Pathway-Derived Features for Machine Learning-Based Prediction of Multidrug Resistance in Pseudomonas Aeruginosa

Jerry Gao *

Iolani School, HI Honolulu, US

* Corresponding Author Email: jerryogao@gmail.com

Abstract. Antimicrobial resistance (AMR) in Pseudomonas aeruginosa poses a critical threat to healthcare, particularly due to surgical site infections. Current phenotypic testing methods are time-consuming, while existing antimicrobial resistance gene-based machine learning approaches suffer from limited data availability, poor multi-antibiotic prediction performance, and lack of interpretability. This study presents a novel framework using KEGG pathway-derived features to predict multidrug resistance in P. aeruginosa. We annotated whole genome sequences to identify AMR-related metabolic and regulatory pathways, then evaluated these pathway-based features across multiple algorithms, including traditional machine learning methods like logistic regression, random forest, and KNN; deep learning models like MLP, CNN, LSTM, and GRU; and the tabular foundation model TabPFN. All algorithms achieved an accuracy above 0.85, demonstrating the robustness of using pathway-derived features for multi-drug AMR prediction. TabPFN showed superior performance with an accuracy of 0.926, successfully detecting resistance phenotypes across multiple antibiotics simultaneously while providing interpretable insights. In conclusion, the pathway-based features capture complex biological mechanisms underlying resistance beyond simple gene presence/absence, offering a meaningful framework for rapid resistance profiling. We hope this method enables accurate antibiotic selection in clinical settings, potentially reducing inappropriate antibiotic use and combating the spread of the AMR phenomenon.

Keywords: Antimicrobial resistance (AMR); Pseudomonas aeruginosa; KEGG pathway-based features; Machine learning, TabPFN

1. Introduction

1.1. Backgroun

Pseudomonas aeruginosa is an opportunistic, gram-negative pathogen, which can cause a wide range of infections such as pneumonia, urinary tract infections, and septicemia, often associated with high mortality rates, particularly among immunocompromised and surgical patients [1,2]. Unfortunately, while antibiotics are the primary form of treatment against these infections, the effectiveness of these treatments is critically threatened by the rise of antimicrobial resistance (AMR). Without the proper action, AMR can become a serious threat. According to the World Health Organization, AMR killed 4.95 million people in 2019 and can surpass cancer as the leading cause of death by 2050 if no measures are taken [3]. Specifically, the resistance of P. aeruginosa is from two key factors: the intrinsic defense mechanisms of bacteria, such as efflux pumps and β-lactamases [4], and the accelerated evolution of acquired resistance driven by the overuse of antibiotics.

Currently, addressing this global health challenge requires the continued development of novel antibiotics and the rational use of existing antimicrobial agents to limit further resistance evolution. However, both strategies face significant difficulties. The development of new antibiotics is increasingly demanding due to the rapid evolution of bacterial resistance mechanisms and the extended timeframes required for conventional drug discovery [5,6]. At the same time, inappropriate prescribing and excessive antibiotic usage remain widespread and difficult to regulate in clinical treatment and agricultural farming. To overcome this limitation, a promising strategy is the rapid identification of bacterial AMR phenotype patterns, which can enable targeted antibiotic selection for individual patients, thereby reducing the effect of poor antibiotic usage and combating the AMR

phenomenon. Current methods mainly depend on culturing and testing bacteria for their minimum inhibitory concentration (MICs) with various antibiotics, but this process is often time-consuming and incredibly resource intensive [7].

Use of whole genomic sequencing (WGS) and the power of artificial intelligence (AI) models promises a new avenue of attack for the issue of AMR [8–12]. AI models, such as random forest and CNN, can rapidly predict resistance phenotypes in silico, enabling the identification of antibiotics to which a bacterial pathogen is likely resistant without the need for time-consuming laboratory testing. As AI algorithms continue to advance, prediction accuracy and generalizability have significantly improved. Also, model interpretability and feature engineering may help identify key genomic markers associated with resistance mechanisms, offering insights into the molecular basis of AMR [13–15].

That said, recent research still has several limitations. First, there is not enough heterogeneous data available for P. aeruginosa. With such a biased dataset, machine learning models may become unbalanced and less accurate after training [11,12,14]. To improve model robustness, it is important to combine different public datasets. In addition, with a model using these current methods for P. aeruginosa, one would need to construct multiple models for each antibiotic to get a high accuracy; many of these current methods and models are not accurate enough when it comes to making predictions for multiple antibiotics at once. In addition, many studies focus only on predicting AMR based on gene regions, mainly using the presence of antimicrobial resistance genes as the key feature, while large parts of the WGS data, such as non-coding RNAs and regulatory regions, which may also influence resistance phenotypes, are not fully used. Although these methods have identified some important genes related to resistance and can predict AMR phenotypes with reasonable accuracy, they often lack the ability to explain the underlying mechanisms beyond the presence of antibiotic resistance genes. Therefore, it is better to include more biologically-informed features in the AI models from a systems perspective. Finally, most existing studies rely on traditional machine learning models, which have shown good predictive performance in AMR classification tasks. However, with recent advances in artificial intelligence, more complex models such as large language models (LLMs) and foundation models could also be introduced to AMR prediction, offering new opportunities for capturing deeper biological patterns.

1.2. Purpose

This study integrated multiple genomic resources from BV-BRC and proposed a new prediction method based on KEGG pathway features to identify multi-drug antimicrobial resistance in different P. aeruginosa strains. Full P. aeruginosa genomes were annotated, and the resulting gene pathway information was used to train and evaluate various machine learning models, including basic models such as random forests and logistic regression, deep learning like CNN and LSTM, and foundational models like TabPFN. We believe these results will provide both predictive accuracy and biological discoveries and show potential for clinical use in guiding antibiotic selection for P. aeruginosa infections.

2. Materials and Methods

2.1. Data

We collected relevant P. aeruginosa bacterial data from the BV-BRC database with a total of 11,049 strains and their resistance phenotypes to certain antibiotics. To download the data, we used BV-VRC's FTP format and ran a python code to download the data into a computer. We then removed any zero-byte files or abnormal data. This quality control prevents faulty data from negatively affecting the performance of the model and ensures the model works as it should. The bar graph below illustrates the percentages of P. aeruginosa strains that were resistant and susceptible to each of the drugs the model later predicted resistance for.

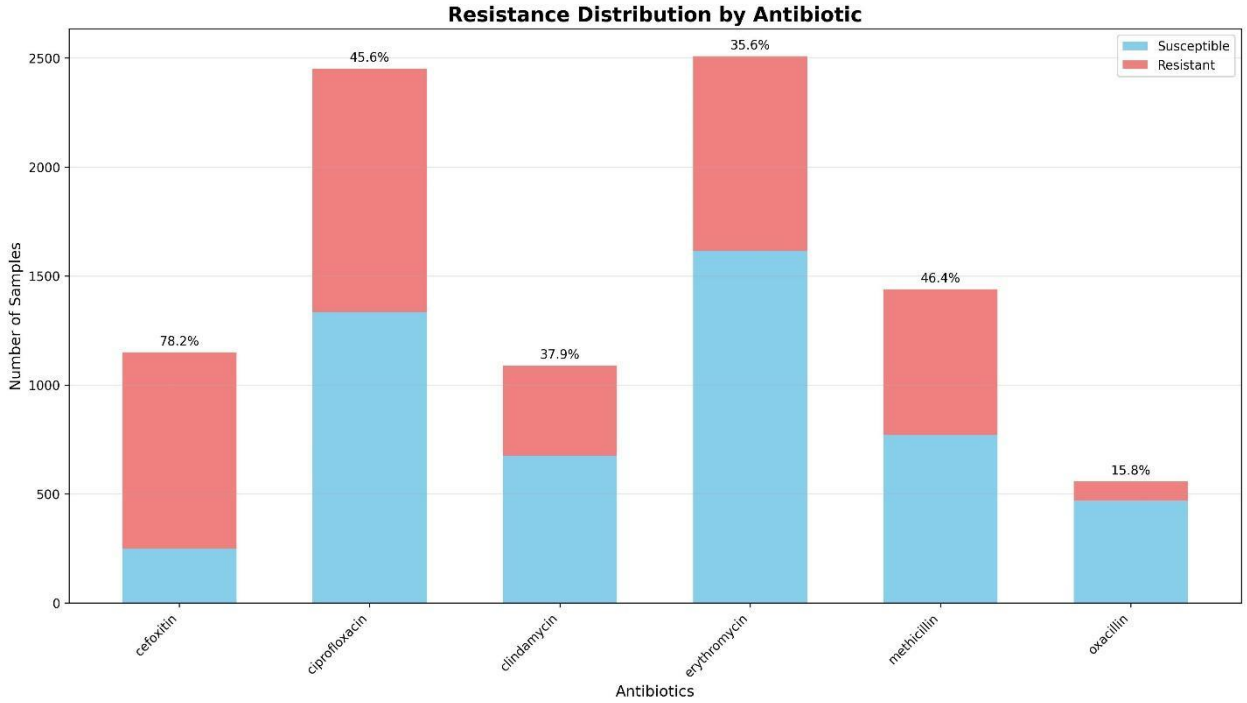

Figure 1: Resistance Distribution by Antibiotic

2.2. Methods

2.2.1 KEGG annotation

To functionally annotate the WGS of all P. aeruginosa strains, we first used Prokka [16], a widely used bacterial genome annotation tool. Prokka predicts and annotates genomic features such as coding sequences (CDSs), rRNAs, and tRNAs and outputs protein sequences in a .faa file. Each CDS in this file represents a predicted gene. In past studies, researchers used the presence or absence of these predicted genes to construct a binary gene matrix as the input for further analysis.

However, Prokka provides only preliminary annotations, and many predicted proteins are labeled as “hypothetical proteins”, meaning they could not be matched to known proteins in reference databases. These may include functionally unknown genes or false positives. To improve annotation quality and gain more functional information, we used eggNOG-mapper, a tool that performs fast functional annotation based on the eggNOG orthology database. It predicts gene functions and assigns categories such as KEGG pathways, cluster of orthologous group (COG) functions, and gene ontology terms. The .faa files generated by Prokka were used as the input to eggNOG-mapper. Since this step involves extensive database comparisons, it usually took longer to run [17].

After additional annotation with eggNOG-mapper, we extracted the KEGG pathway annotations associated with each CDS and generated a KEGG frequency matrix, where each row corresponds to a genome and each column corresponds to a KEGG pathway. The matrix values represent the number of CDSs in each genome assigned to that specific KEGG pathway. This matrix was then used as the feature input for the AMR prediction models.

2.2.2 AI models

To evaluate the predictive performance of different modeling strategies, we tested a range of machine learning models on the KEGG pathway-derived features. These included traditional classifiers, such as logistic regression, random forest, and -nearest neighbors (KNN), which are widely used for their simplicity, interpretability, and robustness on structured tabular data. We also implemented deep learning models, including multilayer perceptrons (MLP), convolutional neural networks (CNN), long short-term memory networks (LSTM), and gated recurrent units (GRU), to explore the nonlinear patterns in biological features. Finally, we tested TabPFN, a recently developed

tabular foundation model based on transformer architecture, which performs end-to-end prediction and gets state-of-the-art performance.

Logistic Regression: This type of model predicts the probability of a binary outcome by fitting a sigmoid curve to the data. Despite its name, it's a useful model for classification problems due to its efficiency and interpretability.

Random Forest: Random forest models combine multiple decision trees trained on random subsets of the data and features to make strong predictions. It reduces overfitting and improves accuracy compared to individual trees, making it useful for both classification and regression.

KNN: This model classifies a sample based on the majority label of its closest neighbors as defined by a set value "k" in the feature space. It is a non-parametric method that makes no assumption about the data distribution and is easy to implement, though performance may vary based on the chosen value of k and feature scaling.

MLP: An MLP (multilayer Perceptron) is a stack of fully connected layers with nonlinear activations. It can model complex relationships in data and serves as a fundamental architecture for many supervised learning tasks. While not specialized for sequences or images, it remains widely used in tabular data and as a building block in larger architectures.

CNN: CNNs (Convolutional Neural Networks) process data using filters that capture local patterns, especially in images. They cut down on parameters by reusing weights and are commonly used for modern computer vision models.

LSTM: LSTM (Long Short-Term Memory) models are a type of recurrent neural network (RNN) that can remember long-term dependencies using gated memory cells. They are widely used in complex tasks like language modeling and time series where earlier inputs affect later predictions.

GRU: GRUs (Gated Recurrent Units) simplify the LSTM design by combining some of its gates, making them faster and often just as effective. They are useful for sequence modeling or simpler tasks where speed is a priority.

Foundation Model: A foundation model is a large-scale neural network (like GPT or BERT) trained on broad data that can be adapted to many downstream tasks. It enables transfer learning at scale and has become a cornerstone of modern AI systems, particularly in natural language processing and vision.

TabPFN: TabPFN (Tabular Prior-Data Fitted Network) is a pretrained transformer model built for tabular classification. Instead of learning from each new dataset, it predicts directly by drawing on prior training across many synthetic tasks. It's fast, only requires some fine tuning, and works well on small datasets [18].

2.2.3 Evaluation index

We tested the models by predicting the AMR phenotypes of P. aeruginosa strains based on KEGG pathway features. For each antibiotic, the model classified whether a strain was resistant or susceptible. Model performance was evaluated using precision, recall, accuracy, and F1-score.

- Precision: Precision quantifies the proportion of positive predictions that are actually correct, calculated as the true positives divided by the sum of the true and false positives.
- Recall: Recall measures the model's ability to capture all actual positive cases, calculated as the true positives divided by the sum of the true positives and false negatives.
- Accuracy: Accuracy reflects the overall proportion of correct predictions.
- F1-Score: The F1-score is the harmonic mean of precision and recall, balancing the trade-off between the two.
- AUC: AUC is the area under the curve, with the "curve" being the receiver operating characteristic (ROC) curve. It measures the model's ability to distinguish the positive and negative outputs.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Precision} = \frac{TP}{TP+FP}$$
$$\text{Recall (Sensitivity)} = \frac{TP}{TP+FN} \qquad \text{F1} - \text{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

3. Results

3.1. Co-Resistance Pattern

To find the relationships between resistance phenotypes, we calculated the pairwise co-resistance rate among six antibiotics based on binary resistance profiles across all P. aeruginosa strains. From Figure 2, we are able to see that ciprofloxacin and cefoxitin have the highest co-resistance rate of 0.61 while methicillin and oxacillin have the lowest co-resistance rate of 0.00. Meanwhile, the co-resistance rate between most of the other pairs of antibiotics ranged from 0.05 to 0.44. These co-resistance patterns reflect potential shared resistance pathways or selective pressure and may inform more effective combination treatment strategies.

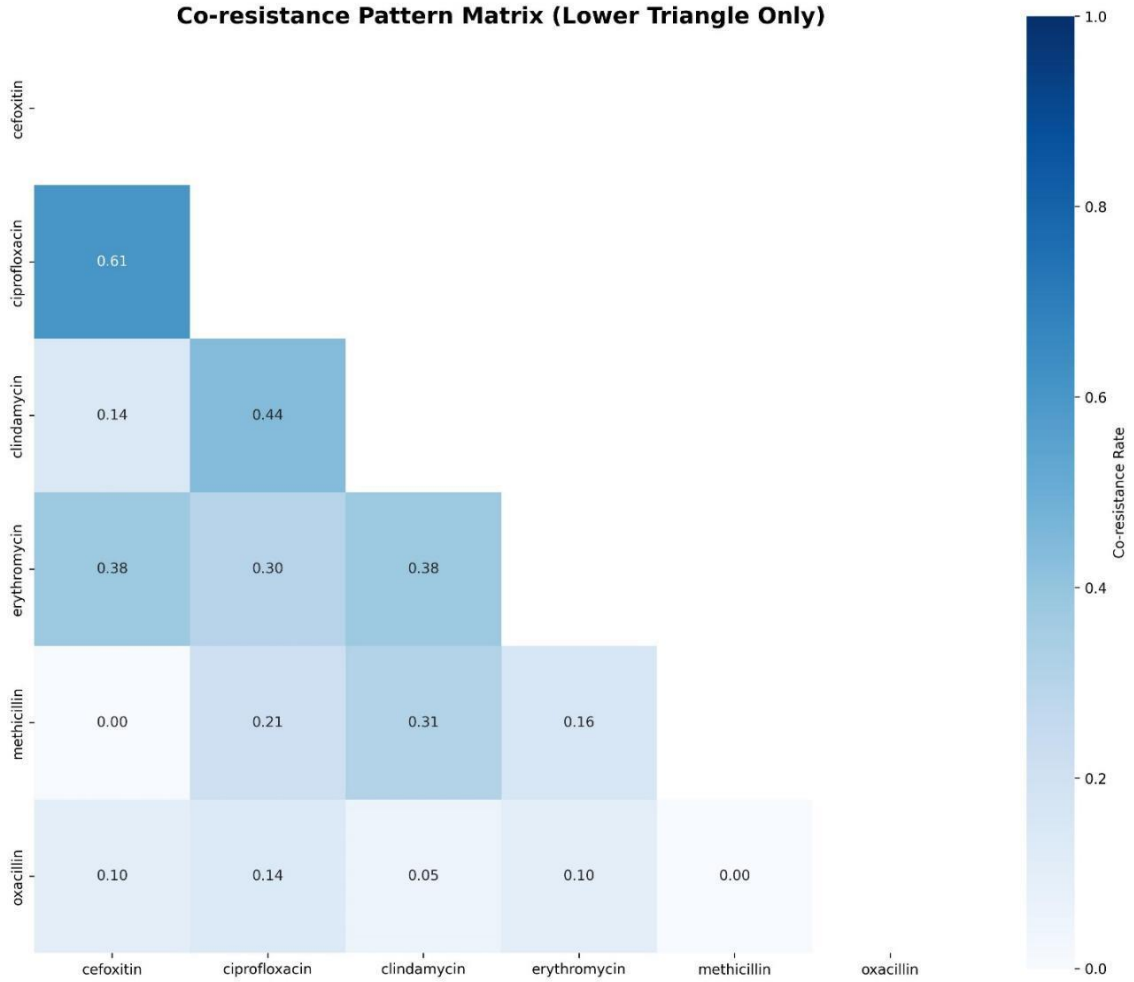

Figure 2: Co-Resistance Pattern Matrix

3.2. KEGG Pathways Results

To visualize the distribution of resistant and susceptible strains, we performed t-distributed stochastic neighbor embedding (t-SNE) on the KEGG pathway-based feature matrix for each antibiotic. As shown in Figure 3, each point represents a strain, colored by its resistance label (red: resistant, blue: susceptible). For some antibiotics like cefoxitin, ciprofloxacin, and clindamycin, resistant and susceptible strains show some separation. For others like erythromycin and oxacillin, the two groups are more mixed. These results suggest that KEGG pathway features can reflect differences in AMR phenotypes.

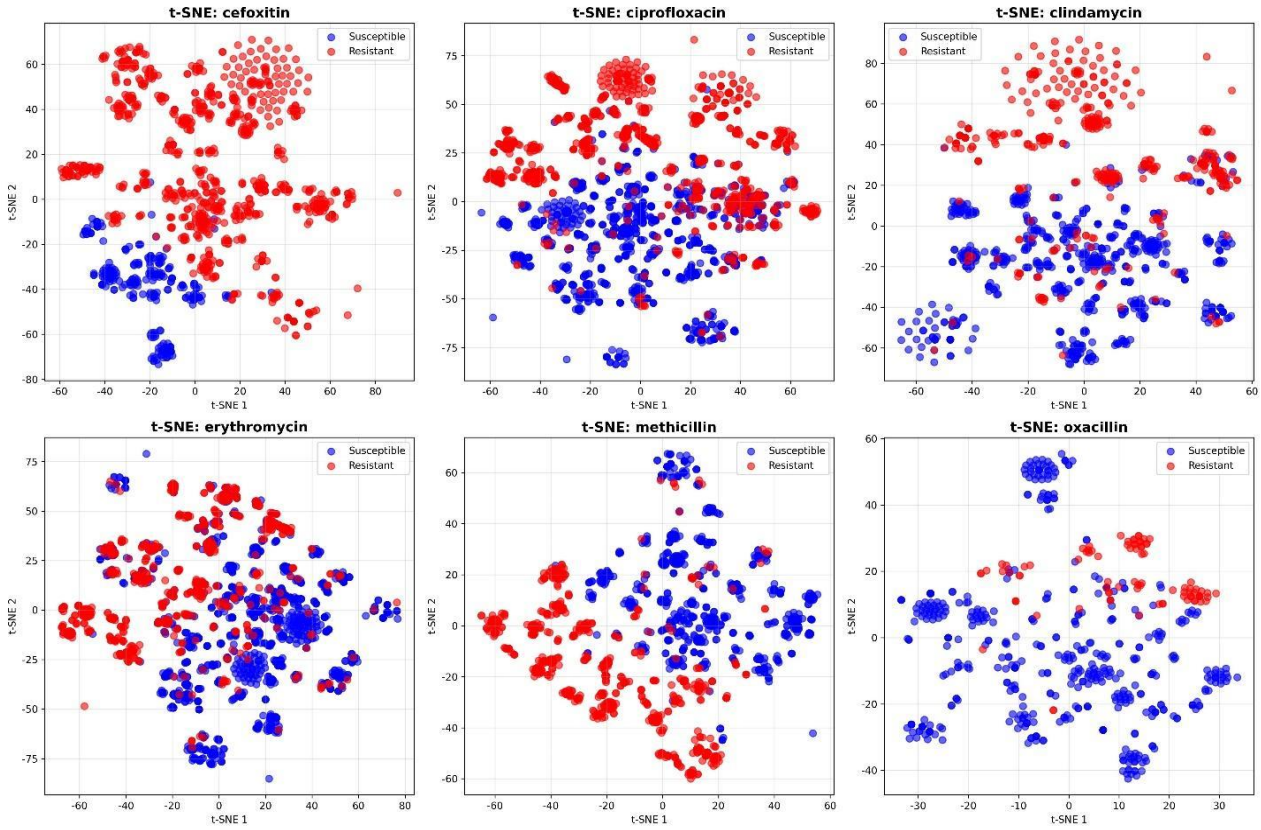

Figure 3: t-SNE Graphs of KEGG Pathways by Antibiotic

3.3. Model Performance

In this section, we compare different algorithms to identify the best-performing model for AMR prediction. The results for how well each model performed are shown in Figure 4 and 5.

The overall performance of each type of model used was decent: the accuracy, precision, and recall scores were high across all the models. All of the models had an accuracy higher than 0.86, as seen in the figures below. That said, the model constructed using TabFPN had the best results overall, having the highest accuracy, recall, f1-score, and AUC. This was most likely because the model had been pretrained on various other data types but could adapt to the KEGG pathway dataset due to how TabFPN worked; the model just had to be finetuned to meet the needs of AMR prediction using the KEGG pathway data.

Surprisingly, models that are commonly regarded as simple, such as logistic regression and random forests, had great results to the point of being better than the results of more complicated deep learning models such as Transformers, LSTM, and GRU. While this seems strange, looking into how each model works explains such results. Traditional prediction models use the presence of AMR genes, but the way the different genes interact can complicate the prediction process and call for more advanced models. However, the models in this experiment used the presence of KEGG pathways for prediction. Since the functions of a single KEGG pathway are often unique to that pathway, the features were chosen well and allowed models like random forest and logistic regression to capitalize on the uniqueness of the features and make very accurate predictions. On the other hand, complex models like Transformers and similar models did not perform well with this type of data. One reason is the limited dataset size, which restricted the learning capacity of these models. Not only that, models like Transformers heavily rely on comparing patterns and relationships between input samples to make accurate predictions. However, in this case, the KEGG pathway features were highly unique across samples, making it difficult for such models to find shared patterns between each of the features.

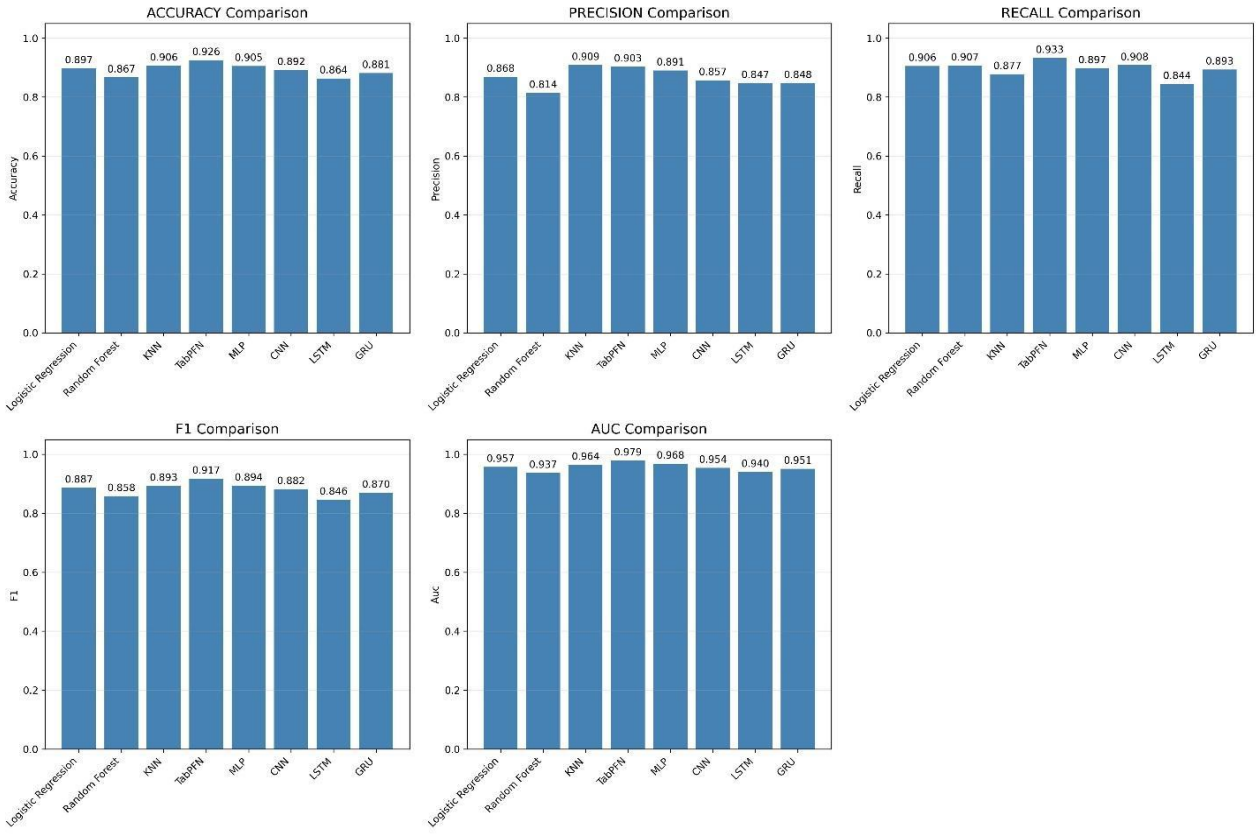

Figure 4: Performance Comparison of Models for AMR Prediction

For single-drug predictions, this can be seen more clearly when the accuracy of each model for each antibiotic is shown in a figure. While all models performed well in predicting antibiotic resistance phenotypes for P. aeruginosa, some models worked better for specific antibiotics. For example, random forest achieved the highest accuracy for most antibiotics, likely because the KEGG pathway-based features effectively captured the dimensional structure of the data. In contrast, MLPs and Transformer-based models did not perform as well. These results are shown in Figure 5 below. Also, these models still showed good accuracy for oxacillin. Logistic regression, decision trees, random forests, and TabPFN often performed as well as or better than models like LSTM and GRU, which usually need more data. Random forest and TabPFN even reached 100% accuracy for cefoxitin. Although KEGG pathway features helped simpler models do well, the small sample size means it is still unclear whether they are truly better than complex models like Transformers. Overall, TabPFN had the best performance across all drugs and metrics.

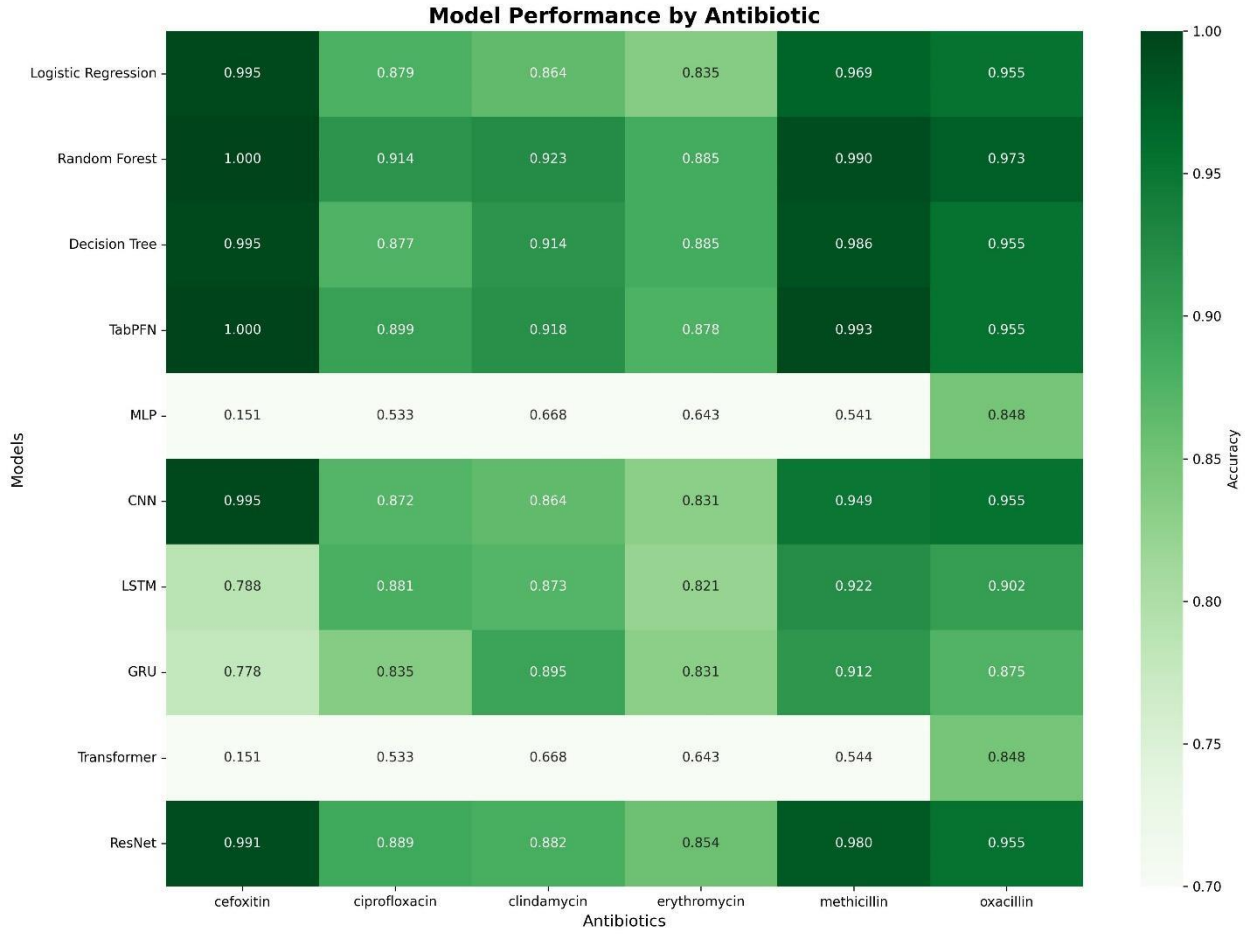

Figure 5. Model performance for AMR prediction across single antibiotic

3.4. AMR-related KEGG Feature Analysis

While the use of KEGG pathways as features in the prediction model resulted in high-quality predictions, it also enabled the identification of important pathways associated with AMR in P. aeruginosa. By leveraging the way foundation models like TabPFN process KEGG data and applying lasso regression to estimate the contribution of each pathway to the final prediction, we produced the results shown in Figure 6.

The KEGG pathways determined to be important for AMR makes sense. For example, the highest ranking KEGG pathway, K02545, is a pathway called penicillin-binding protein 2 prime symbolized by mecA. This pathway is involved in peptidoglycan synthesis and has metabolic functions, but it also takes part in building resistance against beta-lactum antibiotics such as penicillin and methicillin. The second and third KEGG pathways also have functions in providing resistance against certain antibiotics.

Finally, while some of the other KEGG pathways in this list aren't clearly related with AMR, further research can perhaps reveal other genes these pathways are related to that have resistance-inducing roles.

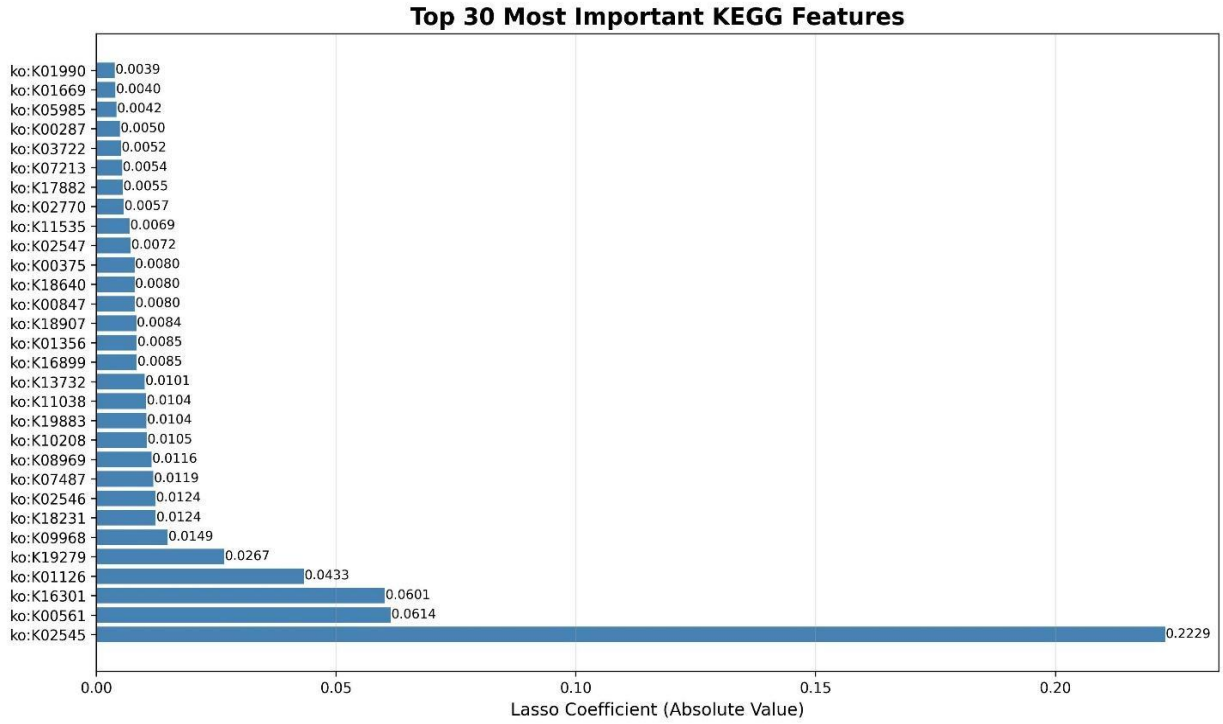

Figure 6: Top 30 Most Important KEGG Features Related to AMR

4. Discussion and Conclusion

Many past studies used AMR genes to build models for predicting antimicrobial resistance. These methods often had problems such as limited data, poor performance across different antibiotics, and low interpretability. In this study, we used KEGG pathway features instead, which helped improve model performance in predicting resistance in P. aeruginosa. All models reached over 0.85 accuracy, and TabPFN showed the best results across all metrics. This shows that KEGG pathways are useful features for AMR prediction, even when the data is limited. In addition, the model is able to show the most important KEGG pathways, such as how K02545, a pathway known for causing antimicrobial resistance in other bacterial species, has a lasso coefficient of 0.2229. Future studies can look into the important KEGG pathways found by the model and apply the method to other resistant bacteria, which may help test its general use across different species as well as discover the AMR mechanisms.

References

[1] Krell T, Matilla MA. Pseudomonas aeruginosa. Trends Microbiol 2024; 32:216–218

[2] Elfadadny A, Ragab R F, AlHarbi M, et al. antimicrobial resistance of Pseudomonas aeruginosa: navigating clinical impacts, current resistance trends, and innovations in breaking therapies[J]. Frontiers in microbiology, 2024, 15: 1374466.

[3] Murray CJL, Ikuta KS, Sharara F, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet 2022; 399:629–655

[4] Holmes AH, Moore LSP, Sundsfjord A, et al. Understanding the mechanisms and drivers of antimicrobial resistance. The Lancet 2016; 387:176–187

[5] de la Fuente-Nunez C. Antibiotic discovery with machine learning. Nat Biotechnol 2022; 40:833–834

[6] Harbottle H, Thakur S, Zhao S, et al. Genetics of Antimicrobial Resistance. Anim Biotechnol 2006; 17:111–124

[7] Fluit AC, Visser MR, Schmitz F-J. Molecular Detection of Antimicrobial Resistance. Clin Microbiol Rev 2001; 14:836–871

[8] Bilal H, Khan MN, Khan S, et al. The role of artificial intelligence and machine learning in predicting and combating antimicrobial resistance. Comput Struct Biotechnol J 2025; 27:423–439

[9] Köser CU, Ellington MJ, Peacock SJ. Whole-genome sequencing to control antimicrobial resistance. Trends in Genetics 2014; 30:401–407

[10] Green AG, Yoon CH, Chen ML, et al. A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis. Nat Commun 2022; 13:3817

[11] Zhao S, Tyson GH, Chen Y, et al. Whole-Genome Sequencing Analysis Accurately Predicts Antimicrobial Resistance Phenotypes in Campylobacter spp. Appl Environ Microbiol 2016; 82:459–466

[12] Mahfouz N, Ferreira I, Beisken S, et al. Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. Journal of Antimicrobial Chemotherapy 2020; 75:3099–3108

[13] Jia H, Li X, Zhuang Y, et al. Neural network-based predictions of antimicrobial resistance phenotypes in multidrug-resistant Acinetobacter baumannii from whole genome sequencing and gene expression[J]. Antimicrobial Agents and Chemotherapy, 2024, 68(12): e01446-24.

[14] Kim J I, Maguire F, Tsang K K, et al. Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective[J]. Clinical microbiology reviews, 2022, 35(3): e00179-21.

[15] Wang S, Zhao C, Yin Y, et al. A practical approach for predicting antimicrobial phenotype resistance in Staphylococcus aureus through machine learning analysis of genome data[J]. Frontiers in Microbiology, 2022, 13: 841289.

[16] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014; 30:2068–2069

[17] Cantalapiedra CP, Hernández-Plaza A, Letunic I, et al. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 2021; 38:5825–5829

[18] Hollmann N, Müller S, Purucker L, et al. Accurate predictions on small data with a tabular foundation model. Nature 2025; 637:319–326