A Machine Learning Model to Forecast Future Antibiotic-Resistant Genes

Grant Proposal

Karena Peterson

Massachusetts Academy of Math and Science

Worcester, MA

Author Note

If needed, write notes here with an indented first line. Be mindful, the text in your lit. review should be between 10-12 font size with a chosen font of Caibri, Times New Roman, or Arial. A Table of Contents is *optional*; however, you should format the section headers appropriately to have them show up in the TOC. Lines should be double-spaced.

Abstract

Antimicrobial resistance is an increasingly paramount issue that deserves a variety of methods to combat. The use of machine learning models for surveillance of antimicrobial resistance utilizes the ability of learning algorithms to recognize patterns in evolution that have otherwise been exploited in microbial evolution modeling; however, current models focus on retrospective knowledge rather than describing future antibiotic resistance genes, thus creating a need. This model focuses on forecasting likely antibiotic resistance genes in *E. coli.*

Keywords: Machine learning models, antibiotic resistant genes, antimicrobial crisis

A Machine Learning Model to Forecast Future Antibiotic-Resistant Genes

Antimicrobial resistance (AMR) is the adaptation of bacteria to antimicrobial medicine and was estimated to be directly responsible for 1.27 million global deaths in 2019 (World Health Organization, 2023). Furthermore, global AMR-related deaths are continuously growing, with an annual increase of 68% to 75% each year (Patra et al., 2025). The prevalence of global resistance levels has also shown an increase, for instance, one study reported the median rates from 76 countries for third-generation cephalosporin-resistant *E. coli* in 2022 was 42%, a rise from the 36% reported by 49 countries in 2020 (World Health Organization, 2020). These increases are driven by several factors, including misuses of antibiotics in humans, plants, and animals, and improper dosing frequency (Patra et al., 2025

Despite the growing ineffectiveness of antibiotics, they remain critical to modern medicine. Notably, antibiotics are vital for surgery, childbirth, infection treatment, and cancer chemotherapy (World Health Organization, 2023). Considering this, the noted increases are even more concerning. Therefore, a diverse set of tools are continuously being developed to fight against the crisis.

Impact of COVID-19 Pandemic on AMR

In the years before the SARS-CoV-2 pandemic, more commonly called COVID-19, attempts to mitigate the growing resistance had been proving more successful (Lee et al., 2013). However, a systematic review of 23 studies done during the pandemic revealed increased resistance in multiple bacterial species (Sulayyim et al., 2022). For instance, one study found that resistance to the antibiotics imipenem, meropenem, and ciprofloxacin had absolute

increases of 32.3%, 36.8%, and 22.6%, respectively (Despotovic et al., 2021). These increases were likely due to self-medication, premature antibiotic administration and improper use of antibiotics within the healthcare system (Sulayyim et al., 2022). This spike resulting from the SARS-CoV-2 pandemic makes AMR even more critical. Thus, the need for tools to combat AMR becomes increasingly vital.

Strategies to combat antibiotic resistance

There are multiple strategies for combatting AMR, such as the development of novel antibiotics. However, this is often expensive and not financially rewarding for pharmaceutical companies, thus discouraging production altogether (Gargate et al., 2025). Alternative treatments have also been explored, but face developmental hurdles (Alaoui Mdarhri et al., 2022). Antibiotic stewardship, the measurement and improvement of how antibiotics are prescribed, is also a valid tactic, though it has been shown to be less effective in recent years (Sulayyim et al., 2022). Finally, many medical professionals and researchers alike utilize surveillance of AMR.

Surveillance of AMR involves the collecting, analyzing, and reporting of data on organism susceptibility patterns, and can inform health policies, such as antibiotic stewardship, and emergency response, improve patient care, and identify long-term trends (Robillard et al., 2024, Rannon et al., 2025). While surveillance systems themselves are not a perfect tactic for AMR, they still provide critical information as to the current state of microbial resistance.

Emerging computational strategies

One way in which surveillance is done is through Machine Learning Models (MLM), which excel at pattern recognition and thus have been applied to the AMR crisis in a multitude of ways. For instance, Machine Learning algorithms have been developed to recognize antibiotic resistance genes (ARG) without the need for a predefined database. Importantly, a MLM that also utilizes biological information to identify truly novel ARGs within a sequence, named Detection of Resistance to AntiMicrobials using Machine-learning Approaches (DRAMMA), demonstrated high performance and was able to successfully annotate gene segments with no known function as ARGs, suggesting their novelty (Rannon et al., 2025). The success of this model in predicting entirely unknown, novel purposes of gene segments suggests the capabilities of MLM for novel gene prediction. However, little to no known research has exploited this to forecast future antibiotic resistance genes, instead focusing on retrospective analysis.

Relevance of Microbial Evolution Modeling

In a separate but related field to AMR surveillance, microbial evolutional modeling also exploits patterns across long periods of time. This allows for the ability of models and principals to characterize evolution quantitively in microbes, with a forward-looking predictive lens. A well-established method, for example, called the Wright-Fisher process can accurately represent the long-term behavior of two competing genetic variants within a population, given certain numerical data under specific conditions. (Good & Hallatschek, 2018). Importantly, these models use known variants to describe future populations, instead of describing future variants as this project proposes. However, these models demonstrate patterns within

microbial evolution, and the ability to quantitively characterize microbial evolution, albeit it with limits. This thus creates a need to test whether a MLM, which excels at pattern recognition, could exploit retrospectively observed structures to make predictions about future variants.

Section II: Specific Aims

This proposal's objective is to create and test if a multi-modal Machine Learning Model will be able to forecast likely characteristics of future ARGs.

Our long-term goal is to proactively inform and improve health responses, where the central hypothesis of this proposal is to test if a Machine Learning Model that considers sequence, function, and mobility will be able to forecast likely characteristic of future ARGs with accuracy exceeding chance-level. These characteristics will include likely mutations spots, likely base-pair mutations, and likely functions. The rationale is that recent advances in MLM for AMR have shown the capabilities of MLM to predict completely novel information, as well as the demonstrated patterns in microbial evolution observed in microbial evolution modeling. The work we propose here will create a multi-modal machine learning model for predicting likely characteristic of future antibiotic-resistant genes.

Specific Aim 1: Train a Tri-Modal Model on ARG Data

Specific Aim 2: Model Forecasting and Validation

Specific Aim 3: Test And Compare Different Model Type Accuracies, Extract Important Information

The expected outcome of this work is a MLM framework based on three distinct types of biological data sources, the forecasting of likely future ARG characteristics with an above chance-level accuracy, and comparison of accuracies of different model types and their resulting information, such as important features for prediction.

Section III: Project Goals and Methodology

Relevance/Significance

The proposed project explores an under researched area in Machine Learning Models in predicting and describing future Antibiotic Resistance Genes. Such descriptions provide highly relevant and applicable information for the healthcare field, especially considering the global struggle with antibiotic resistance.

Innovation

Currently, no known Machine Learning Models address future characteristics of Antibiotic Resistance Genes, despite the evidence of evolution patterns and MLM's capabilities. This proposes project will attempt to address this gap and open future research for forwards forecasting instead of solely retrospective prediction for ARGs.

Methodology

Five datasets for training and validation were downloaded from publicly available NCBI Sequence Read Archive (SRA) and focuses on laboratory *E. coli.* Two datasets consisting of laboratory *E. coli* strains in long-term evolution experiments will be used for training, and one will be split into training and testing data for later use (Crozat et al., 2005; Knöppel et al., 2018). A separate dataset will be collected from a longitudinal study of clinical *E. coli* strains to

help improve versality and avoid overfitting the model during training and will also be split into testing and validation (Kallonen et al., 2017). The last training dataset is on the evolutionarily distinct bacteria, Bacillus subtilis lab strains, to avoid extreme focus on *E. coli*, alone (Brown et al., 2011). An additional dataset not used previously used on bacteria Acinetobacter baylyi will be used to test versatility of the model (Jezequel et al., 2013). These datasets will be annotated using well-established and tested models for gene information, protein function, and mobility indicators. A base line linear regression and random forest model will be tested to find which is optimal. Statistical analysis will be done to assuage accuracy.

Specific Aim #1:

Determine the optimal number of features and optimal feature selection. The objective is to train a tri-modal model on ARG data. Our approach will download four datasets for training from publicly available NCBI Sequence Read Archive (SRA) and focuses on laboratory *E. coli*. Two datasets consisting of laboratory *E. coli* strains in long-term evolution experiments will be used for training. A separate dataset will be collected from a longitudinal study of clinical *E. coli* strains to help improve versality and avoid overfitting the model. Finally, a final dataset on the evolutionarily distinct Bacillus subtilis lab strains will also be used to avoid extreme focus on *E. coli*, alone. These datasets will be annotated using well-established and tested models for gene information, protein function, and

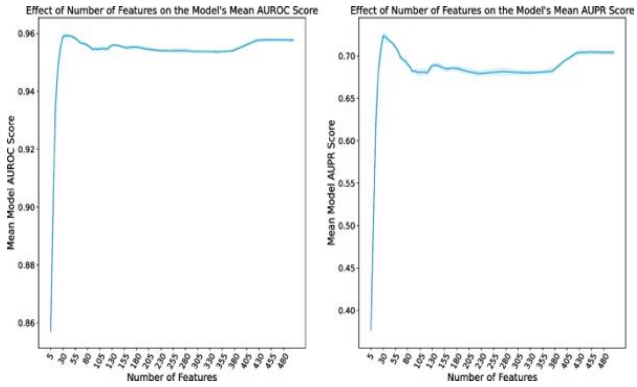

Figure 1 Example effect of number of features on model performance. Effect of number of features on DRAMMA's mean AUROUC score, left. Effect of number of features on the model mean AUPR score, right. (Rannon et al., 2025)

mobility indicators. Our rationale for this approach is that these steps are typically used for Machine Learning Models. Additionally, the data downloaded is from a credible, national source, adding to the integrity. The use of multiple datasets of *E. coli* avoids overfitting the model to one population, and the use of a clinical dataset also helps avoid this. Additionally, while the model more so acts as a proof of concept in laboratory strains, as they are much less noisy, the use of clinical data will act as an indicator of clinical applicability. The separation of the training datasets into both training and testing ensures the testing results are valid and not a result of overfitting. The additional bacteria strain dataset for validation also tests versatility.

The specific number and which features are utilized will be determined by the Machine Learning Model, itself. For instance, Random Forest Models have a built-in feature importance, known as impurity-based importance, which indicates the optimal number of features and which features are most indicative based on the percentage of correct sorting (Figure 1) (Rannon et al., 2025).

Justification and Feasibility. The model is multi-modal, meaning it considers three forms of biological input for final forecasts. For these modals, three feature streams will be used : one focusing on genetic sequencing data, including k-mers, GC content, motifs, and known mutation patterns; another focusing on the functional features, such as protein function, protein structure, and enzyme class; and a final focusing on mobility features, including the presence of transposable elements, plasmid association, and proximity to mobile genetic

Figure 1: Pacific sand lance (*Ammodytes hexapterus*) burrowing into the sand. Mandy Lindeberg, NOAA/NMFS/AKFSC - http://www.photolib.noaa.gov/htmls/fish1917.htm

elements. These features and feature grouping are similar to previous successful works (Rannon et al., 2025).

Summary of Preliminary Data. Will be done later

Expected Outcomes. The overall outcome of this aim is to build a machine learning model that can describe future antibiotic resistance genes with an accuracy higher than chance-level. This knowledge will be used for further fine-tuning to improve accuracy and as a proof of concept of the ability of MLM to forecast specific traits of future antibiotic resistance genes.

Specific Aim

Determine the accuracy of the model. My approach is to test with a variety of statistical analysis tests, including PR-AUC and accuracy matrixes. This meets common practices for analyzing performance of machine learning models.

Justification and Feasibility. PR-AUC is especially considered standard practice for machine learning models handling antibiotic resistant genes as they are better with biased data. Given that ARG are much rarer within a genome than genes serving other purposes, this is especially important. Confusion matrixes are intended for binary classification as they display true positives, false positives, true negatives, and false negatives, as shown in the figure of confusion matrix structure (Horák et al., 2017).

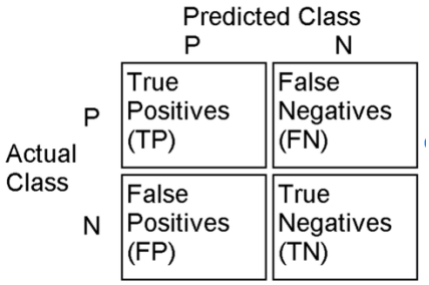

Figure 2 Example confusion matrix structure, where P is positive, N is negative (Horák et al., 2017).

Summary of Preliminary Data. Will be done later

Expected Outcomes. The overall outcome of this aim is to find measures of performance of the model's abilities. This knowledge will be used in the next specific aim to compare substitutional model options performance with this baseline for the optimal options.

Figure 1: Pacific sand lance (*Ammodytes hexapterus*) burrowing into the sand. Mandy Lindeberg, NOAA/NMFS/AKFSC - http://www.photolib.noaa.gov/htmls/fish1917.htm

Specific Aim #3:

Compare and determine multiple model types, different feature selections, and different parameters for optimal accuracy of the model. My approach is to test with a variety of statistical analysis tests, including PR-AUC and accuracy matrixes and then compare from the established baseline in Specific Aims #2. This meets common practices for analyzing performance of machine learning models.

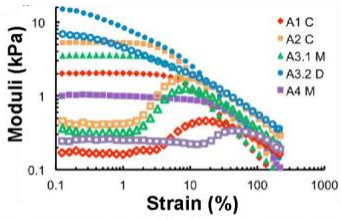

Figure 1.2. Elastic (G', solid symbols) and viscous (G'', hollow symbols) moduli for biofilms grown from the five related bacterial strains that were isolated from one CF patient at different points in time. Strains are listed in order of isolation. Isolates from later timepoints tend to have higher G', except the two mucoid isolates (A3.1 M and A4 M) which have high alginate production and lower G' than their immediate ancestors. The biofilm with highest G' is grown by A 3.2 D which has high Psl production. Figure from

Justification and Feasibility. PR-AUC is especially considered standard practice for machine learning models handling antibiotic resistant genes as they are better with biased data. Confusion matrixes are also especially good for binary classification. The comparison of multiple model types is typical within the field.

Summary of Preliminary Data. Will be done later

Expected Outcomes. The overall outcome of this aim is to find measures of performance of the model's abilities. This knowledge will be used in the next specific aim to compare substitutional model options performance with this baseline for the optimal options.

Figure 1: Pacific sand lance (*Ammodytes hexapterus*) burrowing into the sand. Mandy Lindeberg, NOAA/NMFS/AKFSC - http://www.photolib.noaa.gov/htmls/fish1917.htm

Potential Pitfalls and Alternative Strategies. I expect that data preparation may be difficult and the resulting dataset sizes may be small. This, and downstream analysis for annotated datasets may lead to lower levels of accuracy.

Section III: Resources/Equipment

The model requires very little resources, with the main requirements being datasets and a code space. Google Colab will be used to code the project as it provides necessary storage space. Datasets will be taken from publicly accessible databases.

Section V: Ethical Considerations

As with all Machine Learning Models, it is possible for the gained knowledge to be used in a harmful manner. In consideration of this, code will not be shared freely as to disrupt easy replication.

Section VI: Timeline

Nov 1st - Dec 19th: Build baseline model.

Dec. 19 - Dec. Jan. 1st: Expand upon model

Jan. 1st- Jan. 20th: Use various model types and compare

Jan. 20th – February 10th: Finalizing optimal features, model type, final accuracy testing

Section VII: Appendix

Section VIII: References

Alaoui Mdarhri, H., Benmessaoud, R., Yacoubi, H., Seffar, L., Guennouni Assimi, H., Hamam, M., Boussettine, R., Filali-Ansari, N., Lahlou, F. A., Diawara, I., Ennaji, M. M., & Kettani-Halabi, M. (2022). Alternative therapeutic approaches to conventional antibiotics: Advantages, limitations, and potential application in medicine. *Antibiotics, 11*(12), 1826. https://doi.org/10.3390/antibiotics11121826

Brown, C. T., Fishwick, L. K., Chokshi, B. M., Cuff, M. A., Jackson, J. M., 4th, Oglesby, T., Rioux, A. T., Rodriguez, E., Stupp, G. S., Trupp, A. H., Woollcombe-Clarke, J. S., Wright, T. N., Zaragoza, W. J., Drew, J. C., Triplett, E. W., & Nicholson, W. L. (2011). Whole-genome sequencing and phenotypic analysis of *Bacillus subtilis* mutants following evolution under conditions of relaxed selection for sporulation. *Applied and Environmental Microbiology, 77*(19), 6867–6877. https://doi.org/10.1128/AEM.05272-11

Crozat, E., Philippe, N., Lenski, R. E., Geiselmann, J., & Schneider, D. (2005). Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics, 169*(2), 523–532. https://doi.org/10.1534/genetics.104.035717

Despotovic, A., Milosevic, B., Cirkovic, A., Vujovic, A., Cucanic, K., Cucanic, T., & Stevanovic, G. (2021). The impact of COVID-19 on the profile of hospital-acquired infections in adult intensive care units. *Antibiotics, 10*(10), 1146. https://doi.org/10.3390/antibiotics10101146

Gargate, N., Laws, M., & Rahman, K. M. (2025). Current economic and regulatory challenges in developing antibiotics for Gram-negative bacteria. *NPJ Antimicrobials and Resistance, 3*(1), 50. https://doi.org/10.1038/s44259-025-00123-1

Good, B. H., & Hallatschek, O. (2018). Effective models and the search for quantitative principles in microbial evolution. *Current Opinion in Microbiology, 45*, 203–212. https://doi.org/10.1016/j.mib.2018.11.005

Jezequel, N., Lagomarsino, M. C., Heslot, F., & Thomen, P. (2013). Long-term diversity and genome adaptation of *Acinetobacter baylyi* in a minimal-medium chemostat. *Genome Biology and Evolution, 5*(1), 87–97. https://doi.org/10.1093/gbe/evs120

Kallonen, T., Brodrick, H. J., Harris, S. R., Corander, J., Brown, N. M., Martin, V., Peacock, S. J., & Parkhill, J. (2017). Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Research, 27*(8), 1437–1449. https://doi.org/10.1101/gr.216606.116

Knöppel, A., Knopp, M., Albrecht, L. M., Lundin, E., Lustig, U., Näsvall, J., & Andersson, D. I. (2018). Genetic adaptation to growth under laboratory conditions in *Escherichia coli* and *Salmonella enterica*. *Frontiers in Microbiology, 9*, 756. https://doi.org/10.3389/fmicb.2018.00756

Lee, C. R., Cho, I. H., Jeong, B. C., & Lee, S. H. (2013). Strategies to minimize antibiotic resistance. *International Journal of Environmental Research and Public Health, 10*(9), 4274–4305. https://doi.org/10.3390/ijerph10094274

Naddaf, M. (2024). *40 million deaths by 2050: Toll of drug-resistant infections to rise by 70%.* Nature, 633, 747–748. https://doi.org/10.1038/d41586-024-03033-w

Patra, M., Gupta, A. K., Kumar, D., & Kumar, B. (2025). Antimicrobial resistance: A rising global threat to public health. *Infection and Drug Resistance, 18*, 5419–5437. https://doi.org/10.2147/IDR.S530557

Rannon, E., Shaashua, S., & Burstein, D. (2025). DRAMMA: A multifaceted machine learning approach for novel antimicrobial resistance gene detection in metagenomic data. *Microbiome, 13*(1). https://doi.org/10.1186/s40168-025-02055-4

Robillard, D. W., Sundermann, A. J., Raux, B. R., & Prinzi, A. M. (2024). Navigating the network: A narrative overview of AMR surveillance and data flow in the United States. *Antimicrobial Stewardship & Healthcare Epidemiology, 4*(1), e55. https://doi.org/10.1017/ash.2024.64

Salam, M. A., Al-Amin, M. Y., Salam, M. T., Pawar, J. S., Akhter, N., Rabaan, A. A., & Alqumber, M. A. A. (2023). *Antimicrobial resistance: A growing serious threat for global public health.* Healthcare, 11(13), 1946. https://doi.org/10.3390/healthcare11131946

Sulayyim, H. J. A., Ismail, R., Hamid, A. A., & Ghafar, N. A. (2022). Antibiotic resistance during COVID-19: A systematic review. *International Journal of Environmental Research and Public Health, 19*(19), 11931. https://doi.org/10.3390/ijerph191911931

Tang, R., Luo, R., Tang, S., Song, H., & Chen, X. (2022). Machine learning in predicting antimicrobial resistance: A systematic review and meta-analysis. *International Journal of Antimicrobial Agents, 60*(6), 106684. https://doi.org/10.1016/j.ijantimicag.2022.106684

Ventola, C. L. (2015). *The antibiotic resistance crisis: Part 1: Causes and threats.* Pharmacy and **Therapeutics, 40**(4), 277–283. https://pmc.ncbi.nlm.nih.gov/articles/PMC4378521/

World Health Organization. (2020). *Global antimicrobial resistance surveillance system (GLASS) report: Early implementation 2020* (CC BY-NC-SA 3.0 IGO). World Health Organization. https://apps.who.int/iris/handle/10665/332081

World Health Organization. (2023, November 21). *Antimicrobial resistance*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance