 Methylation	Hypermethylated (exon 1F)	Normative	Lower in outliers

Table 2. Collected results comparing variables in relation to ACE count, separated by predetermined threshold. Variables of depressive symptoms, Insomnia symptoms and PCE relate the effects and potential reversibility of ACE consequences (n = 2,841) (Wu et al., 2024).

Outcome Metric	ACE < 4	ACE ≥ 4
Depressive symptoms	1.59 [1.34, 1.89]	1.09 [0.75, 1.59]
Insomnia symptom	1.73 [1.44, 2.09]	1.76 [0.96. 3.23]
PCE	0.93 [0.88, 0.99]	0.82 [0.66, 1.02]

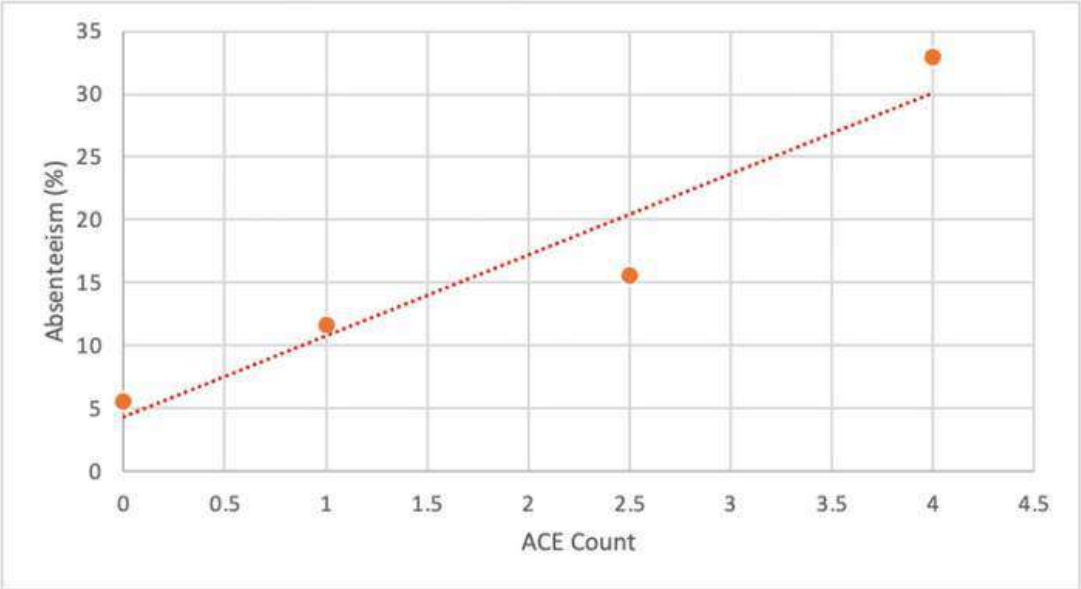


Figure 1. Association between ACE count and high school absenteeism. Scatter plot showing the percentage of participants reporting high school absenteeism across ACE categories (0, 1, 2–3, ≥4). ACE categories were converted to numeric values for visualization, and a linear regression line was fitted to examine the relationship between ACE exposure and absenteeism (*Client Challenge*, 2026).

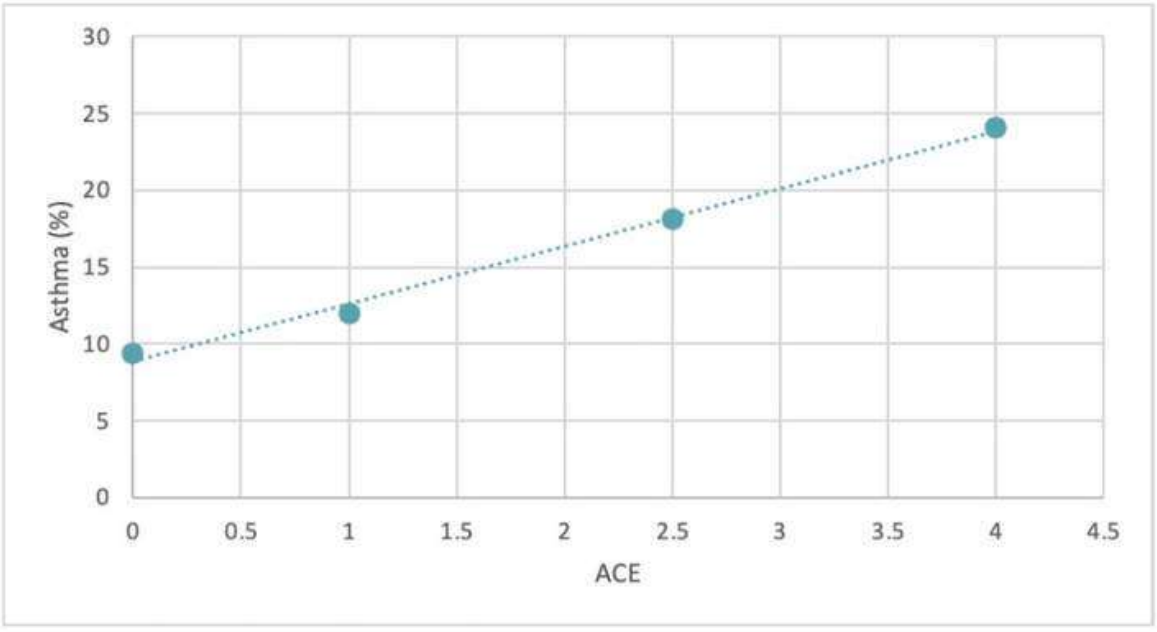


Figure 2. Association between ACE count and asthma prevalence. Scatter plot displaying the percentage of participants reporting asthma across ACE categories (0, 1, 2–3, ≥4). ACE categories were represented numerically, and a linear regression model was applied to assess the relationship between ACE count and asthma prevalence (*Client Challenge*, 2026).

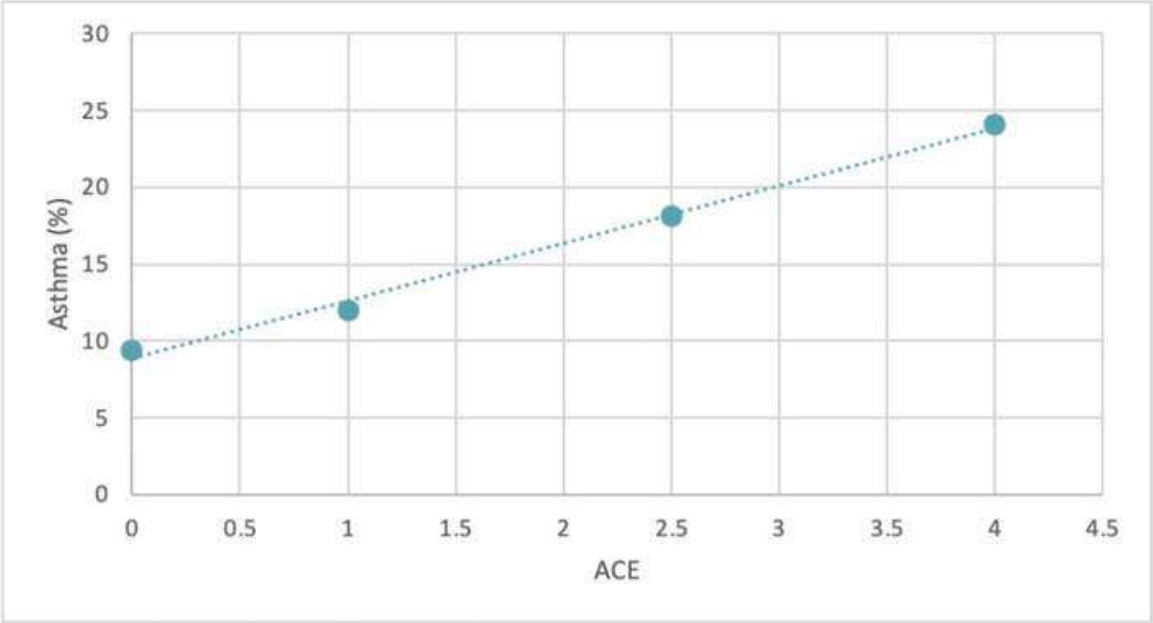


Figure 2. Association between ACE count and asthma prevalence. Scatter plot displaying the percentage of participants reporting asthma across ACE categories (0, 1, 2–3, ≥4). ACE categories were represented numerically, and a linear regression model was applied to assess the relationship between ACE count and asthma prevalence (*Client Challenge*, 2026).

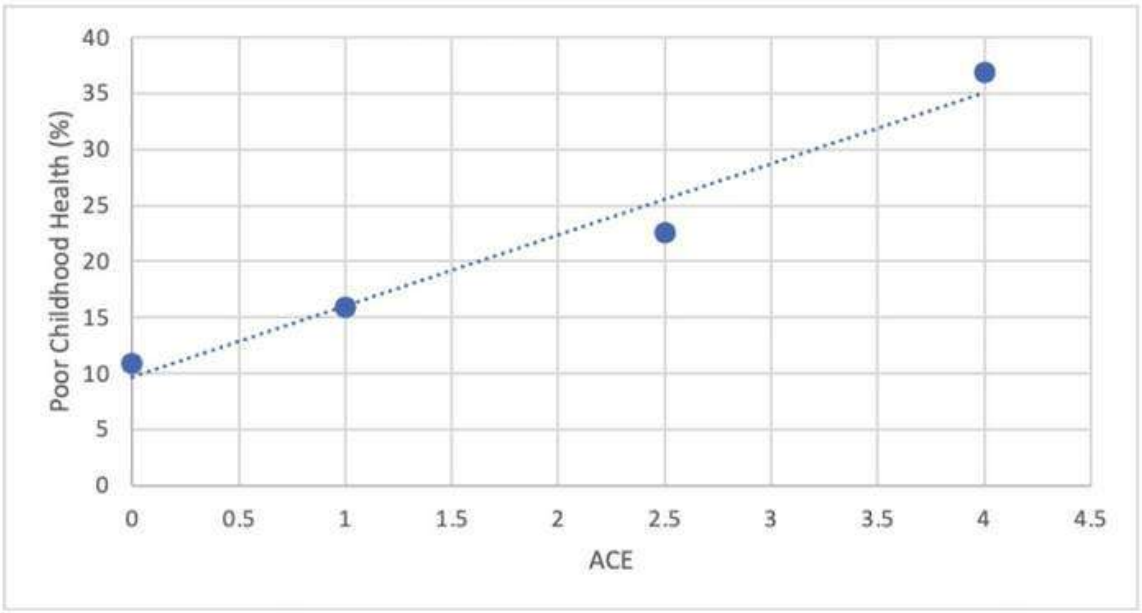


Figure 3. Association between ACE count and poor childhood health. Scatter plot illustrating the percentage of participants reporting poor childhood health across ACE categories (0, 1, 2–3, ≥4). ACE categories were converted to numeric values and a linear regression line was fitted to evaluate the association between ACE exposure and reported childhood health outcomes (*Client Challenge*, 2026).

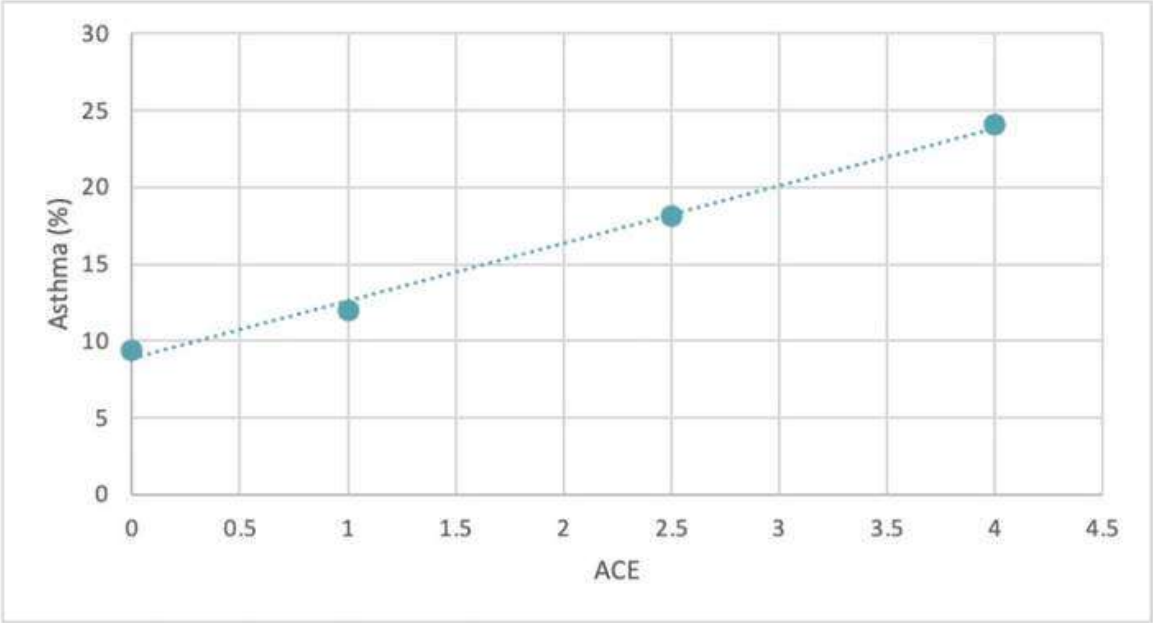


Figure 2. Association between ACE count and asthma prevalence. Scatter plot displaying the percentage of participants reporting asthma across ACE categories (0, 1, 2–3, ≥4). ACE categories were represented numerically, and a linear regression model was applied to assess the relationship between ACE count and asthma prevalence (*Client Challenge*, 2026).

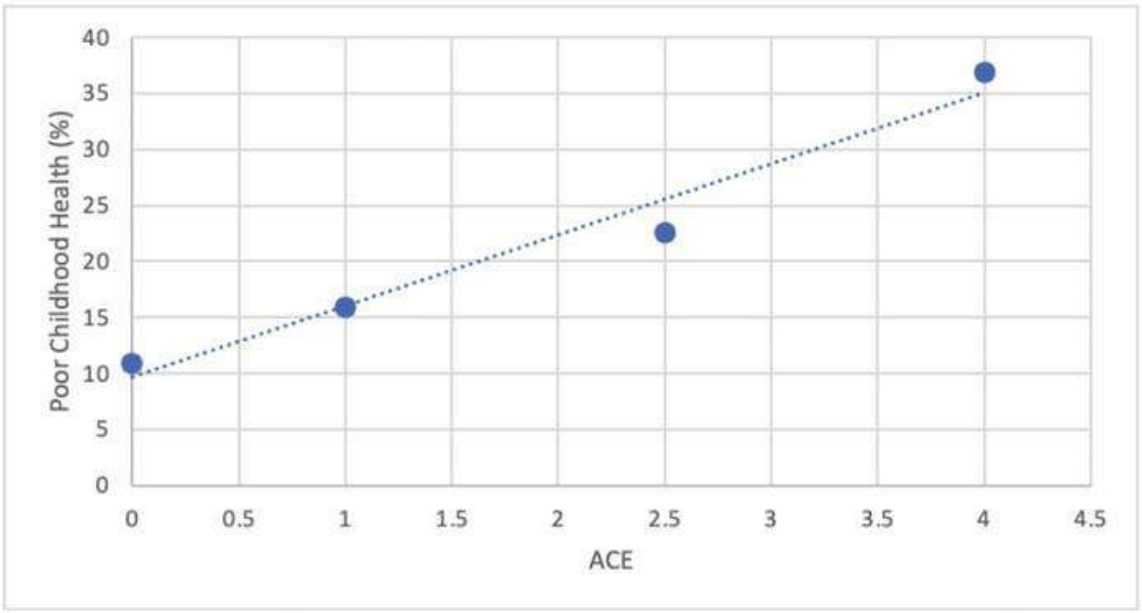


Figure 3. Association between ACE count and poor childhood health. Scatter plot illustrating the percentage of participants reporting poor childhood health across ACE categories (0, 1, 2–3, ≥4). ACE categories were converted to numeric values and a linear regression line was fitted to evaluate the association between ACE exposure and reported childhood health outcomes (*Client Challenge*, 2026).

Discussion

Based on the background research followed by the results focused on outlining the relationship between ACE and PCE, the findings are generally in favor of the hypothesis. Specifically, beyond the established background research highlighting the epigenetic consequences of ACEs on sleep disruption and methylation changes associated with prolonged stress exposure, the compiled datasets demonstrate clear associations between increasing ACE exposure and several health-related outcomes. Linear regression visualizations of selected outcomes, including high school absenteeism, asthma prevalence, and poor childhood health, further illustrate the graded relationship between ACE exposure levels and reported health outcomes across the population datasets analyzed. These patterns align with existing literature suggesting that early-life adversity contributes to cumulative physiological and behavioral risks.

These patterns may be partially explained by underlying biological mechanisms. Early-life stress is associated with epigenetic modifications, particularly hypermethylation of the NR3C1 gene, which can alter glucocorticoid receptor expression and disrupt HPA axis regulation. However, the presence of resilient individuals suggests that these epigenetic changes may not be fixed. Instead, they may be modifiable, allowing for more adaptive stress responses in some individuals despite high ACE exposure (Forum et al., 2025).

However, not all findings aligned with the expected trend. Depressive symptom scores were lower in the ACE ≥ 4 group compared to the ACE < 4 group, contradicting the hypothesis. This may reflect the presence of resilient individuals within the high-ACE group or limitations in the data, such as differences in measurement or self-reported bias. Importantly, this inconsistency points to increased variability in outcomes, suggesting that responses to early-life stress are not uniform (Wu et al., 2024).

This variability, classified as a statistical trend across the previously mentioned categories, suggests the presence of outliers that may represent individuals exhibiting resilience. Some data points display more favorable outcomes than expected based on ACE exposure alone, indicating that certain individuals may develop adaptive responses that mitigate negative effects. These deviations from expected patterns highlight the heterogeneity in responses to early-life stress and reinforce that ACE exposure does not uniformly predict negative outcomes.

Furthermore, the observation that some individuals with ACE ≥ 4 display a greater likelihood of positive childhood experiences (PCEs) than expected suggests a potential buffering or reversal effect. Positive childhood experiences may act as protective factors capable of moderating the biological and behavioral consequences of early adversity. This supports emerging evidence that epigenetic mechanisms, including methylation patterns associated with stress-related genes such as NR3C1, may remain partially reversible under supportive environmental conditions. As a result, the presence of resilient individuals within high ACE exposure groups reinforces the concept that vulnerability and resilience coexist within populations exposed to early-life adversity.

These findings carry direct implications for clinical practice, as it would allow to improve current screening and early interventions. By implementing a standardized 10-item ACE survey, individuals could be screened and a vulnerability profile created, creating a profile that could have tailored interventions aiming and partial methylation reversibility. While the current gap of artificial

demethylation using gene editing technology is constantly shrinking, such implementation would take time to translate to a clinical environment. Thus the proposal for an elaborate screening strategy focused on establishing supportive environments as a means would promote a leading framework that could enhance the current clinical practices in relation to the effects originating from ACE's.

Conclusion

Overall, higher levels of early-life stress were associated with poorer health and sleep outcomes, supporting the hypothesis, but the results also showed clear variability between individuals. The presence of resilient individuals, along with the influence of positive childhood experiences, suggests that the effects of adversity are not fixed and can be shaped by protective factors. Together, these findings point to a more complex relationship between early-life stress and long-term outcomes, where both environmental influences and underlying biological processes play a role.

Future Direction

Despite outlining datasets to further establish the point of resilient individuals through epigenetic differences in contrast to non-resilient individuals, the lack of access to the data, limits the validity of findings to only survey based data. This in turn prevents the ability to generalize and validate our findings. This could be improved by submitting research proposals and initiating a request to be able to view the various genetic data sets we were able to find, as mentioned in our Methods.

Moreover, our main source of data is primarily collected in the region of Taiwan, thus not discussing how geographic difference could affect the reliability or generalizability of the results in applied for other regions. As the findings are heavily reliant on epigenetic based outcomes, drastically different factors such as lifestyle, diet as well as environmental exposure produce distinctive conditions which limits transferability of the finding to the North American region, which most of this papers background and introduction is derived from.As such, validation regarding the transferability of these findings through past comparison of data collection in both regions under one epigenetic study, would be advisable.

References

Bakusic, J., Vrieze, E., Ghosh, M., Bekaert, B., Claes, S., & Godderis, L. (2020). Increased methylation of NR3C1 and SLC6A4 is associated with blunted cortisol reactivity to stress in major depression. Neurobiology of Stress, 13, Article 100272. https://doi.org/10.1016/j.ynstr.2020.100272

Borçoi, A. R., Mendes, S. O., Dos Santos, J. G., de Oliveira, M. M., Moreno, I. A. A., Freitas, F. V., Silva, M. A., & Álvares-da-Silva, A. M. (2020). Risk factors for depression in adults: NR3C1 DNA methylation and lifestyle association. Journal of Psychiatric Research, 121, 24–30. https://doi.org/10.1016/j.jpsychires.2019.11.002

Client Challenge. (2026). Springermedizin.de. https://www.springermedizin.de/adverse-childhood-experiences-and-sources-of-childhood-resilienc/15882250

Cicchetti, D., & Handley, E. D. (2017). Methylation of the glucocorticoid receptor gene, nuclear receptor subfamily 3, group C, member 1 (NR3C1), in maltreated and nonmaltreated children: Associations with behavioral undercontrol, emotional lability/negativity, and externalizing and internalizing symptoms. Development and Psychopathology, 29(5), 1795–1806. https://doi.org/10.1017/S095457941700140X

Daskalakis, N. P., Ruck, C., Nievergelt, C. M., Maihofer, A. X., Logue, M. W., Guintivano, J., Vinkers, C. H., Vermetten, E., Rutten, B. P., Bin Ali, O. T., Radovic-Magno, A. P., Stein, M. B., Ressler, K. J., Uddin, M., McLean, S. A., Wildman, D. E., Liberzon, I., Galea, S., & Binder, E. B. (2015). Environmental and genetic determinants of DNA methylation in the glucocorticoid receptor gene (NR3C1). Biological Psychiatry, 77(11), 940–948. https://doi.org/10.1016/j.biopsych.2014.07.022

DiMarzio, K., Rojo-Wissar, D. M., Hernandez Valencia, E., Ver Pault, M., Denherder, S., Lopez, A., Lerch, J., Metrailer, G., Merrill, S. M., Highlander, A., & Parent, J. (2024). Childhood Adversity and Adolescent Epigenetic Age Acceleration: The Role of Adolescent Sleep Health. *MedRxiv : The Preprint Server for Health Sciences*, 2024.09.02.24312939. https://doi.org/10.1101/2024.09.02.24312939

Dempster, E. L., Burcescu, I., Wigg, K., Kiss, E., Baji, I., Gadoros, J., ... & International Consortium for Childhood-Onset Mood Disorders. (2007). Evidence of an association between the vasopressin V1b receptor gene (AVPR1B) and childhood-onset mood disorders. Archives of general psychiatry, 64(10), 1189-1195.

Forum, D. M. K., Bjerregaard, C., & Thomsen, P. H. (2025). The significance of DNA methylation of the *NR3C1* gene encoding the glucocorticoid receptor for developing resilience in individuals exposed to early life stress. *Nordic Journal of Psychiatry*, *79*(1), 1–14. https://doi.org/10.1080/08039488.2024.2436987

Jang, H. S., Shin, W. J., Lee, J. E., & Do, J. T. (2017). CpG and non-CpG methylation in epigenetic gene regulation and brain function. Genes, 8(6), Article 148. https://doi.org/10.3390/genes8060148

Liu, P. Z., & Nusslock, R. (2018). How stress gets under the skin: Early life adversity and glucocorticoid receptor epigenetic regulation. Current Genomics, 19(8), 653–664. https://doi.org/10.2174/1389202919666171130151801

Lopez, M., Ruiz, M. O., Rovnaghi, C. R., Tam, G. K-Y., Hiscox, J., Gotlib, I. H., Barr, D. A., Carrion, V. G., & Anand, K. J. S. (2021). The social ecology of childhood and early life adversity. *Pediatric Research*, *89*(2).

Murata, Y., Fujii, A., Kanata, S., Fujikawa, S., Ikegame, T., Nakachi, Y., Zhao, Z., Kohda, K., Kasai, K., & Iwamoto, K. (2019). Evaluation of the usefulness of saliva for DNA methylation analysis in cohort studies. Neuropsychopharmacology Reports, 39(4), 301–305. https://doi.org/10.1002/npr2.12073

Papadopoulos, A. S., & Cleare, A. J. (2012). Early life stress and HPA axis changes in CFS. *Nature Reviews Endocrinology*, *8*(8), 502–502. https://doi.org/10.1038/nrendo.2011.153-c2

Roy, B., Shelton, R. C., & Dwivedi, Y. (2017). DNA methylation and expression of stress related genes in PBMC of MDD patients with and without serious suicidal ideation. Journal of Psychiatric Research, 89, 115–124. https://doi.org/10.1016/j.jpsychires.2017.02.015

Turecki, G., Ota, V. K., Belangero, S. I., Jackowski, A., & Kaufman, J. (2014). Early life adversity, genomic plasticity, and psychopathology. *The Lancet Psychiatry*, *1*(6), 461–466. https://doi.org/10.1016/s2215-0366(14)00022-4

van der Knaap, L. J., Riese, H., Hudziak, J. J., Verbiest, M. M. P. J., Verhulst, F. C., Oldehinkel, A. J., & van Oort, F. V. A. (2014). Glucocorticoid receptor gene (NR3C1) methylation following stressful events between birth and adolescence. The TRAILS study. *Translational Psychiatry*, *4*(4), e381–e381. https://doi.org/10.1038/tp.2014.22

Wu, M.-H., Chiao, C., & Lin, W.-H. (2024). Adverse childhood experience and persistent insomnia during emerging adulthood: do positive childhood experiences matter? *BMC Public Health*, *24*(1). https://doi.org/10.1186/s12889-024-17774-w

In Silico Identification of Cas12a crRNA Targets in the HIV-1/SIVcpz *pol* Region

Tahreem Sheikh, Areesha Baig, Isra Akhter

Abstract

Human Immunodeficiency Virus (HIV) is a global health challenge, impacting millions everyday. The management of the virus highlights the importance of molecular technologies to detect and target viral sequences. This study aimed to identify conserved HIV-1 sequences suitable for CRISPR-Cas12a targeting to be used as a diagnostic tool with minimal off-target interactions with the human genome. Publicly available pre-aligned HIV-1 sequences were obtained from the Los Alamos National Laboratory. The sequences were retrieved from the highly conserved region in the HIV-1 pol gene due to its low mutability rate. Despite several PAM distributed through the sequence, only 13 crRNA targets were considered viable considering GC content and secondary structure formation, which was determined using the RNAfold program. The candidate crRNAs were then scanned against the reference human genome (GRCh38) to find the ideal target crRNA with minimal off-target matches using BLAST. From this search, a sequence of 23 nucleotides with 22.87% hairpin formation was considered a viable target. However, this sequence had a 57% match to the reference human genome. Although the cleavage site of Cas12a is outside the off-target match, the human genome should still be scanned for a PAM near the matched sequence. The interaction between the Cas12a enzyme and the sequence was then modeled using UCSF Chimera.

Introduction/Rationale

Human immunodeficiency viruses (HIV) are rampant viruses with various strains that

have affected millions worldwide, with approximately 30 million people living with the infection (Deeks et al., 2015). It is transmitted via infected human bodily fluids, such as blood or sexual fluids. HIV primarily targets CD4+ T-cells, an activating cell that is responsible for coordinating the body's adaptive immune responses (Kirchhoff, 2013). CD4+ T-cells activate macrophages, B-cells, and CD8+ T-cells through cytokine signalling, allowing the immune system to enact effective responses to pathogens (Kirchhoff, 2013). By depleting these cells, HIV severely compromises the body's defense system, leaving infected individuals vulnerable to various viruses and infections. If left untreated, it can progress to acquired immunodeficiency syndrome (AIDS), characterized by leaving the individual susceptible to opportunistic infections and cancers (Deeks et al., 2015).

As of today, HIV has no cure. Antiretroviral therapy (ART) has transformed HIV infection into a manageable diagnosis, almost entirely preventing HIV replication and reducing the chance of developing AIDS. However, if one was to stop taking the medication, the virus would rebound within weeks. Regardless, immediate usage of ART is essential for preventing further development of HIV and limiting the risk of spreading the infection to uninfected partners. Early detection and viral monitoring are therefore critical in improving patient outcomes. Rapid HIV methods, such as blood from a finger-stick or swabbing saliva from the mouth, while useful, have limited sensitivity for detecting acute HIV infection (Deeks et al., 2015).

Clustered Regularly Integrated Short Palindromic Repeats, CRISPR, and CRISPR-associated (cas) genes provide bacteria with their defense against phages. CRISPR arrays undergo transcription and processing to produce tiny CRISPR RNAs (crRNA) that associate with cas proteins to create crRNA-guided surveillance complexes (Bondy-Denomy et al., 2015). While typically defined as a gene editing technology derived from bacterial systems, for the purpose of this experiment, CRISPR will be used as a molecular detector

rather than a gene-editing tool. This will not treat or remove HIV, but instead act as a method for early detection and diagnosis. It can be used to measure the amount of virus present and whether the prescribed treatment is working. To do this, CRISPR will bind to HIV-1 crRNA sequences and cut nearby reporters, releasing a signal that can be detected. CRISPR requires a CRISPR-associated protein (Cas) to perform the actual cleavage of DNA. For HIV detection, the associated protein is typically Cas12a (Tong et al., 2021; Bagi et al., 2025). Cas12a requires that the crRNA sequences contain a PAM; this is because CRISPR is a system derived from bacterial immune systems which evolved to cleave and destroy viruses through binding of Cas proteins. The PAM allow the CRISPR system to identify only foreign viral DNA, unwind it, and cleave if the crRNA spacer sequence is complementary to the viral target sequence (Horvath & Barrangou, 2010; Hillary & Ceasar, 2022).

This study aims to propose CRISPR–Cas12a as a diagnostic tool for the detection of HIV-1 RNA. This study hypothesizes that the conserved polymerase consensus sequences from HIV-1 can identify structurally viable crRNA targets across a variety of strains with minimal off-target matches to the reference human genome (GRCh38).

Methods

First, candidate crRNA sites were found for Cas12a on February 12, 2026. A pre-made alignment of HIV-1/SIVcpz on the Los Alamos National Laboratory of HIV sequences was used to find ancestral/consensus polymerase sequences (Los Alamos National Laboratory, n.d.). The pre-made alignment retrieved a consensus sequence of 4.2 kb (Figure 1A). Using the CRISPOR tool, this consensus was scanned for candidate crRNAs (CRISPOR, n.d.). An ideal candidate had a protospacer adjacent motif (PAM), 23 base pairs following the PAM, medium GC content, and no long AT stretches. The PAM for a Cas12a enzyme is TTTV, where V can be A, G, or C (Bandyopadhyay et al., 2020). Then, using

RNAfold, the secondary structures of the sequences were determined at 37°C, and the candidate list was further shortlisted (Lorenz et al., 2011). Sequences with a minimum free energy structure that formed a hairpin less than 60% of the time were shortlisted as candidates.

Then, on February 13, 2026, to minimize off-target matches, the DNA-equivalent sequences were screened against the reference human genome (GRCh38) using BLASTn (National Centre for Biotechnology Information, n.d.), a tool that searches a nucleotide database with a nucleotide query (Figure 2A). The database chosen, human genomic and transcript databases (Human G+T), is curated from human reference sequences, which explicitly includes GRCh38. Due to NCBI BLAST limitations, the lowest word size permitted was 7, resulting in an inexact screening that detected extended sequence similarities. All other algorithm parameter values were optimized for somewhat similar sequences (blastn- short), an expected threshold of 1000, with low-complexity regions off, and match/mismatch scores of 1, -3. The sequences were compared to one another graphically using GC content, hairpin formation, query coverage, and E-value.

On February 15, the three-dimensional structure of the catalytically inactive Francisella novicida Cas12a (FnCas12a) in complex with a crRNA guide and a double-stranded DNA target was retrieved from the RCSB Protein Data Bank (PDB ID: 5B43) and used as the structural template. The crystal structure was determined by X-ray diffraction at a resolution of 2.65 Å, providing sufficient atomic detail for visualization. No mutations were present in the structure; therefore, the sequence corresponds to the actual wild-type FnCas12a protein as expressed in *Escherichia coli*. To prepare the data for modelling, water and non-relevant small molecules, including small ligands and biomolecules, were removed to reduce steric interference and improve visual clarity. X-ray diffraction, the method used to determine the structure, relies on the scattering of X-rays by

electrons in the crystal lattice to reconstruct an electron density map, allowing precise modelling of atomic positions (Kristo, 2012). Further discrepancies between the structure and energy structure that formed a hairpin less than 60% of the time were shortlisted as candidates.

Then, on February 13, 2026, to minimize off-target matches, the DNA-equivalent sequences were screened against the reference human genome (GRCh38) using BLASTn (National Centre for Biotechnology Information, n.d.), a tool that searches a nucleotide database with a nucleotide query (Figure 2A). The database chosen, human genomic and transcript databases (Human G+T), is curated from human reference sequences, which explicitly includes GRCh38. Due to NCBI BLAST limitations, the lowest word size permitted was 7, resulting in an inexact screening that detected extended sequence similarities. All other algorithm parameter values were optimized for somewhat similar sequences (blastn- short), an expected threshold of 1000, with low-complexity regions off, and match/mismatch scores of 1, -3. The sequences were compared to one another graphically using GC content, hairpin formation, query coverage, and E-value.

On February 15, the three-dimensional structure of the catalytically inactive Francisella novicida Cas12a (FnCas12a) in complex with a crRNA guide and a double-stranded DNA target was retrieved from the RCSB Protein Data Bank (PDB ID: 5B43) and used as the structural template. The crystal structure was determined by X-ray diffraction at a resolution of 2.65 Å, providing sufficient atomic detail for visualization. No mutations were present in the structure; therefore, the sequence corresponds to the actual wild-type FnCas12a protein as expressed in *Escherichia coli.* To prepare the data for modelling, water and non-relevant small molecules, including small ligands and biomolecules, were removed to reduce steric interference and improve visual clarity. X-ray diffraction, the method used to determine the structure, relies on the scattering of X-rays by

electrons in the crystal lattice to reconstruct an electron density map, allowing precise modelling of atomic positions (Kristo, 2012). Further discrepancies between the structure and HIV-targeted crRNA sequences were accounted for by substituting the HIV-1-pol-derived spacer sequences while maintaining the PAM to ensure accuracy for subsequent modelling. Modelling of the structure was performed using UCSF Chimera. The FnCas12a structure was selected because it contains the complete Cas12a-crRNA-DNA complex, including the PAM-interacting domain and RuvC catalytic site, enabling accurate evaluation of crRNA-target DNA binding and structural compatibility.

Results/Findings

Table 1. crRNA candidate sequences

Sequence Number	Sequence	Hairpin formation at 37 °C?	Shortlisted?
1	*TTTG* GCAACGACCCCTCG TCACAATAA	41.31%	Yes
2	*TTTC* CATCTTCCTGGCAA ACTCATTTC	0%	Yes
3	*TTTC* CCATTAGCCCTATTG AGACTGTA	0%	Yes
4	*TTTA* CTGGTACAGTCTCA ATAGGGCTA	62.24%	No
5	*TTTG* GGCCATCCATTCCTG GCTTTAAT	72.08%	No
6	*TTTC* CATCCCTGTGGAAG CACATTGTA	73.37%	No
7	*TTTG* GAATATTGCTGGTGA TCCTTTCC	56.37%	Yes
8	*TTTA* TCAGGATGGAGTTC ATAACCCAT	48.13%	Yes
9	*TTTC* TGCCAGTTCTAGCTC TGCTTCTT	22.87%	Yes
10	*TTTA* GGAGTCTTTCCCCAT ATTACTAT	41.62%	Yes

11	*TTTC* TGCTCCTACTATGGG TTCTTTCT	33.86%	Yes
12	*TTTA* TCTAGCTTTGCAGG ATTCGGGAT	23.79%	Yes
13	*TTTA* TTACAGGGACAGCA GAAATCCAC	58.29%	Yes
14	*TTTG* GAAAGGACCAGCAA AGCTCCTCT	31.64%	Yes
15	*TTTC* CAAAGTGGATTTCT GCTGTCCCT	46.66%	Yes
16	*TTTC* CAGAGGAGCTTTGC TGGTCCTTT	30.59%	Yes

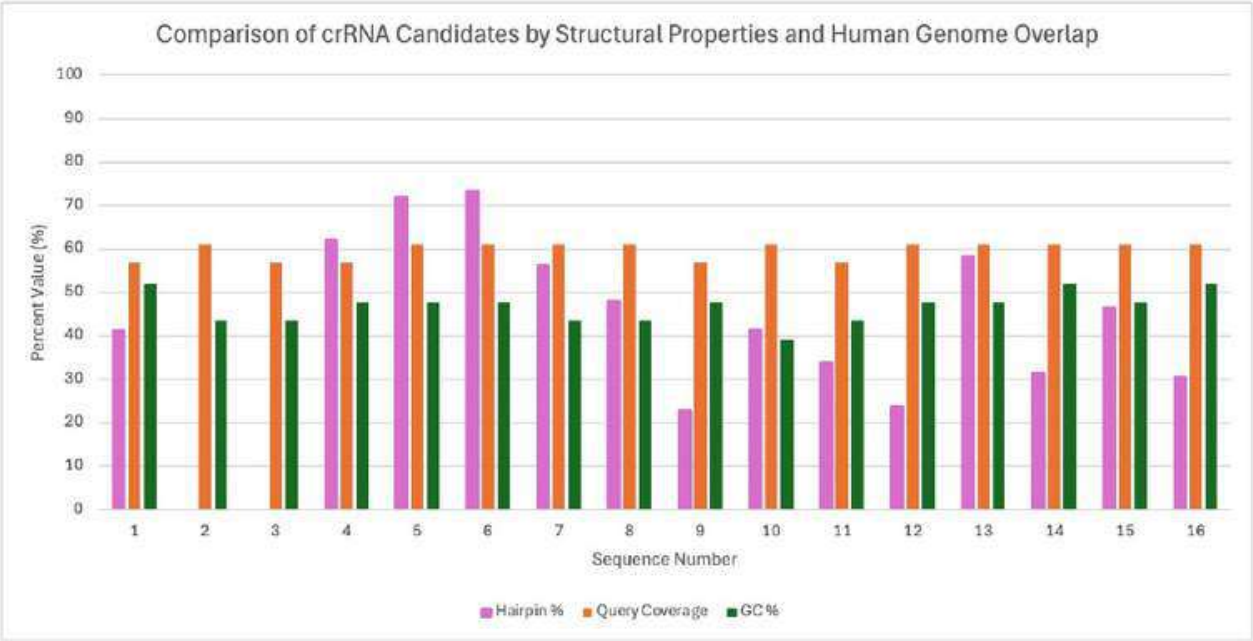


Figure 1. Comparison of GC content (green), hairpin formation (pink), and query coverage (orange) across 16 crRNA candidates derived from the HIV-1 *pol* consensus sequence. GC% and hairpin% reflect structural suitability for Cas12a binding, while query coverage reflects the degree of similarity to the reference human genome (GRCh38).

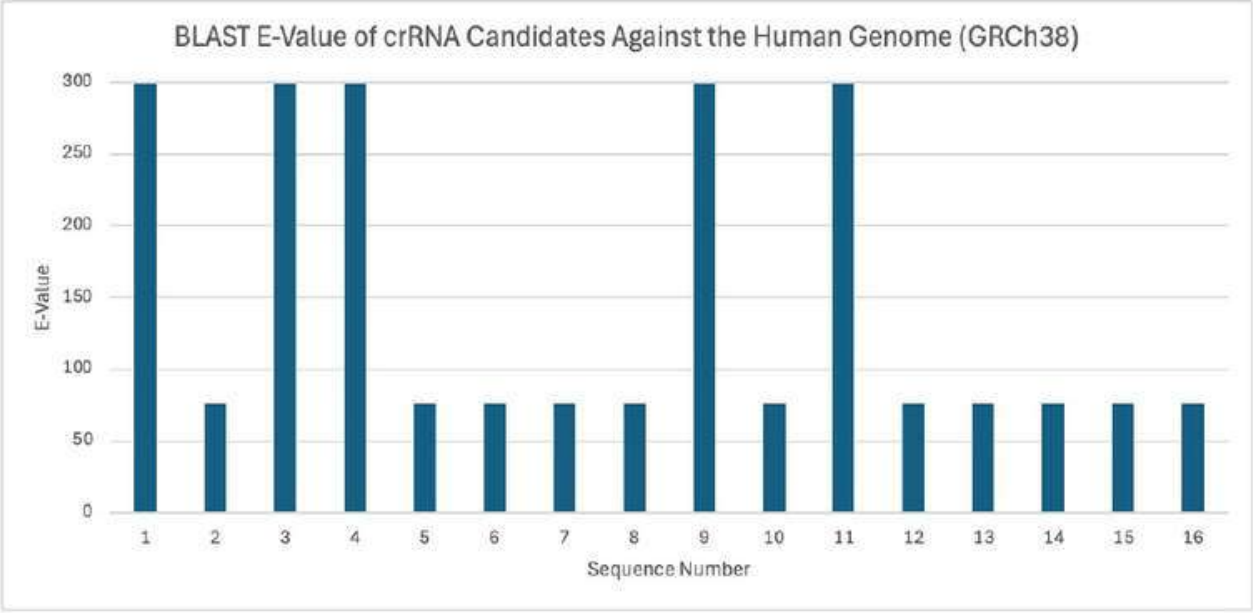


Figure 2. BLAST E-values for 16 crRNA candidates screened against the reference human genome (GRCh38). A higher E-value indicates a weaker match to the human genome, reflecting greater sequence specificity for HIV-1. Sequences 1, 3, 4, 9, and 11 returned the

highest E-value (299), with Sequence 9 selected as the optimal candidate based on its combined structural and specificity scores.

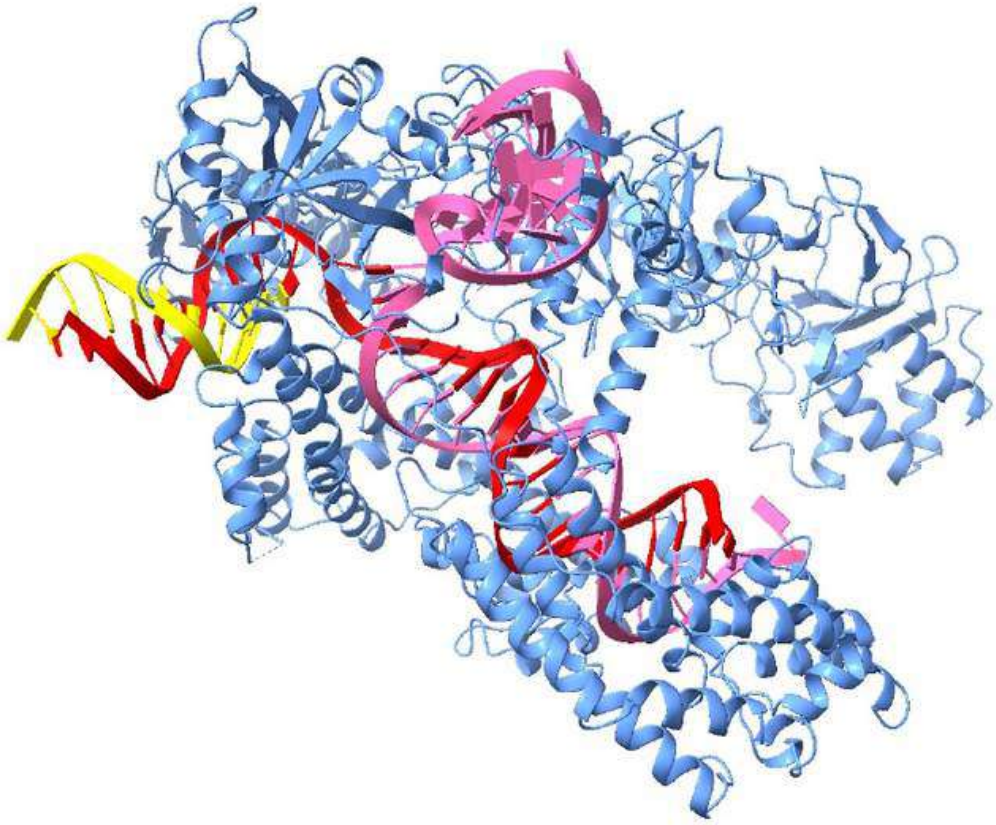


Figure 3. Crystal structure of the Cas12a-crRNA-target DNA complex (PDB ID: 5B43). Cas12a protein (Chain A, blue) is shown bound to crRNA (Chain B, pink), target DNA (Chain C, red), and the complementary strand of target DNA (Chain D, yellow).

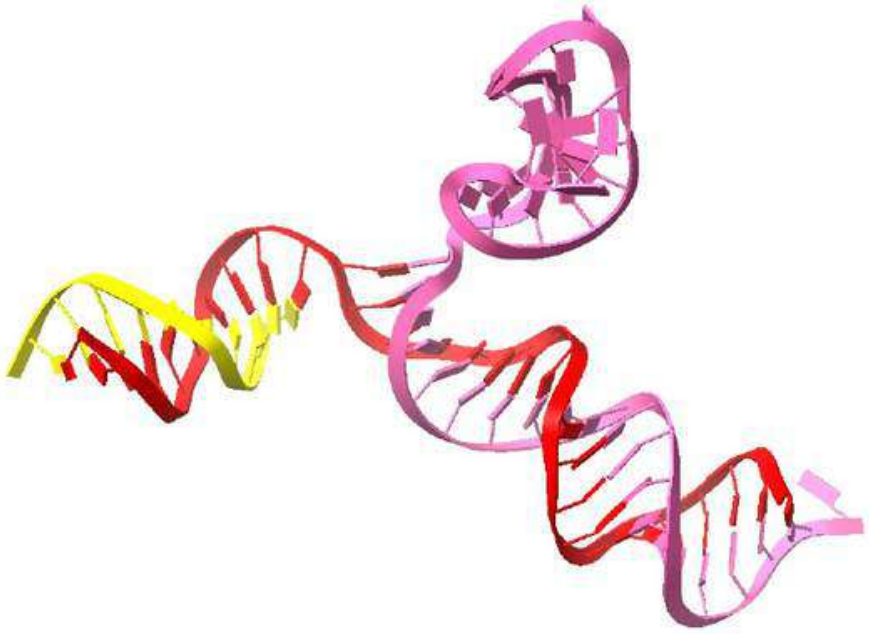


Figure 4. Interaction between the crRNA and the target DNA in the Cas12a-crRNA-target DNA complex (PDB ID: 5B43). Cas12a is hidden for clarity, allowing visualization of the cRRNA base-pairing with the target DNA duplex containing double-stranded DNA. This view emphasizes the RNA-DNA recognition interface.

Discussion

This study aimed to computationally identify ideal crRNA candidates for HIV detection using CRISPR-Cas12a. The sequences were derived from a conserved HIV-1 polymerase consensus sequence. This is because previous studies show that the pol region has essential enzymatic roles and high conservation across HIV-1 strains (Morris et al., 2005). The conservativity makes it ideal for application across a wide variety of strains. Specifically, the HIV pol gene encodes for reverse transcriptase, integrase, and protease, all necessary enzymes for viral replication. This means that mutations in pol are absolutely lethal to the genome. Evolutionary flexibility is limited due to these constraints. Other HIV regions, such as env, vif, tat, and nef are not as robust as the pol region (Keating et al., 2008). Typically, mature crRNAs of Cas12a are 42-44 nucleotides in length, containing a repeat

sequence of 19-20 nucleotides and a spacer sequence that is 23-25 nucleotides (Paul & Montoya, 2020). Due to this, candidate spacer sequences of 23 nucleotides following the PAM were chosen. The spacer sequences were then scanned to maintain 40-60% GC content to minimize secondary hairpin structure formations (Morris et al., 2005; Creutzburg et al., 2020). The conserved pol consensus sequence derived was 4.2 kb with several PAM scattered through, indicating that Cas12a does not target a single locus on the genome, but instead a wide variety of sequences across the genome, meaning that targeting can be multiplexed across functionally constrained regions of the genome. With the filters mentioned above, 16 candidate crRNAs were identified (Table 1), suggesting that the pol region is suitable for Cas12a target sites. These sequences were computationally tested for structural viability, ensuring that hairpin formation occurred less than 60% of the time. This parameter was set due to the fact that the binding of Cas12a overrides moderate secondary structures. With this extra guideline, 13 crRNA candidates were deemed viable. This reinforces the idea that Cas12a diagnostic design is feasible within the conserved polymerase consensus region of HIV-1. It is important to remember that computational filtering can narrow candidate targets; however, computational predictions cannot effectively describe functional Cas12a activity in vitro. Another limitation to this screening is the fact that the parameter of 60% hairpin formation was set arbitrarily, with the knowledge that Cas12a binding overrides moderate secondary structural formations. Cas12a unwinds its target DNA processively, displacing moderate secondary structures during the seed region scanning process (Creutzburg et al., 2020). Sequences forming hairpins more than 60% of the time at physiological temperature are likely to be predominantly folded under assay conditions, reducing effective concentration of the open, bindable form. Below 60%, the equilibrium favours enough open conformation that Cas12a binding is plausible. Additionally, this limitation has minimal bearing on the final outcome considering that the selected sequence (sequence 9) has a

for ranking candidates by structural viability (National Centre for Biotechnology Information, n.d.). Using these parameters, sequence 9 was chosen as the best crRNA target and modelled using UCSF Chimera. However, sequence 9 had a 57% match to the reference human genome. Due to the fact that the human genome was not scanned for nearby PAM, the selected crRNA target is limited in its functionality (Bandyopadhyay et al., 2020). The Cas12a-crRNA-target DNA complex model substantiated the compatibility of the sequences in the detection framework and also showed guide target recognition.

For future studies, it is recommended to extend the screening of candidates to the gag region of the HIV-1 genome. This is because prior cross-clade studies have shown that gag and pol are two highly conserved regions that can limit mutation tolerance (Morris et al., 2005). The gag region can be screened for crRNA targets with high structural suitability and lower off-target matching. If better crRNA targets are found, this experiment can be expanded to an in vitro analysis of CRISPR-Cas12a with the crRNA targets to experimentally verify cleavage results. Then, this system can be used to determine the viral load of HIV-1 RNA, acting as a scale to govern the effectiveness of prescribed treatment.

Ethical Integrity Statement

We confirm that this submission is our original work completed without the use of AI tools, that all sources are properly cited in APA 7 format, and that we understand BACSA is not responsible for any plagiarism or academic misconduct.

References

Bagi, M., Jamalzadegan, S., Steksova, A., & Wei, Q. (2025). CRISPR–CAS based platforms for RNA detection: Fundamentals and applications. *Chemical Communications*, *61*(72), 13571–13600. https://doi.org/10.1039/d5cc03257a

Bandyopadhyay, A., Kancharla, N., Javalkote, V. S., Dasgupta, S., & Brutnell, T. P. (2020). CRISPR-CAS12A (Cpf1): A versatile tool in the plant genome editing tool box for Agricultural Advancement. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.584151

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M. F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K. L., & Davidson, A. R. (2015). Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature*, *526*(7571), 136–139. https://doi.org/10.1038/nature15254

Creutzburg, S. C., Wu, W. Y., Mohanraju, P., Swartjes, T., Alkan, F., Gorodkin, J., Staals, R. H., & van der Oost, J. (2020). Good guide, bad guide: Spacer sequence-dependent cleavage efficiency of cas12a. *Nucleic Acids Research*, *48*(6), 3228–3243. https://doi.org/10.1093/nar/gkz1240

CRISPOR. (n.d.). CRISPOR: Design and evaluate CRISPR/Cas guide RNAs. Retrieved February 12, 2026, from https://crispor.gi.ucsc.edu/

Deeks, S. G., Overbaugh, J., Phillips, A., & Buchbinder, S. (2015). HIV infection. *Nature Reviews Disease Primers*, *1*(1). https://doi.org/10.1038/nrdp.2015.35

References

Bagi, M., Jamalzadegan, S., Steksova, A., & Wei, Q. (2025). CRISPR–CAS based platforms for RNA detection: Fundamentals and applications. *Chemical Communications*, *61*(72), 13571–13600. https://doi.org/10.1039/d5cc03257a

Bandyopadhyay, A., Kancharla, N., Javalkote, V. S., Dasgupta, S., & Brutnell, T. P. (2020). CRISPR-CAS12A (Cpf1): A versatile tool in the plant genome editing tool box for Agricultural Advancement. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.584151

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M. F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K. L., & Davidson, A. R. (2015). Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature*, *526*(7571), 136–139. https://doi.org/10.1038/nature15254

Creutzburg, S. C., Wu, W. Y., Mohanraju, P., Swartjes, T., Alkan, F., Gorodkin, J., Staals, R. H., & van der Oost, J. (2020). Good guide, bad guide: Spacer sequence-dependent cleavage efficiency of cas12a. *Nucleic Acids Research*, *48*(6), 3228–3243. https://doi.org/10.1093/nar/gkz1240

CRISPOR. (n.d.). CRISPOR: Design and evaluate CRISPR/Cas guide RNAs. Retrieved February 12, 2026, from https://crispor.gi.ucsc.edu/

Deeks, S. G., Overbaugh, J., Phillips, A., & Buchbinder, S. (2015). HIV infection. *Nature Reviews Disease Primers*, *1*(1). https://doi.org/10.1038/nrdp.2015.35

Multimodal Prediction of Alzheimer's Disease Conversion in in Mild Cognitive Impairment: Integrating Cognitive Assessments, APOE Genotype, and Plasma Biomarkers to Predict Alzheimer’s Disease Conversion

Siya Gera, Shirley Zhang, Nivedita Nair

Abstract

Alzheimer’s disease (AD) progression from mild cognitive impairment (MCI) involves complex biological mechanisms including amyloid deposition, tau pathology, and neurodegeneration. While cognitive testing is commonly used for prognosis, blood-based biomarkers may improve early prediction. This study investigated whether integrating cognitive assessments, APOE ε4 genotype, and plasma biomarkers improves prediction of progression from MCI to AD. Baseline data from 767 MCI participants in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) were analyzed, with 201 individuals having complete plasma biomarker data. Logistic regression modeling incorporated ADAS13, CDRSB, MMSE (cognitive scales), APOE ε4 count (genetic risk), plasma pTau217, Aβ42/Aβ40 ratio, neurofilament light (NfL) (axonal damage), and GFAP (inflammation). The multimodal model achieved an accuracy of approximately 0.69 and an AUC of 0.79, demonstrating good discrimination between converters and non-converters. CDRSB, NfL, and pTau217 emerged as the strongest predictors of progression. The amyloid ratio showed a protective effect, while APOE ε4 provided moderate predictive value. These findings demonstrate that integrating plasma biomarkers with cognitive measures enhances prognostic accuracy and highlights the central role of neurodegeneration and tau pathology in early AD transition.

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide, accounting for approximately 60–70% of all dementia cases (Alzheimer's Association, 2025). Pathologically, AD is defined by the extracellular accumulation of amyloid-β (Aβ) plaques and intracellular aggregation of hyperphosphorylated tau protein forming neurofibrillary tangles (Jack et al., 2018). These pathological processes are accompanied by synaptic dysfunction, neuronal loss, and progressive brain atrophy, ultimately resulting in cognitive and functional decline (De Strooper & Karran, 2016).

Mild cognitive impairment (MCI) represents a transitional stage between normal aging and dementia. Individuals with MCI exhibit measurable cognitive decline that exceeds age-related expectations but retain relative independence in daily functioning (Petersen, 2018). However, MCI is clinically heterogeneous: while some individuals remain stable or revert to normal cognition, others progress to Alzheimer's disease at varying rates. Annual conversion rates from MCI to AD range between 10–15% in clinical cohorts (Albert et al., 2011). The ability to accurately predict which individuals will convert is therefore critical for early intervention, risk stratification, and clinical trial enrichment.

Traditional prognostic approaches rely heavily on neuropsychological assessments such as the Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS13), and Clinical Dementia Rating – Sum of Boxes (CDRSB). These instruments quantify global cognition and functional impairment and are widely used in both clinical and research settings. However, cognitive tests measure downstream clinical manifestations of disease rather than the underlying biological pathology driving neurodegeneration (Sperling et al., 2011). As a result, reliance on cognitive measures alone may limit early predictive sensitivity.

Advances in biomarker research have transformed understanding of AD pathophysiology. The amyloid cascade hypothesis posits that Aβ accumulation initiates downstream tau pathology and

neurodegeneration (Hardy, 2006). More recently, the AT(N) framework has categorized AD biomarkers into amyloid (A), tau (T), and neurodegeneration (N) groups, emphasizing a biologically defined disease model. Plasma biomarkers have emerged as minimally invasive tools capable of capturing these pathological domains. Phosphorylated tau (pTau217) reflects tau aggregation and correlates strongly with brain tau pathology (Palmqvist et al., 2020). Neurofilament light (NfL) serves as a marker of axonal injury and neurodegeneration (Mattsson et al., 2017). Glial fibrillary acidic protein (GFAP) indicates astroglial activation and neuroinflammatory processes (Shir et al., 2022) Additionally, the plasma Aβ42/Aβ40 ratio reflects cerebral amyloid burden, with lower ratios associated with increased amyloid deposition.

Genetic risk further contributes to disease susceptibility. The apolipoprotein E (APOE) ε4 allele is the strongest common genetic risk factor for late-onset AD, influencing amyloid accumulation and disease progression (Corder et al., 1993). However, genotype reflects inherited risk rather than current disease activity and may not fully capture short-term conversion risk.

From a computational perspective, machine learning models offer the capacity to integrate heterogeneous data types (including clinical, genetic, and biomarker features) into unified predictive frameworks. Logistic regression and ensemble learning methods have been applied to AD classification and progression modeling with promising results (Moradi et al., 2015). Nonetheless, many prior studies have focused on either cognitive variables or neuroimaging biomarkers, while fewer investigations have evaluated fully integrated multimodal models incorporating blood-based biomarkers alongside clinical and genetic predictors.

Although individual biomarker domains have demonstrated predictive value, there remains a need to determine whether combining cognitive assessments, genetic risk, and plasma biomarkers meaningfully improves prediction of progression from MCI to Alzheimer's disease within a single interpretable model. In particular, it remains unclear which biological domains, amyloid deposition, tau pathology, or neurodegeneration, contribute most strongly to short-term clinical transition.

This study aims to determine whether integrating cognitive assessments, APOE ε4 genotype,

and plasma biomarkers improves prediction of progression from mild cognitive impairment to Alzheimer's disease. We hypothesize that a multimodal model incorporating both clinical and biological features will outperform models relying solely on cognitive measures, and that markers of neurodegeneration and tau pathology will demonstrate stronger predictive influence than amyloid burden alone.

Methods

Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, a longitudinal multicenter study designed to investigate clinical and biological markers of Alzheimer's disease progression. Data was accessed on January 15, 2026. Participants were included if they had a baseline diagnosis of mild cognitive impairment (MCI), documented longitudinal follow-up diagnostic status, available baseline cognitive assessment scores, and APOE genotype data. A total of 767 individuals met these criteria. Among these participants, 285 (37.1%) progressed from MCI to Alzheimer's disease during follow-up. Plasma biomarker data, including phosphorylated tau 217 (pTau217), Aβ42/Aβ40 ratio, neurofilament light (NfL), and glial fibrillary acidic protein (GFAP), were available for a subset of 201 participants, which formed the final multimodal modeling cohort. While the reduction in sample size for the biomarker subset (n=201) is a limitation to statistical power, it remains a representative sample of the broader MCI cohort for these exploratory analyses. Within this subset, 62 individuals (30.8%) converted to AD during follow-up. Progression status was defined based on diagnostic classification at subsequent clinical visits and encoded as a binary outcome (1 = converter, 0 = stable MCI). All analyses were conducted using Python (version 3.11) with the pandas and NumPy libraries for data manipulation and numerical computation, scikit-learn for machine learning modeling and validation, and matplotlib for visualization in Jupyter Notebook. Data preprocessing was performed prior to model training. Participants with missing outcome labels were excluded, and observations with incomplete plasma biomarker data were removed from the multimodal subset. Continuous predictor variables, including ADAS13, CDRSB, MMSE, pTau217, Aβ42/Aβ40 ratio, NfL, and GFAP, were standardized using z-score normalization to ensure

comparability across scales. APOE genotype was encoded as ε4 allele count (0, 1, or 2), representing genetic risk dosage. A logistic regression model was constructed to predict progression from MCI to AD using baseline cognitive measures (ADAS13, CDRSB, MMSE), APOE ε4 allele count, and plasma biomarkers (pTau217, Aβ42/Aβ40 ratio, NfL, GFAP). Logistic regression was selected due to its interpretability, suitability for binary classification, and robustness in moderate sample sizes. Model coefficients were exponentiated to obtain odds ratios for clinical interpretability.

Model performance was evaluated using 5-fold cross-validation to estimate generalization ability. In each iteration, four folds were used for training and one fold for validation, with performance metrics averaged across folds. Primary evaluation metrics included classification accuracy and area under the receiver operating characteristic curve (AUC-ROC), with AUC serving as the principal discrimination measure due to its robustness to class imbalance. Statistical significance was defined as $p < .05$.

Results

A total of 767 individuals with baseline mild cognitive impairment (MCI) were identified, of whom 285 participants (37.1%) progressed to Alzheimer's disease during longitudinal follow-up. Plasma biomarker measurements were available for 201 individuals, constituting the final multimodal modeling cohort. Within this subset, 62 participants (30.8%) converted to Alzheimer's disease over the observation period. The multimodal logistic regression model integrating cognitive assessments, APOE ε4 genotype, and plasma biomarkers achieved an overall classification accuracy of 0.69. The model demonstrated good discriminatory performance, with an area under the receiver operating characteristic curve (AUC) of 0.789 (95% CI: 0.72-0.85), indicating effective differentiation between individuals who converted to Alzheimer's disease and those who remained stable. Statistical significance across the primary predictors was maintained ($p < .05$), and the strongest individual separations were observed in CDRSB and pTau217.

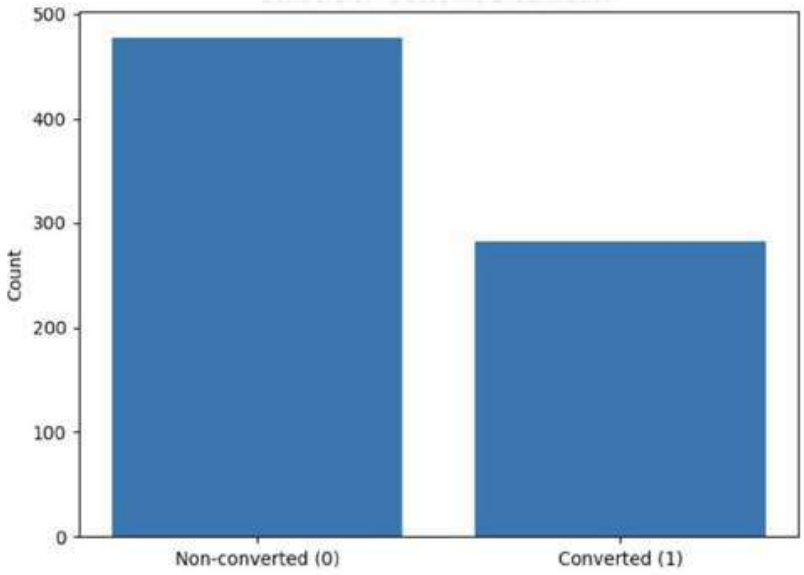


Figure 1: Conversion Outcome Distribution in the Baseline MCI Cohort. This figure shows the number of participants who remained stable versus those who progressed to Alzheimer's disease during follow-up.

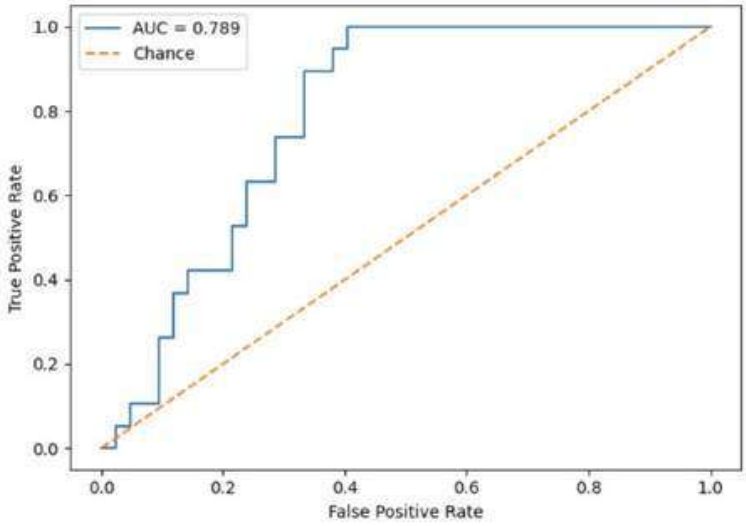


Figure 2: Receiver Operating Characteristic (ROC) Curve for Multimodal Logistic Regression Model. The ROC curve demonstrates model discrimination performance, with an AUC of 0.789 compared to the chance line (AUC = 0.50).

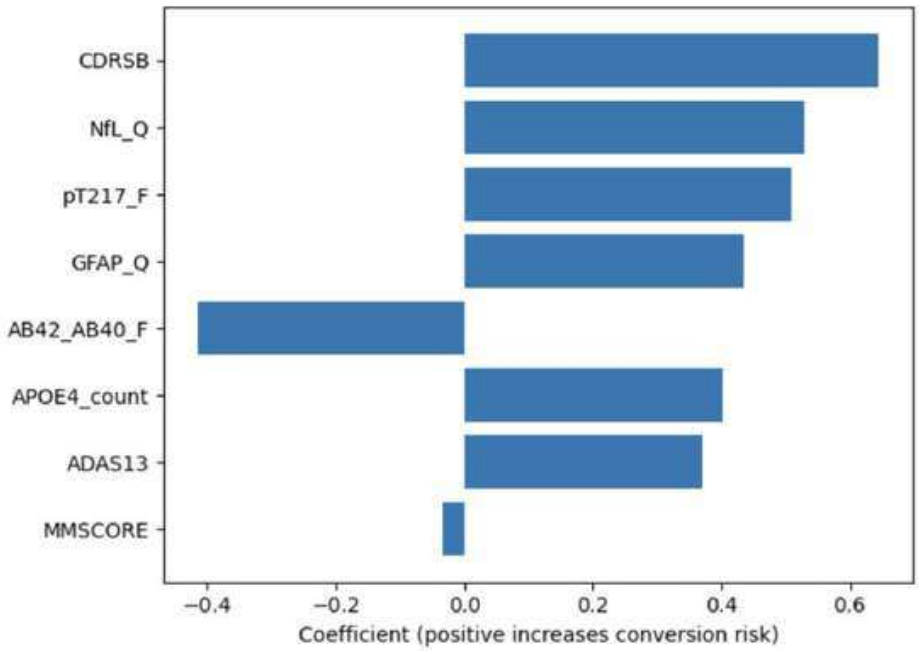


Figure 3: Scaled logistic regression coefficients demonstrated the following direction and relative magnitude of effects. Positive coefficients indicate increased risk of conversion; negative coefficients indicate protective effects.

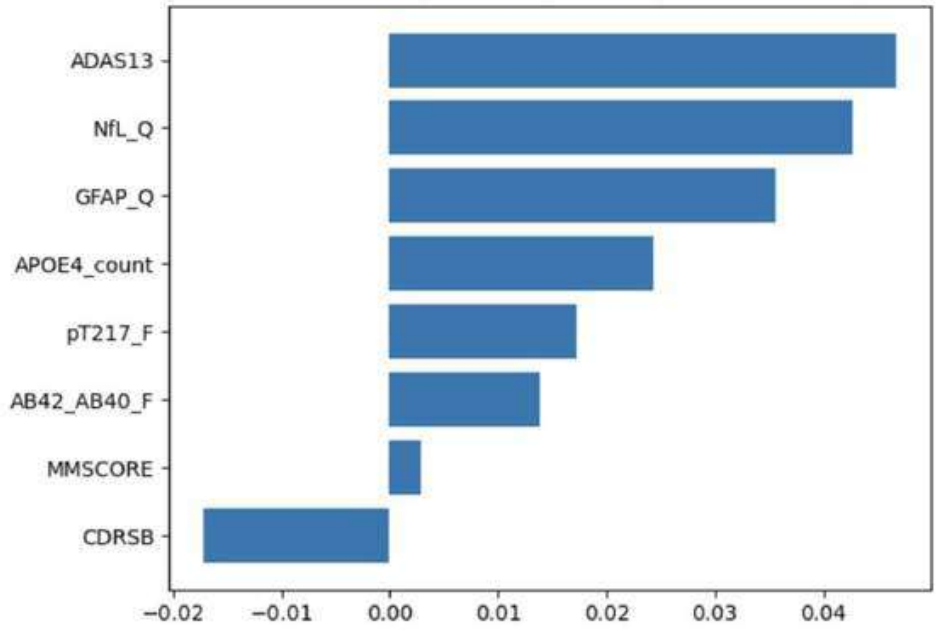


Figure 4: Higher values indicate greater reduction in model performance when the feature is permuted.

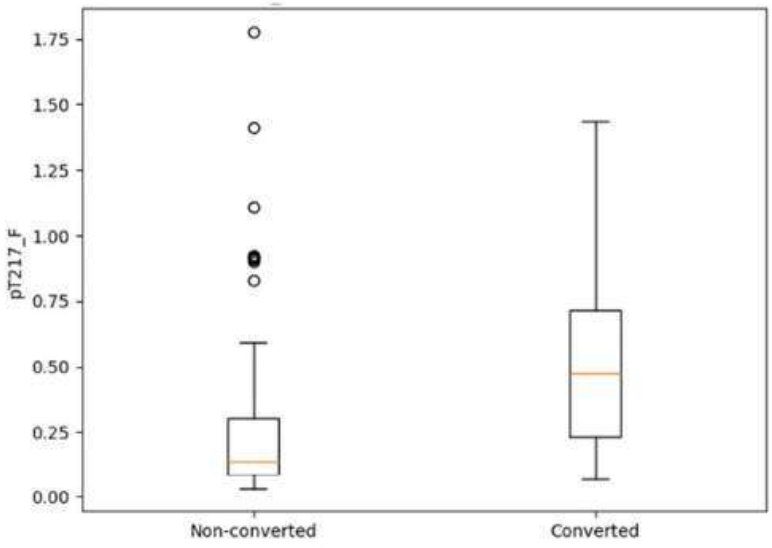


Figure 5: Baseline pTau217 Levels by Conversion Status. Boxplots demonstrate higher median pTau217 levels among converters relative to non-converters.

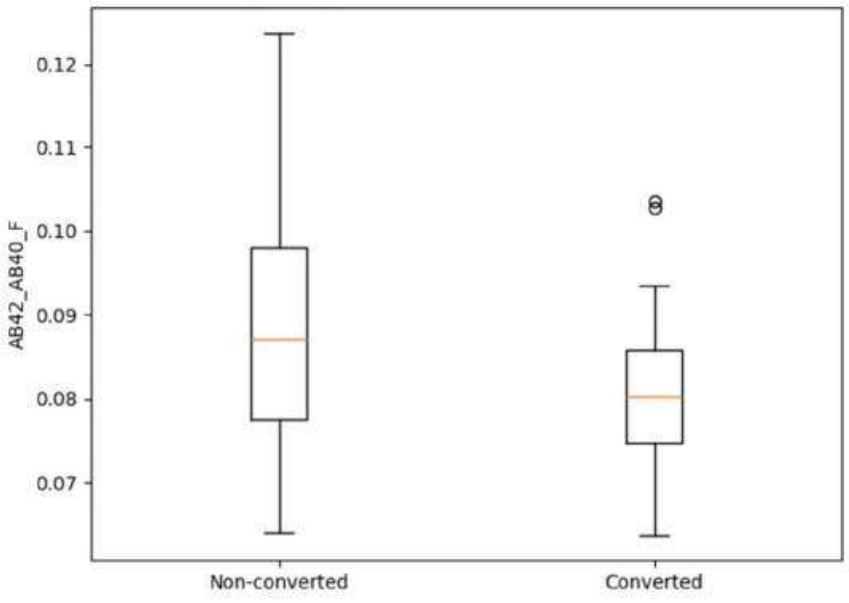


Figure 6: Baseline NfL Levels by Conversion Status. Converters exhibited higher NfL distributions compared to non-converters.

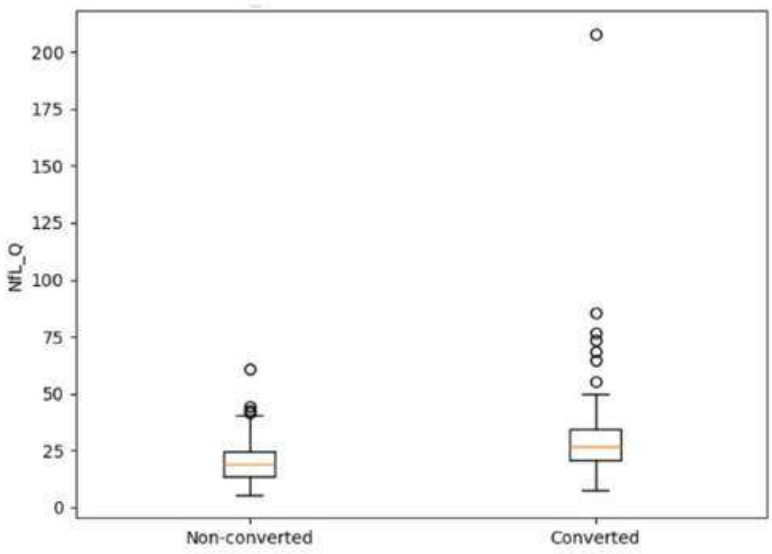


Figure 7: Baseline Aβ42/Aβ40 Ratio by Conversion Status. Converters showed lower amyloid ratios compared to non-converters.

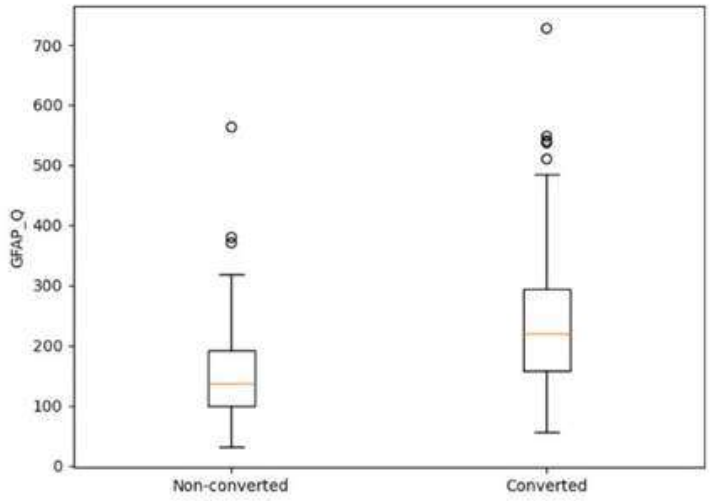


Figure 8: Baseline GFAP Levels by Conversion Status. Boxplots comparing baseline glial fibrillary acidic protein (GFAP) levels between individuals who remained stable and those who converted to Alzheimer's disease. Converters exhibit higher median GFAP values and a wider distribution range.

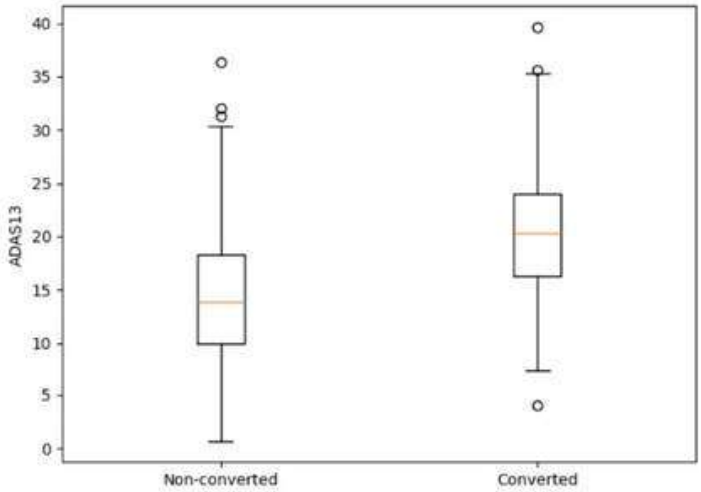


Figure 10: Baseline ADAS13 Scores by Conversion Status. Boxplots showing baseline Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS13) scores stratified by conversion outcome. Converters demonstrate higher median ADAS13 scores compared to non-converters.

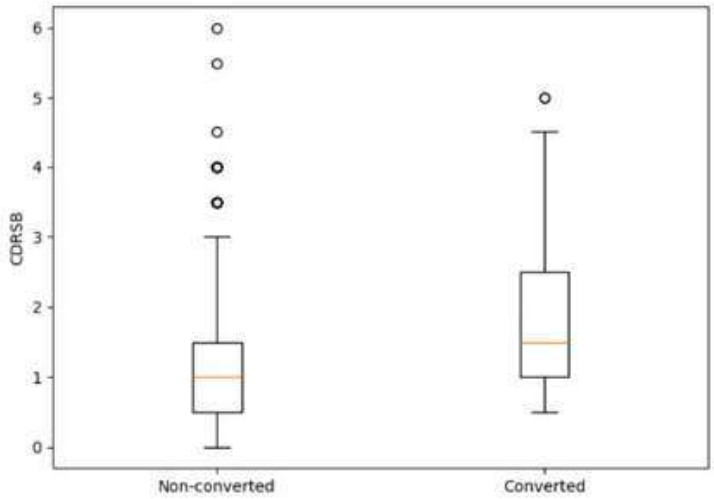


Figure 11: Baseline CDRSB Scores by Conversion Status. Boxplots comparing baseline Clinical Dementia Rating – Sum of Boxes (CDRSB) scores between groups. Higher median CDRSB values are observed among converters.

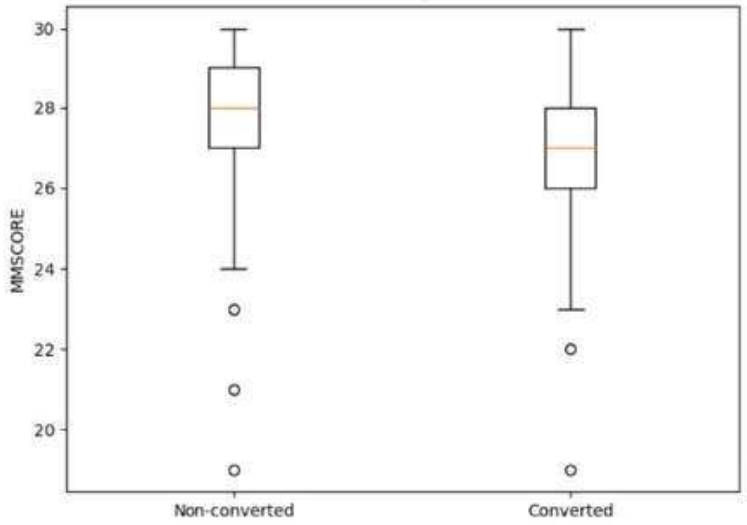


Figure 12: Baseline MMSE Scores by Conversion Status. Boxplots illustrating baseline Mini-Mental State Examination (MMSE) scores by conversion status. Converters show slightly lower median MMSE values relative to non-converters, with overlapping distributions.

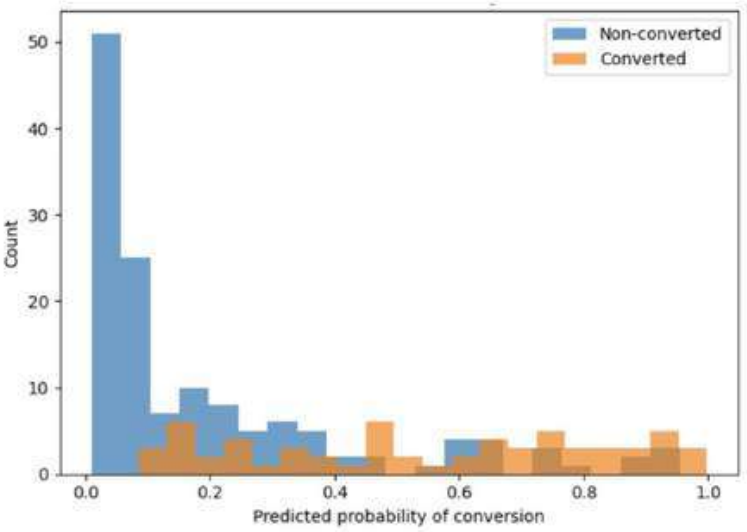


Figure 13: Predicted Conversion Probability Distribution. Histogram of predicted probabilities for converters and non-converters generated by the multimodal logistic regression model.

Discussion

The present study demonstrates that integrating cognitive assessments, APOE ε4 genotype, and plasma biomarkers provides meaningful predictive value for progression from mild cognitive impairment (MCI) to Alzheimer's disease (AD). Within the multimodal subset (n = 201), the logistic regression model achieved an accuracy of 0.69 and an AUC of 0.789 (Figure 2), indicating good discrimination between converters and non-converters. The predicted probability distribution (Figure 13) further supports this finding, as converters were more frequently assigned higher predicted probabilities, although overlap between groups remained. This study acknowledges the high interpretability of the single modelling approach with logistic regression, though future interactions may benefit from comparing these results against non-linear machine learning methods to further validate model stability. The presence of distribution overlap highlights the biological and clinical heterogeneity of MCI and indicates that while multimodal integration improves discrimination, progression risk remains probabilistic rather than deterministic.

Baseline group comparisons reveal consistent biological and clinical differences between converters and non-converters. Converters demonstrated higher ADAS13 scores (Figure 10) and higher CDRSB scores (Figure 11), reflecting greater baseline cognitive and functional impairment. In contrast, MMSE scores showed only modest separation with substantial overlap (Figure 12), suggesting that MMSE may be less sensitive to early progression risk when compared to more granular cognitive and functional measures. Plasma biomarkers exhibited systematic directional differences aligned with AD pathophysiology. Individuals who converted had higher baseline pTau217 levels (Figure 5), elevated NfL (Figure 6), and higher GFAP concentrations (Figure 8), while the Aβ42/Aβ40 ratio was lower among converters (Figure 7). These findings support the biological plausibility of the model and align with established frameworks describing amyloid dysregulation, tau pathology, neurodegeneration, and astroglial activation as central components of AD progression.

The regression coefficient plot (Figure 3) reinforces these observations, demonstrating positive associations between conversion risk and CDRSB, NfL, pTau217, GFAP, ADAS13, and APOE ε4

count, while the Aβ42/Aβ40 ratio shows a protective negative association. The minimal contribution of MMSE in the multivariable setting mirrors its limited baseline separation (Figure 12), suggesting redundancy once more sensitive cognitive measures are included. Notably, permutation-based feature importance (Figure 4) shows that certain predictors produce larger reductions in AUC when shuffled, highlighting their relative contribution to model discrimination. Differences between coefficient magnitude (Figure 3) and permutation importance (Figure 4) likely reflect intercorrelations among predictors; when multiple features encode overlapping information, permuting one may not dramatically reduce performance because related variables retain partial signal. This emphasizes the value of examining multiple interpretability methods rather than relying on a single metric of importance.

Biologically, the findings suggest that conversion from MCI to AD is closely associated with markers of tau pathology and neurodegeneration. Elevated pTau217 (Figure 5) indicates increased tau-related processes among converters, while elevated NfL (Figure 6) reflects greater axonal injury. Increased GFAP levels (Figure 8) further suggest astroglial activation and neuroinflammatory involvement in individuals at higher risk of progression. Although the lower Aβ42/Aβ40 ratio among converters (Figure 7) supports the role of amyloid pathology, the comparatively strong contributions of tau and neurodegeneration markers are consistent with emerging models in which amyloid accumulation may precede but not directly determine short-term clinical decline. In this context, markers of active neuronal injury and tau dysfunction may provide stronger near-term prognostic information during the MCI stage.

From a computational standpoint, the results demonstrate that a relatively simple and interpretable model can achieve clinically meaningful discrimination. The AUC of 0.789 (Figure 2) indicates performance substantially above chance, and the predicted probability gradients (Figure 13) suggest that the model can stratify individuals along a risk continuum rather than only generating binary classifications. Logistic regression offers the additional advantage of transparent effect interpretation through coefficients and odds ratios (Figure 3), which enhances potential clinical

translation compared to more complex black-box models. However, the presence of distribution overlap across nearly all features (Figures 5–12) underscores the inherent complexity of AD progression and the limits of baseline-only modeling.

Several limitations should be acknowledged. First, the final multimodal cohort was restricted to participants with complete plasma biomarker data (n = 201), reducing statistical power and potentially affecting the stability of feature importance estimates. Second, conversion was defined using clinical diagnostic status, which may be influenced by variability in follow-up duration or diagnostic criteria. Third, the ADNI cohort may not fully represent broader clinical populations, limiting generalizability. Finally, although the model demonstrates good discrimination, the overlapping probability distributions (Figure 13) indicate that prediction remains imperfect and should not be interpreted as definitive for individual patients.

Future work should expand the sample size and incorporate external validation to assess generalizability. Longitudinal modeling of biomarker trajectories, rather than baseline values alone, may improve predictive accuracy by capturing dynamic disease progression. Additional multimodal features, including neuroimaging markers or digital cognitive measures, could further enhance predictive performance. Model calibration and clinical threshold optimization will also be important for translating predicted probabilities into actionable risk categories. Collectively, these findings support the value of multimodal integration while emphasizing the need for continued refinement to

Conclusion/Future Directions

This study demonstrates that integrating cognitive assessments, APOE ε4 genotype, and plasma biomarkers improves prediction of progression from mild cognitive impairment to Alzheimer's disease. The multimodal logistic regression model achieved good discriminatory performance (AUC = 0.789), indicating that combining clinical and biological features provides meaningful prognostic value beyond cognitive measures alone. Across analyses, individuals who converted to Alzheimer's disease exhibited greater functional impairment, elevated tau and neurodegeneration markers, and lower amyloid ratios at baseline, supporting the biological validity

of the model.

The primary contribution of this work lies in demonstrating that blood-based biomarkers, particularly pTau217, NfL, and GFAP, can be effectively integrated with traditional cognitive assessments within an interpretable predictive framework. By combining clinical, genetic, and circulating biomarker data, this study reinforces emerging models of Alzheimer's disease that emphasize the role of active neurodegeneration and tau pathology during early disease transition.

Broader implications include the potential for scalable, minimally invasive risk stratification using plasma biomarkers rather than relying solely on neuroimaging or cerebrospinal fluid measures. With further validation in larger and more diverse cohorts, this multimodal approach could support earlier identification of high-risk individuals, improve clinical trial enrichment strategies, and contribute to personalized prognostic modeling in Alzheimer's disease.

Ethical Integrity Statement

We confirm that this submission is our original work completed without the use of AI tools , that all sources are properly cited in APA 7 format, and that we understand BACSA is not responsible for any plagiarism or academic misconduct.

References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the National Institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 270–279. https://doi.org/10.1016/j.jalz.2011.03.008

Alzheimer's disease facts and figures. Alzheimer's Association. (2025). https://www.alz.org/alzheimers-dementia/facts-figures

Alzheimer's Disease Neuroimaging Initiative. (2023). ADNI data repository. https://adni.loni.usc.edu

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L., & Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. *Science*, *261*(5123), 921–923. https://doi.org/10.1126/science.8346443

De Strooper, B., & Karran, E. (2016). The cellular phase of alzheimer's disease. *Cell*, *164*(4), 603–615. https://doi.org/10.1016/j.cell.2015.12.056

Hardy, J. (2006). Alzheimer's disease: The amyloid cascade hypothesis: An update and reappraisal. *Journal of Alzheimer's Disease*, *9*(s3), 151–153. https://doi.org/10.3233/jad-2006-9s317

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jack, C., Bennett, D., Blennow, K., Carrillo, M., Dunn, B., Haeberlein, S., Holtzman, D., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J., Montine, T., Phelps, C., Rankin, K., Rowe, C., Scheltens, P., Siemers, E., Snyder, H., … Silverberg, N. (2018). Nia-AA Research Framework: Toward a biological definition of alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535–562. https://doi.org/10.1016/j.jalz.2018.02.018

Mattsson, N., Andreasson, U., Zetterberg, H., & Blennow, K. (2017). Association of plasma neurofilament light with neurodegeneration in patients with alzheimer disease. *JAMA Neurology*, *74*(5), 557. https://doi.org/10.1001/jamaneurol.2016.6117

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based alzheimer's conversion prediction in MCI subjects. *NeuroImage*, *104*, 398–412. https://doi.org/10.1016/j.neuroimage.2014.10.002

NumPy Developers. (2023). *NumPy (Version 1.x)* [Computer software]. https://numpy.org/

Palmqvist, S., Janelidze, S., Quiroz, Y. T., Zetterberg, H., Lopera, F., Stomrud, E., Su, Y., Chen, Y., Serrano, G. E., Leuzy, A., Mattsson-Carlgren, N., Strandberg, O., Smith, R., Villegas, A., Sepulveda-Falla, D., Chai, X., Proctor, N. K., Beach, T. G., Blennow, K., … Hansson, O. (2020). Discriminative accuracy of Plasma Phospho-Tau217 for alzheimer disease vs other neurodegenerative disorders. *JAMA*, *324*(8), 772. https://doi.org/10.1001/jama.2020.12134

Petersen, R. C. (2018). Practice guideline update summary: Mild cognitive impairment [retired]. *Neurology*, *90*(3), 126–135. https://doi.org/10.1212/wnl.0000000000004826

Project Jupyter. (2023). *Jupyter Notebook (Version 6.x)* [Computer software]. https://jupyter.org/

Python Software Foundation. (2023). *Python (Version 3.11)* [Computer software]. https://www.python.org/

Scikit-learn Developers. (2023). *Scikit-learn (Version 1.x)* [Computer software]. https://scikit-learn.org/

Shir, D., Graff-Radford, J., Hofrenning, E. I., Lesnick, T. G., Przybelski, S. A., Lowe, V. J., Knopman, D. S., Petersen, R. C., Jack, C. R., Vemuri, P., Algeciras-Schimnich, A., Campbell, M. R., Stricker, N. H., & Mielke, M. M. (2022, February 28). *Association of plasma glial fibrillary acidic protein (GFAP) with neuroimaging of alzheimer's disease and vascular pathology*. Alzheimer's & dementia (Amsterdam, Netherlands). https://pmc.ncbi.nlm.nih.gov/articles/PMC8883441/

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jack, C. R., Kaye, J., Montine, T. J., Park, D. C., Reiman, E. M., Rowe, C. C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M. C., Thies, B., Morrison-Bogorad, M., … Phelps, C. H. (2011). Toward defining the preclinical stages of alzheimer's disease: Recommendations from the National Institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 280–292. https://doi.org/10.1016/j.jalz.2011.03.003

The pandas development team. (2023). *pandas (Version 2.x)* [Computer software]. https://pandas.pydata.org/

Structural Modeling of the CFTR G500D Variant: An AlphaFold-Based Analysis of Protein Folding Defects

Omer Yimaz, Ekambir Singh, Anisa Zangiabadi

Abstract

Cystic Fibrosis (CF), an inherited genetic disorder, is caused by recessive deleterious mutations associated with CFTR gene. The ΔF508 deletion is the most commonly known CFTR mutation, but several rare mutants exist which remain untreatable due to a lack of data personalization. In this study, we will be focusing on one of these rare mutations: Glycine to Aspartic Acid at position 500 (G500D). The study utilized AlphaFold and ChimeraX to perform a comparative analysis between wild-type (WT) CFTR and the G500D variant. High-confidence models were generated (average pLDDT > 70), showing a low pruned Root Mean Square Deviation (RMSD) of 0.838 Å, which indicates that the G500D mutation does not prevent the synthesis of a full-length protein. However, a significant global RMSD of 15.544 Å was observed, centered on localized misalignments within the nucleotide-binding domains (NBDs). These structural deviations suggest that while the protein is expressed, the G500D variant suffers from a distinct local folding defect that disrupts its local functional conformation. Atomic-level structural analysis using ChimeraX (1.11.1) also indicated that Aspartic Acid at position 500 (ASP500) potentially forms a new hydrogen bond (2.741 Å) not present in the WT structure, which may perhaps locally hinder flexibility and help explain the inhibited functional gating mechanism. These findings characterize G500D as a potentially modifiable target, providing a structural basis for future screening of small-molecule correctors.

1. Introduction

1.1 Genetic Foundations and and Protein Synthesis

The transition from genetic information (DNA to RNA) to a functional protein is a process called the central dogma of molecular biology. Genetic mutations are caused due to modifications in the genome, often caused by replication errors or exposure to environmental mutagens [12]. In the context of the CFTR gene, single-nucleotide polymorphisms (SNPs) can happen where a single base substitution changes the codon to encode a different amino acid.

These primary structural changes are critical because the chemical properties of an amino

acid.

These primary structural changes are critical because the chemical properties of an amino acid, such as size, charge, and hydrophobicity, dictate the folding pathway of the polypeptide chain [13]. In the case of CFTR, a substitution of the Glycine to Aspartic Acid at position 500 (G500D) introduces a charged residue into the sequence potentially optimized for a small, neutral glycine, thereby disrupting the non-covalent interactions (hydrogen bonding, van der Waals forces) necessary for the protein's tertiary stability [14].

1.2 The Molecular Physiology of Cystic Fibrosis

Cystic Fibrosis (CF) is an autosomal recessive genetic disorder characterized by a malfunctioning CFTR protein, a cAMP-regulated chloride channel located in the apical membrane of epithelial cells [1]. This protein is responsible for maintaining the salt-water balance on mucosal surfaces by facilitating the transport of chloride and bicarbonate ions. When this channel is defective, ion transport is hindered, leading to the accumulation of thick, dehydrated, and hyperviscous mucus [5].

1.3 Pathology, Immunochemistry, and Disease Burden

The pathology of CF is multi-systemic, meaning it affects multiple systems of the body. In the lungs, the "stagnant mucus" phenotype provides a fertile environment for chronic bacterial colonization, most notably by *Pseudomonas aeruginosa* and *Staphylococcus aureus* [15]. This leads to a persistent inflammatory state characterized by an influx of neutrophils and the release of pro-inflammatory cytokines (e.g., IL-8, TNF-α), which progressively destroy lung tissue [16].

Despite advancements in care, CF significantly reduces life expectancy and requires a rigorous daily treatment regimen of chest physiotherapy and pharmacological interventions. In the pancreas, mucus blockage prevents the secretion of digestive enzymes, leading to malabsorption and CF-related diabetes [17]. The high cost of treatment and the psychological impact of a chronic, life-limiting condition underscore the urgent need for breakthrough therapies that address the underlying protein defect rather than just the symptoms.

1.4 Classification of CFTR Mutations

The therapeutic approach to CF is largely determined by the specific class of mutation contained in the protein. Mutation classes are defined below:

- **Class I Mutations:** These mutations result in premature termination codons (PTC). This typically leads to no functional CFTR protein being produced. Examples include frameshift mutations such as c.3623delG and c.3755delG [6].
- **Class II Mutations:** These mutations lead to protein misfolding and subsequent deterioration by the cell's quality control mechanisms. The cellular quality control

mechanism, primarily the endoplasmic reticulum-associated degradation (ERAD) pathway, identifies the protein as defective and degrades it before it reaches the cell membrane [2]. The most common mutation, ΔF508 deletion, falls into this category.

- **Rare and Missense Mutations:** These are mutations such as single-nucleotide substitutions that change one amino acid in a protein often leading to detrimental effects on protein structure, stability, and function. Glycine to Aspartic Acid at position 500 (G500D) mutation belongs to this category and may result in a full-length protein that is synthesized but structurally malformed and possibly dangerous [2].

1.5 The Need for Computational Data

Although highly effective CFTR modulators, including both correctors and potentiators, are available for prevalent mutations, rare variants such as G500D lack approved pharmacological interventions. The absence of high-resolution experimental structures for these specific variants limits the efficacy of traditional structure-based drug discovery. To advance personalized therapeutic strategies, computational modeling and three-dimensional structural analysis are required to characterize the conformational changes induced by specific mutations.

2. Hypothesis and Objectives

Research Question: To what extent can in silico modeling via AlphaFold characterize domain-level structural deviations, specifically localized RMSD shifts and alterations in inter-residue contacts, caused by the G500D mutation in CFTR?

Hypothesis: It is hypothesized that AlphaFold-generated models of the G500D variant will exhibit significant localized structural deviations (RMSD > 2.0 Å) compared to the wild-type, specifically within the Nucleotide Binding Domains (NBDs), suggesting a conformational instability that is potentially targets for small-molecule correctors.

Specific Objectives:

1. **Generate** high-confidence 3D structural models of wild-type and G500D CFTR using AlphaFold.
2. **Quantify** global and domain-specific structural variance between the wild-type and mutant proteins using Root Mean Square Deviation (RMSD)
3. Identify sites of misfolding within the G500D variant.
4. Evaluate the structure of shortened proteins from frameshift mutations

3. Methods

3.1 Data Sources and Sequence Preparation

The primary amino acid sequence for the human CFTR protein was retrieved from the UniProtKB database (Accession ID: P13569). This 1480-residue sequence served as the wild-type (WT) reference. To generate the mutant dataset, the WT sequence was manually edited to incorporate a missense substitution at position 500, replacing the native Glycine (G) with Aspartic Acid (D), creating the G500D variant [8].

3.2 Computational Modeling with AlphaFold

Structural predictions were performed using AlphaFold v2.3.2. For the G500D variant, the edited FASTA sequence was processed using the standard AlphaFold pipeline with default parameters as shown in Figure 1 [3,10, 11].


```
{
    "num_queries": 1,
    "use_templates": true,
    "num_relax": 0,
    "relax_max_iterations": 200,
    "relax_tolerance": 2.39,
    "relax_stiffness": 10.0,
    "relax_max_outer_iterations": 3,
    "msa_mode": "mmseqs2_uniref_env",
    "model_type": "alphafold2_ptm",
    "num_models": 5,
    "num_recycles": 3,
    "recycle_early_stop_tolerance": null,
    "num_ensemble": 1,
    "model_order": [1, 2, 3, 4, 5],
    "initial_guess": null,
    "keep_existing_results": false,
    "rank_by": "plddt",
    "max_seq": 512,
    "max_extra_seq": 5120,
    "pair_mode": "unpaired_paired",
    "pairing_strategy": "greedy",
    "host_url": "https://api.colabfold.com",
    "user_agent": "colabfold/google-colab-main",
    "stop_at_score": 100.0,
    "random_seed": 0,
    "num_seeds": 1,
    "recompile_padding": 10,
    "commit": "905fdfddf3acb97a86c97e34ccee06882139cc25",
    "use_dropout": false,
    "use_cluster_profile": true,
    "use_fuse": true,
    "use_bfloat16": true,
    "version": "1.5.5",
    "calc_extra_ptm": false,
    "use_probs_extra": true,
    "max_template_date": "2100-01-01",
    "max_template_hits": 20
}
```


Figure 1: AlphaFold parameters used.

3. Methods

3.1 Data Sources and Sequence Preparation

The primary amino acid sequence for the human CFTR protein was retrieved from the UniProtKB database (Accession ID: P13569). This 1480-residue sequence served as the wild-type (WT) reference. To generate the mutant dataset, the WT sequence was manually edited to incorporate a missense substitution at position 500, replacing the native Glycine (G) with Aspartic Acid (D), creating the G500D variant [8].

3.2 Computational Modeling with AlphaFold

Structural predictions were performed using AlphaFold v2.3.2. For the G500D variant, the edited FASTA sequence was processed using the standard AlphaFold pipeline with default parameters as shown in Figure 1 [3,10, 11].


```
{
    "num_queries": 1,
    "use_templates": true,
    "num_relax": 0,
    "relax_max_iterations": 200,
    "relax_tolerance": 2.39,
    "relax_stiffness": 10.0,
    "relax_max_outer_iterations": 3,
    "msa_mode": "mmseqs2_uniref_env",
    "model_type": "alphafold2_ptm",
    "num_models": 5,
    "num_recycles": 3,
    "recycle_early_stop_tolerance": null,
    "num_ensemble": 1,
    "model_order": [1, 2, 3, 4, 5],
    "initial_guess": null,
    "keep_existing_results": false,
    "rank_by": "plddt",
    "max_seq": 512,
    "max_extra_seq": 5120,
    "pair_mode": "unpaired_paired",
    "pairing_strategy": "greedy",
    "host_url": "https://api.colabfold.com",
    "user_agent": "colabfold/google-colab-main",
    "stop_at_score": 100.0,
    "random_seed": 0,
    "num_seeds": 1,
    "recompile_padding": 10,
    "commit": "905fdfddf3acb97a86c97e34ccee06882139cc25",
    "use_dropout": false,
    "use_cluster_profile": true,
    "use_fuse": true,
    "use_bfloat16": true,
    "version": "1.5.5",
    "calc_extra_ptm": false,
    "use_probs_extra": true,
    "max_template_date": "2100-01-01",
    "max_template_hits": 20
}
```


Figure 1: AlphaFold parameters used.

3.3 Scientific Integrity and pLDDT Validation

To ensure the accuracy of our models, we calculated the pLDDT (predicted Local Distance Difference Test) scores. pLDDT is a per-residue measure of local confidence in the predicted structure on a scale of 0 to 100. A score above 90 indicates very high confidence; 70-90 is considered high; scores below 50 suggest the region may be poorly predicted. Our wild-type CFTR model (AF-P13569-F1-v6) provided an average pLDDT of 75.62, indicating a high overall level of confidence in the model's structural framework.

3.4 Structural Alignment and Quantitative Analysis

Quantitative comparison between the wild-type and mutant models was conducted in UCSF ChimeraX v1.6 using the following protocol:We calculated the RMSD for the backbone Cα atoms.

1) To get the sequence alignment a Needleman-Wunsch alignment was performed to map the G500D sequence to the WT.
2) The MatchMaker tool was utilized to superimpose the Cα backbones.
3) A pruned RMSD calculation was performed to identify the 'best-fit' core alignment by iteratively removing outlier residue pairs with deviations exceeding 2.0 Å. This subset of residue pairs represents the structural regions that remain most stable after alignment. In contrast, a global RMSD calculation (covering all 1480 residues) was conducted to measure the total degree of conformational divergence and potential misfolding across the entire protein structure.

4. Results

4.1 Sequence Alignment and Molecular Stability

The alignment of the G500D variant sequence was compared with the WT CFTR sequence (UniProt P13569) with the parameters indicated in Table 1 [7]. The comparison yielded a sequence alignment score of 7398.1, reflecting a high degree of primary structural homology despite the single amino acid substitution. Consequently, the AlphaFold2 models for both the WT and G500D variant achieved high-confidence scores (pLDDT > 70), indicating that both sequences are capable of forming a full-length tertiary structure.

Table 1: Detailed comparison of the G500D variant against the WT CFTR (P13569), showing the primary structural homology required for high-confidence in silico folding.

Parameters	
Chain pairing	bb
Alignment algorithm	Needleman-Wunsch
Similarity matrix	BLOSUM-62
SS fraction	0.3
Gap open (HH/SS/other)	18/18/6
Gap extend	1
SS matrix	H S O H 6 -9 -6 S 6 -6 O 4
Iteration cutoff	2

4.2 Localized vs. Global Structural Divergence (RMSD)

To measure the structural impact of the G500D mutation, a global superposition of the AlphaFold-generated models was performed. The Root Mean Square Deviation (RMSD) results are summarized in Table 2.

Table 2: Summary of Root Mean Square Deviation values, distinguishing between the stable protein core (Pruned) and the significant conformational misfolding (Global) induced by the G500D mutation

	Number of Atom Pairs Analyzed in RMSD Calculation	RMSD Result (to 3 decimal places)
Pruned Atom Pairs	673 pruned atom pairs	0.838 Å
Global Alignment	1480 atom pairs (full length of the protein)	15.544 Å

As indicated in Table 2, the analysis revealed a low pruned RMSD of 0.838 Å. This indicates that the fundamental scaffold of the CFTR protein remains largely congruent between the WT and the mutant.

In contrast, the global RMSD reached a significant value of 15.544 Å. This high value indicates that while the "core" is stable, there are massive conformational shifts in the flexible loops or peripheral domains. For a protein of this size, an RMSD over 10 Å suggests a "misfolded" state where the domains are not oriented correctly to facilitate ion transport.

4.3 Atomic-Level Analysis at the Mutation Site (Position 500)

Visual analysis in ChimeraX allows for a direct comparison of the residue environment at the

mutation site. Figure 2 shows a global superposition of the WT and the G500D mutant. The overall folding is largely preserved, with no major global abnormalities seen.

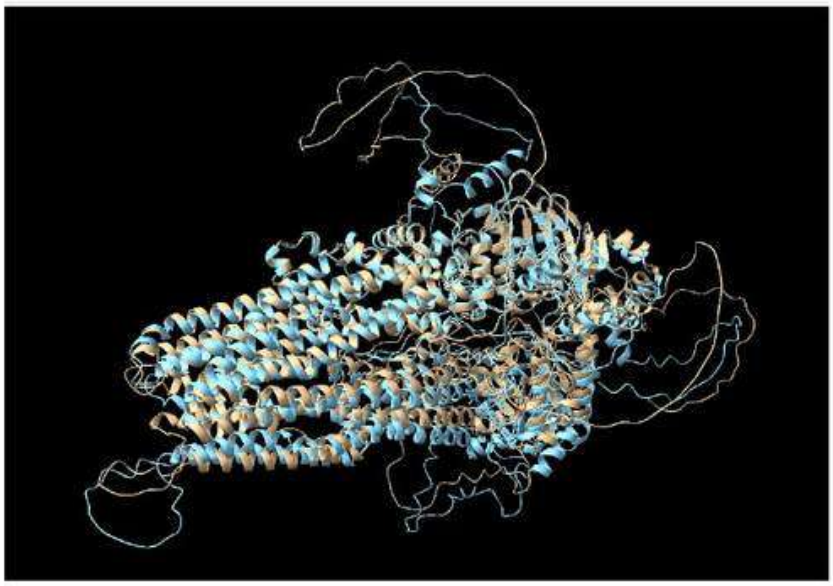


Figure 2: Global Structural Superposition of WT and G500D CFTR Models. A full-length ribbon representation showing the alignment of the AlphaFold-generated WT (tan) and G500D variant (blue) structures. The transmembrane domains show high stability, although significant conformational divergence is observed in the peripheral loops, resulting in a global RMSD of 15.544 Å.

However, while the global backbone shifts are subtle, the substitution of the tiny Glycine with the larger, negatively charged Aspartic Acid introduces local steric hindrance and altered electrostatic interactions as shown by the formation of a new, strong hydrogen bond of length 2.741Å between ASP500 and THR501in Figure 3. This local change causes a series of misalignments in the neighboring residues.

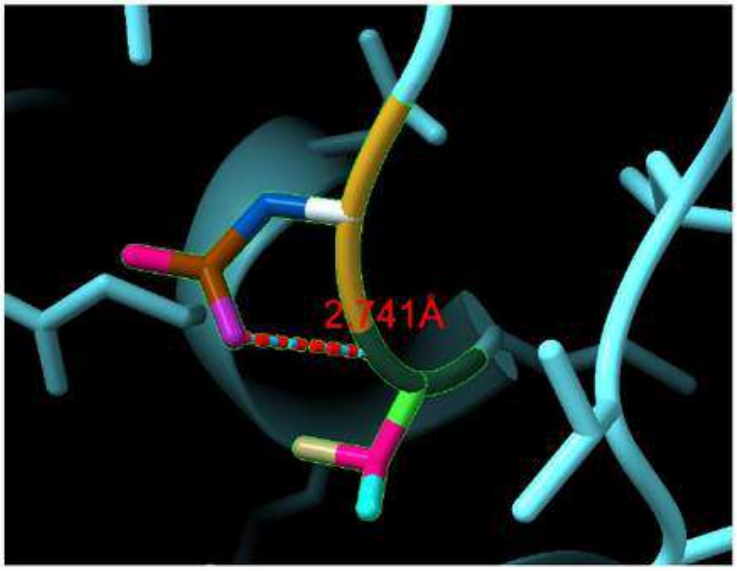


Figure 3: Zoom-in on the region surrounding the G500D mutation in CFTR - residues 500 and 501. Substitution of Gly with Asp at position 500 introduces a new hydrogen bond (2.741 Å) between ASP500 (Mustard) and THR501 (Dark Green), displayed as the red dotted line. This bond is absent in the WT structure and highlights a local change in residual side-chain interactions caused by the mutation.

5. Discussion

5.1 Analysis of G500D Misfolding and Therapeutic Potential

The G500D mutation involves the substitution of a Glycine residue with an Aspartic Acid residue at position 500. Glycine is a small, neutral amino acid, whereas Aspartic Acid is larger and carries a negative charge. Our data indicates that this substitution does not cause a collapse of the entire protein structure. This is supported by the low pruned RMSD of 0.838 Å which shows that a significant portion of the protein backbone remains stable and aligned with the wild-type structure.

However, the mutation leads to a significant domain-level shift, as reflected in the high global RMSD of 15.544 Å. The contrast between these two values provides specific insights into the nature of the G500D defect. The localized stability suggested by the pruned RMSD indicates that the transmembrane domains (TMDs) and major globular regions maintain a wild-type-like fold. In contrast, the high global RMSD suggests that, while individual domains are stable, their relative positions or the folding of flexible loops—specifically within Nucleotide Binding Domain 1 (NBD1) where the mutation is located—are significantly disrupted. This disruption likely affects the "ATP-binding pocket" or the interface between NBD1 and the intracellular loops (ICLs) of the TMDs.

This finding is critical because it suggests that the G500D variant likely reaches the cell membrane but remains in a non-functional or misaligned state. Because the core scaffold is intact, this mutation is a strong candidate for "corrector" molecules. These small-molecule drugs could potentially stabilize the NBD1-TMD interface to restore proper channel gating

rather than having to address a complete loss of protein integrity.

In addition to these global and domain-level observations, analysis of the local environment close to the mutation site revealed the formation of a new hydrogen bond between ASP500 and THR501 (2.741 Å), a key observation which is not observed in the WT structure. This finding exposes a potentially significant change in the interaction network within NBD1. Because GLY500 does not have a side-chain and does not contribute to local electrostatic interactions with other residues and side-chains, its replacement with a polar, negatively charged ASP contributes to steric load and the opportunity for potentially new electrostatic interactions. This new, strong hydrogen bond being present suggests that a local stabilizing interaction forms, which may limit structural flexibility in this region. Considering the role of NBD1 in being flexible and dynamic for ATP-dependent conformational changes, we speculate that this interaction could act as a structural 'lock', limiting the flexibility required for proper domain movements and contributing to the misalignment indicated by the high global RMSD. This allows for a possible structural mechanism by which the mutation mostly preserves overall global folding while hampering functional gating mechanisms at the same time.

5.2 Evaluation of AlphaFold as a Diagnostic Tool

The results demonstrate that AlphaFold-generated models can effectively identify specific sites of structural failure. By comparing pruned and global RMSD, we can distinguish between mutations that cause total protein loss and those that cause localized, "correctable" shifts. In the case of G500D, the identification of a structural shift exceeding the 2.0 Å threshold provides a clear target for stabilization.

This ability to pinpoint localized defects distinguishes G500D from Class I mutations, such as the frameshifts analyzed in this study. These mutations result in a total loss of essential domains and cannot be addressed by folding stabilizers. Therefore, AlphaFold serves as a valuable tool for categorizing rare variants based on whether they are structurally salvageable, which is a key step toward personalized medicine in Cystic Fibrosis treatment.

5.3 Technical Demands and Limitations

The average pLDDT score of 75.62 confirms the technical reliability of our models for the structured regions of the CFTR protein. However, it is important to note the limitations of these computational predictions. AlphaFold provides a static "snapshot" of a protein's significantly disrupted. This disruption likely affects the "ATP-binding pocket" or the interface between NBD1 and the intracellular loops (ICLs) of the TMDs.

This finding is critical because it suggests that the G500D variant likely reaches the cell membrane but remains in a non-functional or misaligned state. Because the core scaffold is intact, this mutation is a strong candidate for "corrector" molecules. These small-molecule drugs could potentially stabilize the NBD1-TMD interface to restore proper channel gating rather than having to address a complete loss of protein integrity.

significantly disrupted. This disruption likely affects the "ATP-binding pocket" or the interface between NBD1 and the intracellular loops (ICLs) of the TMDs.

This finding is critical because it suggests that the G500D variant likely reaches the cell membrane but remains in a non-functional or misaligned state. Because the core scaffold is intact, this mutation is a strong candidate for "corrector" molecules. These small-molecule drugs could potentially stabilize the NBD1-TMD interface to restore proper channel gating rather than having to address a complete loss of protein integrity.

In addition to these global and domain-level observations, analysis of the local environment close to the mutation site revealed the formation of a new hydrogen bond between ASP500 and THR501 (2.741 Å), a key observation which is not observed in the WT structure. This finding exposes a potentially significant change in the interaction network within NBD1. Because GLY500 does not have a side-chain and does not contribute to local electrostatic interactions with other residues and side-chains, its replacement with a polar, negatively charged ASP contributes to steric load and the opportunity for potentially new electrostatic interactions. This new, strong hydrogen bond being present suggests that a local stabilizing interaction forms, which may limit structural flexibility in this region. Considering the role of NBD1 in being flexible and dynamic for ATP-dependent conformational changes, we speculate that this interaction could act as a structural 'lock', limiting the flexibility required for proper domain movements and contributing to the misalignment indicated by the high global RMSD. This allows for a possible structural mechanism by which the mutation mostly preserves overall global folding while hampering functional gating mechanisms at the same time.

5.2 Evaluation of AlphaFold as a Diagnostic Tool

The results demonstrate that AlphaFold-generated models can effectively identify specific sites of structural failure. By comparing pruned and global RMSD, we can distinguish between mutations that cause total protein loss and those that cause localized, "correctable" shifts. In the case of G500D, the identification of a structural shift exceeding the 2.0 Å threshold provides a clear target for stabilization.

This ability to pinpoint localized defects distinguishes G500D from Class I mutations, such as the frameshifts analyzed in this study. These mutations result in a total loss of essential domains and cannot be addressed by folding stabilizers. Therefore, AlphaFold serves as a valuable tool for categorizing rare variants based on whether they are structurally salvageable, which is a key step toward personalized medicine in Cystic Fibrosis treatment.

5.3 Technical Demands and Limitations

The average pLDDT score of 75.62 confirms the technical reliability of our models for the structured regions of the CFTR protein. However, it is important to note the limitations of these computational predictions. AlphaFold provides a static "snapshot" of a protein's

structure. In a living cell, CFTR is a highly dynamic protein that undergoes constant shape changes during ATP binding and hydrolysis.

Furthermore, AlphaFold does not fully account for how the lipid bilayer (the cell membrane) influences protein stability. These results should be considered as a structural baseline. Future research should use these models as a starting point for Molecular Dynamics (MD) simulations to observe how the G500D shift behaves in a simulated membrane environment and to confirm if structural variations result in a permanent loss of channel function.

6. Conclusion

This study successfully used AlphaFold to analyze structural changes in rare CFTR variants. For the G500D mutation, we found a stable core structure (0.838 Å pruned RMSD) but a significant domain-level shift (15.544 Å global RMSD). This supports the conclusion that G500D causes a localized folding defect rather than a total loss of protein structure.

Future research should focus on using these 3D models for virtual drug screening to find molecules that can bind to and stabilize the G500D variant. Additionally, performing Molecular Dynamics simulations will help us see how the protein moves through a membrane. This approach provides a potentially viable path toward the development of personalized treatments for patients with rare Cystic Fibrosis mutations.

7. Ethical Integrity Statement

We confirm that this submission is our original work completed without the use of AI tools , that all sources are properly cited in APA 7 format, and that we understand BACSA is not responsible for any plagiarism or academic misconduct.

8. References

[1] Cutting, G. R. (2015). Cystic fibrosis genetics: from molecular understanding to clinical application. *Nature Reviews Genetics*, 16(1), 45-56.

[2] Farinha, C. M., & Callebaut, I. (2022). Molecular mechanisms of cystic fibrosis - how mutations lead to misfunction and guide therapy. *Bioscience Reports*, 42(7), BSR20212006.

[3] Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.

[4] Pettersen, E. F., Goddard, T. D., Huang, C. C., et al. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1), 70-82

[5] Riordan, J. R., Rommens, J. M., Kerem, B., et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245(4922), 1066-1073.

[6] Sharma, N., & Cutting, G. R. (2020). The genetics and genomics of cystic fibrosis. *Journal of Cystic Fibrosis*, 19, S5-S9..

[7] UCSF ChimeraX: Tools for structure building and analysis (2023) Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, Ferrin TE. *Protein Sci.* 2023 Nov;32(11):e4792.

[8] UniProt Consortium (2023). UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523-D531.

[9] Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

[10] Varadi, M. et al. "AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences." *Nucleic Acids Research*, gkad1011 (2023). DOI: 10.1093/nar/gkad1011

[11] Varadi, M. et al. "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models." *Nucleic Acids Research*, 50(D1), pages D439–D444 (2021). DOI: 10.1093/nar/gkab1061

[12] Alberts, B., et al. *Molecular Biology of the Cell*. 6th ed. (General mutation theory).

[13] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*.

[14] Branden, C., & Tooze, J. *Introduction to Protein Structure*. (Impact of amino acid substitutions).

[15] Boucher, R. C. (2007). Evidence for airway surface dehydration as the initiating event in CF lung disease. *Journal of Internal Medicine*.

[16] Cantin, A. M., et al. (2015). Inflammation in cystic fibrosis: Role of the CFTR mutation. *The International Journal of Biochemistry & Cell Biology*.

[17] Moran, A., et al. (2010). Epidemiology, pathophysiology, and management of cystic fibrosis-related diabetes. *Diabetes Care*.