ML Approach for Predicting the Environmental Origin of Metagenomic Samples Using Resistome Analysis

Nir Borger (ID: 313580920)

Submitted as a project report for David Burstein's Lab, Tel-Aviv University

Abstract

The use of antibiotics in healthcare and agriculture has undeniably saved numerous lives by combatting bacterial infections. However, this excessive consumption has inadvertently led to the emergence of antibiotic resistance, diminishing the efficacy of once-potent drugs. This report employs advanced machine learning models to analyze metagenomic samples and predict their environmental origins based on antibiotic resistance patterns. Results have shown that the model successfully predicted the source of samples from well-represented metagenomic environments. Feature importance analysis showed correlation between the importance of the feature and its frequency within the data distribution.

This research not only enhances our comprehension of how antibiotic resistance spreads in various environments but also offers invaluable insights into the associated risks. These findings underscore the urgency of monitoring and mitigating antibiotic resistance within diverse ecosystems.

Introduction

Antimicrobial drugs have unquestionably transformed the landscape of modern medicine since their discovery in the beginning of the 20th century. Not only saving countless lives by effectively combating bacterial infections, but allowing new medical procedures such as surgery, organ transplantation and cancer chemotherapy[1]. The distinctive capabilities of antibiotics substances lie in their ability to inhibit and even eliminate the growth of bacteria without damaging the host's cells and tissues. The discovery of these drugs marked a pivotal moment in healthcare, ushering in an era where previously fatal diseases could be tamed with a simple prescription.

Antibiotics remain one of the most frequently prescribed classes of drugs, with a 36% increase in antibiotic consumption observed through the years 2000-2010[1]. One facet of the excessive consumption of antibiotics stems from their inappropriate use: treating patients and animals with antibiotics without a necessary need for the useful medicine[2]. Moreover, many antibiotic drugs are being sold over the counter without prescription or a presence of a documented clinical need[2]. Even so, in cases where prescription is needed, doctors often prescribe them as a precaution, without the use of sufficient diagnosis tools[3].

This extensive use of antibiotics has given rise to a concerning counter phenomenon: the emergence of antibiotic resistance genes (ARGs). Bacteria, demonstrating their adaptability, have devised ingenious mechanisms to withstand the presence of antibiotics. These mechanisms, encoded within the genomes of bacterial populations, represent a formidable response to our most potent medical treatments.

While the emergence of antimicrobial resistance (AMR) is not a recent occurrence, its rapid propagation in recent years has catapulted it into a global crisis.

In Europe, it is estimated that approximately 25,000 lives are claimed each year due to infections caused by multidrug-resistant bacteria. The economic repercussions on the European Union amount to an annual burden of around 1.5 billion euros. Meanwhile, in the United States, more than 2 million individuals grapple with infections caused by antibiotic-resistant bacteria annually, resulting in 23,000 deaths directly attributed to these infections[4].

It is also crucial to emphasize that should antibiotics lose their effectiveness, the potential for casualties could rise significantly. Vital medical procedures, such as surgeries and chemotherapy, may become perilous, further complicating the battle against infections.

Finding new antibiotics and antibiotic mechanisms is a formidable task, as demonstrated by the fact that no new class of antibiotics has been discovered for decades[3]. In addition, as the rate of discovery of new antibiotic classes is decreasing, the prevalence of ARGs is increasing. The human race is at a disadvantage in this struggle, and investments in this research are inadequate. Therefore, a different approach is required, such as AMR surveillance.

Antibiotic resistance genes often originate in environmental microbiota due to the presence of antibiotics. These genes can be transferred to pathogens through horizontal gene transfer (HGT), making it crucial to study them. Metagenomics, the study of genetic material recovered directly from environmental samples, is a key method for analyzing these genes in environmental microbiota. Metagenomics allows for the examination of microbial communities and their resistance genes, providing insight into the spread of antibiotic resistance and its impact on health and ecosystems.

Metagenomic studies have substantially enhanced our comprehension of microbial communities and their functional roles. For example, metagenomic approaches have been instrumental in identifying novel enzymes that are vital for biocatalytic processes within microbial communities[5].

Another notable illustration of the significance of metagenomics lies in the study of the human microbiome. Metagenomic research has established a correlation between microbial communities and health conditions, including obesity and antibiotic resistance. Recent advancements in AMR research have led to significant developments, including the discovery of new antibiotics targeting previously resistant bacteria and the identification of environmental factors contributing to AMR spread. These discoveries are pivotal in designing more effective treatments and enhancing global surveillance systems to monitor and respond to AMR threats.

Different environments have distinct impacts on bacterial communities studied through metagenomics. Factors such as temperature, pH, moisture, and nutrient availability can shape the diversity and function of these communities. For instance, extreme environments like hot springs host thermophilic bacteria, while acidic soils favor acidophilic microbes. Similarly, gut microbiomes vary based on diet and host genetics. Understanding these environmental effects is crucial for applications in biotechnology and environmental conservation.

When it comes to antibiotic resistance genes, distinctive characteristics are exhibited in various ecosystems, shaped by selective pressures and co-evolution with bacteria. Understanding these patterns is essential in unraveling the complexities of resistance.

An example for previous efforts to predict environment of samples based on the sequenced alignments made upon these samples is the Metabolomic Analysis of Metagenomes using flux Balance Analysis (FBA) and Optimization (MAMBO) approach. It does so by analyzing microbial genome distributions and their metabolic potential in metagenomic data. MAMBO employs Genome-Scale Metabolic Models (GSMMs) to establish a correlation between the growth of microorganisms and their abundance in metagenomic profiles. This allows the inference of available metabolic resources in various environments. The method has been effectively applied to human body tissues like oral, skin, stool, and vaginal samples, demonstrating its ability to predict and cluster metabolomic data consistent with specific environmental characteristics[6].

Machine learning, a branch of artificial intelligence, is revolutionizing various fields by enabling computers to learn from data and make predictions or decisions without being explicitly programmed. Its applications range from image and speech recognition to medical diagnoses, enhancing both efficiency and accuracy in complex tasks.

In this context, the development of machine learning models that can predict the environmental origins of sequenced samples becomes not only meaningful but imperative. Such models can decipher the subtle but informative signatures of antibiotic resistance factors within protein samples. This predictive capability offers insights into the origins and potential risks associated with antibiotic resistance, bolstering our efforts to combat its spread.

To achieve this goal, a Random Forest model was trained, as it offers several advantages in machine learning. Random forests are an ensemble learning method that combines the predictions of multiple decision trees, resulting in a robust and accurate model. One key benefit is their ability to handle both classification and regression tasks with ease, making them versatile for various types of data analysis. Additionally, random forests are resilient to overfitting, a common concern in complex models, thanks to their built-in mechanism of averaging predictions across multiple trees. Moreover, they are less sensitive to outliers and noisy data, contributing to their reliability in real-world applications. Furthermore, random forest yields good results on a high-dimensional data such as this and provides feature importance scores[7].

The model achieved good performance in both cross-validation evaluation and in a validation on an external dataset. Analysis of the features chosen by the model has surprisingly shown that the most important features were ones related to specific bacterial targets of various antibiotics, while no feature describing drugs class or mechanism of resistance was chosen.

In conclusion, this study provides valuable insights into the transmission dynamics of AMR across diverse ecosystems. The findings could inform strategies to mitigate the spread of AMR, enabling us to identify and manage environmental risk factors more effectively. Furthermore, this research paves the way for a deeper understanding of

the environmental determinants of AMR propagation, which is crucial for developing targeted interventions to safeguard public health.

Methods

Dataset Composition

To construct the dataset, metagenomic samples were taken from various sources: NCBI's WGS dataset[8], the Mgnify dataset[9] and metagenomic samples from various sources that were assembled by the lab ("Denovo" dataset). WGS and Mgnify were used to train the model, while the Denovo dataset was used as the test set. Each database contains protein sequences from different environments.

In order to create features for each sample, a Hidden Markov Model (HMM) search was executed against 173 HMM profiles of resistance genes families. These profiles originate from Resfams, a curated database containing protein families authenticated for antibiotic resistance function and their corresponding profile HMMs[10]. Only hits that attained an e-value of no more than 10^{-10} were kept. The results were then mapped to four criterions: (1) gene families by identifier, (2) gene families by accession, (3) resistance mechanisms and (4) antibiotics to which the genes conferred resistance.

To facilitate dealing with the large number of features, gene families and identifiers were consolidated into groups, resulting in a smaller number of more informative features. Gene families were originally labeled by their Resfams identifiers (e.g., RF0004), therefore, after the feature processing, the families' names were given instead. These consolidated features were added to the table, making 6 criterions. This information was meticulously sourced from the Comprehensive Antibiotic Resistance Database (CARD), which houses a repository of information concerning resistance genes, comprising their familial classification, associated products, and correlated phenotypic manifestations[11].

Upon the collection of all the properties mentioned above, the model's features were generated by calculating, for each sample, the frequency across each criterion. Moreover, the label of each sample (i.e., the environment the sample originated from) was extracted from the sample's metadata and mapped to one of five predefined environments.

During the model training and testing, in order to ensure high-level testing, it was essential to synchronize features for both train set and test set. Therefore, all features that appeared in one of these sets alone were removed.

Since the data was highly unbalanced, the "Human Microbiome" environment was under sampled randomly from ~5300 to ~870, thus ensuring smaller differences between environments. Moreover, both the train and the test sets had an unmistakable small number of samples labeled as "Groundwater", hence, these samples were removed from both groups. The final sizes of training and test datasets were as follows:

	Human Microbiome	Soil	Animal Microbiome	Sewage
Training Set	873	432	115	84
Test Set	447	370	28	229

Data Analysis

Following the creation of the feature table, the distribution of different properties across the various environments was analyzed. The training data was then visualized as bar plots for each criterion, demonstrating the frequency of each feature in each environment.

Concurrently, UMAP, a dimensionality reduction algorithm was applied to the feature table. UMAP serves as an effective tool for dimension reduction, facilitating the visualization of the frequencies table. The resultant UMAP projections elucidate the distribution patterns of the training samples across environments, offering a clear depiction of the relative similarities between different environments.

Model Training & Evaluation

A Random Forest model from the Scikit-learn[12] library in Python was trained on the features extracted from metagenomic samples sourced from Mgnify and WGS.

The training process was divided into two stages: first, evaluation of the model, using 5-fold cross validation (CV), and second, training the model on the entire training set and testing it on the external test set.

5-fold cross-validation involves dividing a dataset into five subsets, training and evaluating the model five times with different subsets as the testing set. This method enhances performance assessment by reducing variability, offering a more reliable estimate of generalization. Benefits include a comprehensive evaluation of model generalization, minimized risk of overfitting or underfitting, and accurate performance representation across diverse scenarios.

During the training process, feature selection was made for each fold to improve model performance. By removing redundant or irrelevant features, and thus reducing dimensionality, feature selection streamlines the model, which can lead to faster training times and better generalization to new data.

In addition, by using the built-in feature importance function of a specific model in the scikit-learn library, the 20 most contributing features were identified. Feature importance quantifies the contribution of each individual feature to the predictive power of a model, highlighting which variables have the most significant impact on the model's decisions. This knowledge not only enhances model interpretability, allowing for more informed decision-making, but also guides the refinement of the model by focusing on the most relevant features.

Testing involved training the model on the training set as a whole, and then testing on the test set by predicting the probabilities of each environment.

The performance of the model during evaluation and testing was measured by the Area Under the Precision-Recall curve (AUPR) and the Area Under the Receiver Operating Characteristics curve (AUROC). These two metrics are utilized to assess the model's abilities.

Results

Distribution of AMR in Metagenomic Samples

After constructing the table from each sample in the training dataset, the frequencies of the features were calculated for each sample and for each criterion separately. The data collected described above yielded the following results: ATP-Binding Cassette (ABC) Transporter is the most frequent genes family, followed by CPT (Ceftaroline) (Figure 1 a). Moreover, ABC Transporter is also the most frequent mechanism while other effluxes are the 2nd most frequent (Figure 1 b). In terms of drug classes, the Glycylcycline is most abundant drug bacteria has developed resistance to (Figure 1 c).

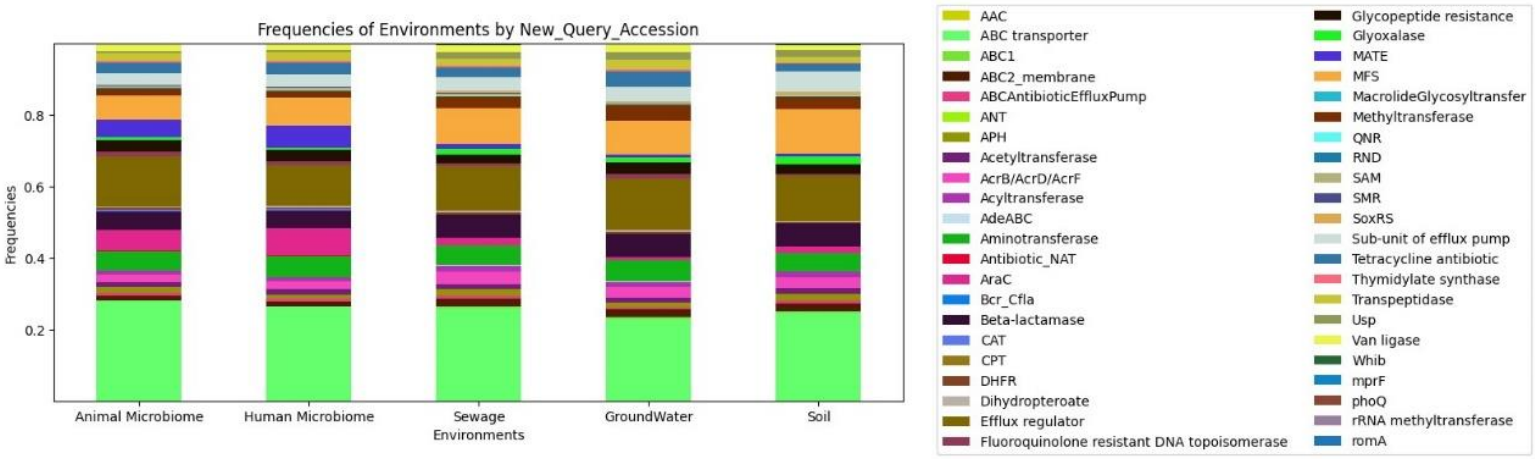

Figure 1 a

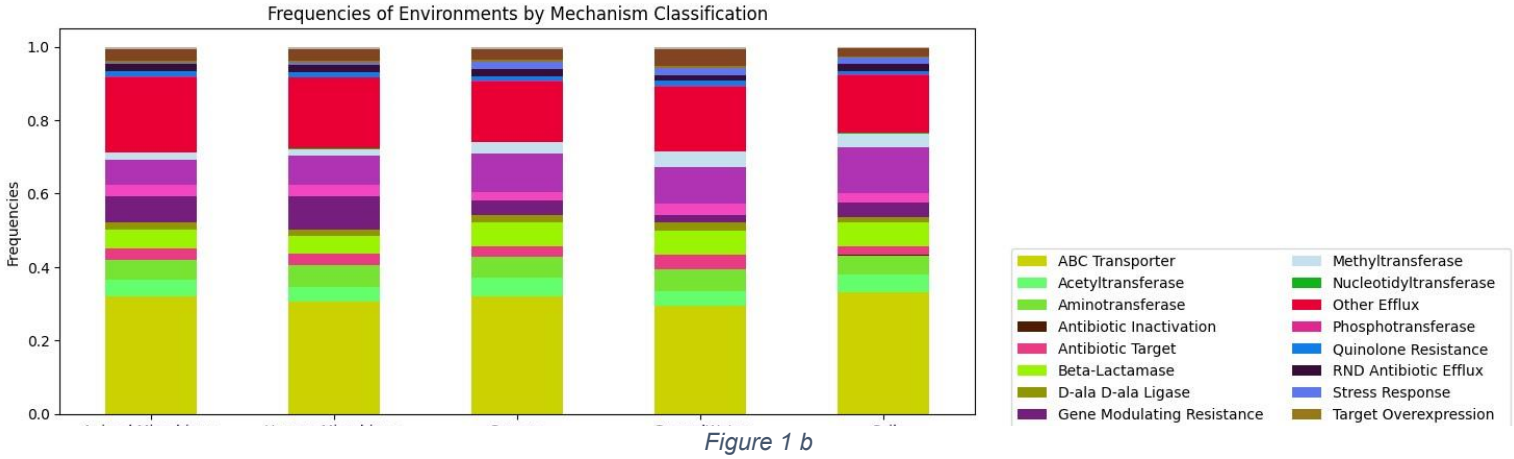

Figure 1 b

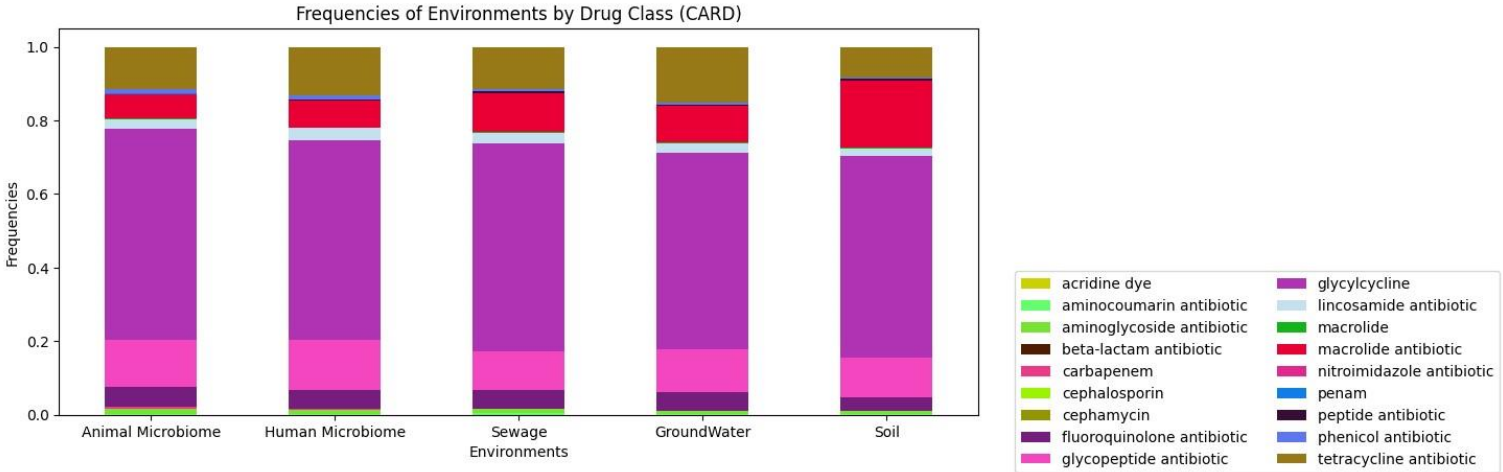

Figure 1 c

Figure 1: Distribution of Frequencies of Environments by Various Criteria

Relativeness of environments

Following the performance of dimensionality reduction on the frequencies table of the training dataset, the UMAP algorithm was applied to explore the relationships in the data. From the projections, it appears that none of the environments cluster together, meaning that they are probably well distinguished by the features collected. Figure 2 is an example for such distance, demonstrated by Drug Class relativeness.

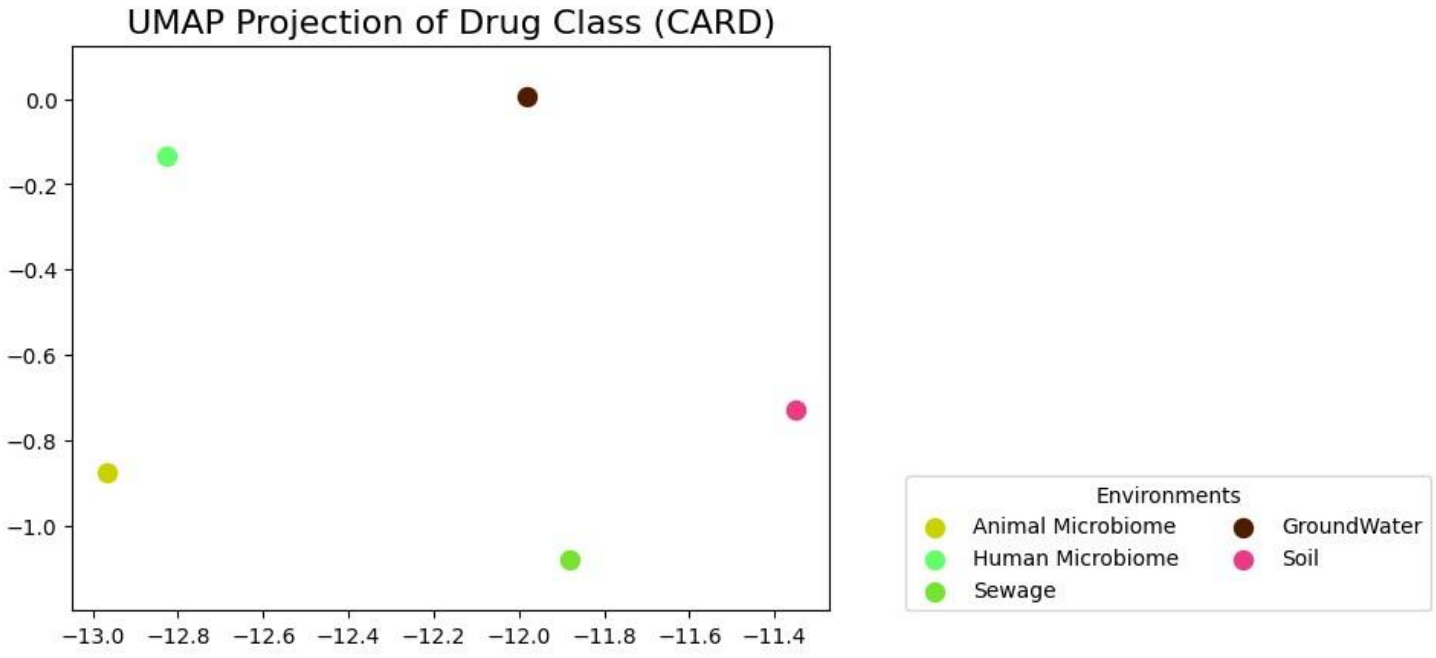

Figure 2: UMAP Projection of Environments Relativeness by Drug Class

Environment Prediction Results on the Training Set

First, the model was evaluated using five-fold cross validation. As results from the model evaluation shows, in terms of Receiver Operating Characteristic (ROC), the model exhibited a good performance across all environments, especially on the human microbiome and soil environment, reaching a mean AUC score of 96% and 94% respectively (Figure 3).

In terms of Precision-Recall (PRC), the model performed best on the human microbiome environment, attaining a mean AUC of 97% with near-zero standard deviation (Figure 4). On the other hand, the model had the worst performance on the sewage environment, presenting a mean AUC score of 40% with a standard deviation of 11% across the five folds. The trade-off between the recall and precision of the model can be well-observed across all environments, but it is especially prominent in the animal microbiome and sewage environments.

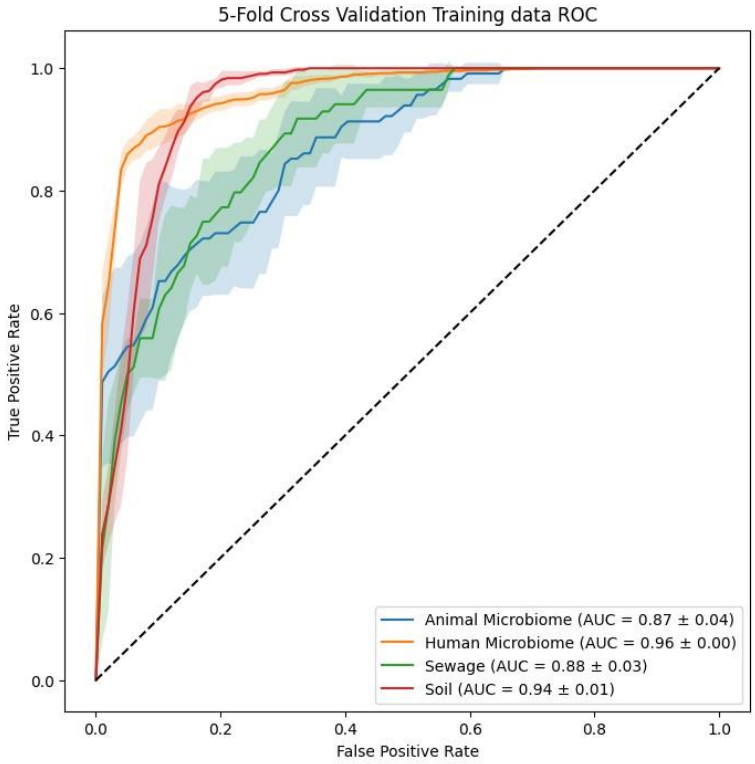

Figure 4: Precision-Recall on Training Set With 5-Fold Cross Validation

Environment Prediction Results on the Testing Set

Training the model on the entire training set and subsequently predicting on the test set, which contains samples gathered by the lab, yielded the testing results. For the test set, the model acts differently. In the ROC graph, the model exhibited a more balanced performance across the different classes. While the model performed the poorest on the animal microbiome with an AUC of 68%, the besr performance was observed on the human microbiome and soil environments, with a mean score 98% and 97% respectively (Figure 5).

In terms of PRC, the performance of the model on the animal microbiome environment was low, achieving a mean AUC score of 4%. It can be shown that the precision-recall curve of this class is very high and drastically drops with a peak-drop pattern that gets more subtle as Recall is getting higher. The model performed well on the other environments, exhibiting a good trade-off between recall and precision (Figure 6).

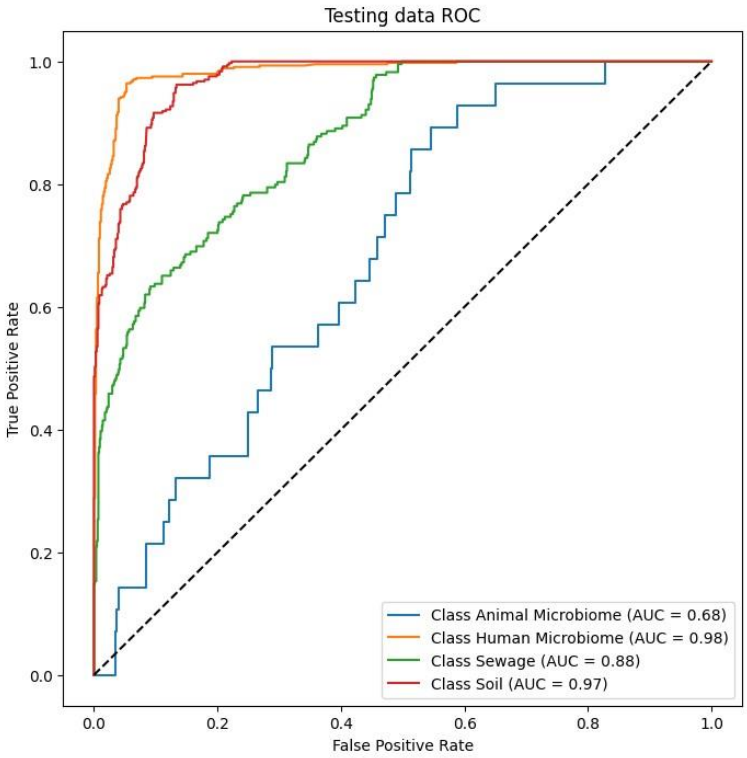

Figure 5: Receiver Operating Characteristic on Test Set

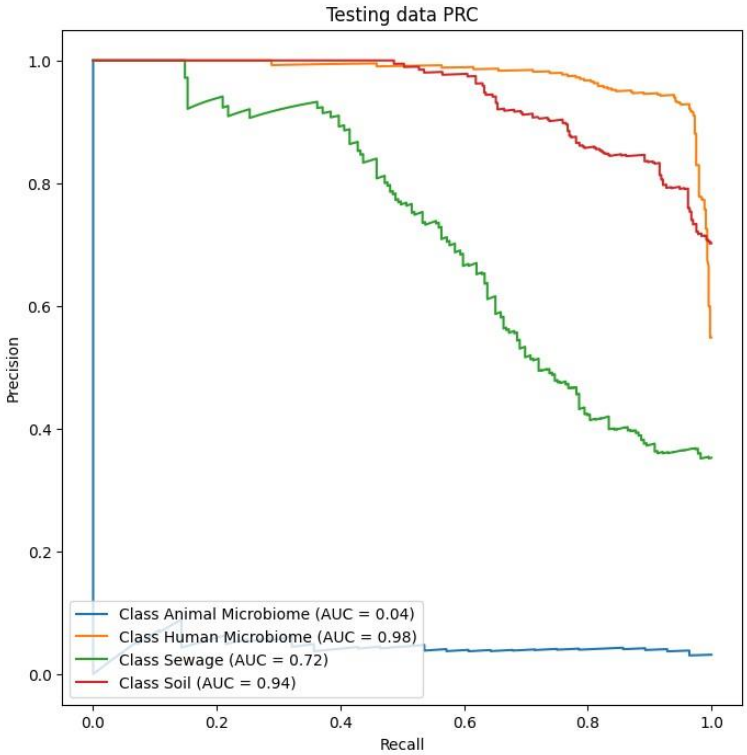

Figure 6: Precision-Recall on Test Set

Feature Importance Analysis

During model evaluation, the built-in feature importance function of scikit-learn was utilized to identify the 20 most influential features. There are four features that stand out in terms of feature importance: (1) Lactamase B, an enzyme produced by bacteria that confers resistance to β-lactam antibiotics, (2) ANT9, a class of aminoglycoside nucleotidyltransferases, (3) RND Efflux, Resistance-Nodulation-Division efflux systems involved in multidrug resistance, and (4) Antibiotic NAT, nucleotidyl transferase, enzymes that inactivate antibiotics by transferring a functional group to the antibiotic molecule. Other features are somewhat similar in importance. These additional features, while not as dominant as the aforementioned ones, contribute to the overall predictive capability of the model. They represent various aspects of antibiotic resistance, such as different mechanisms of action, drug classes, and bacterial targets.

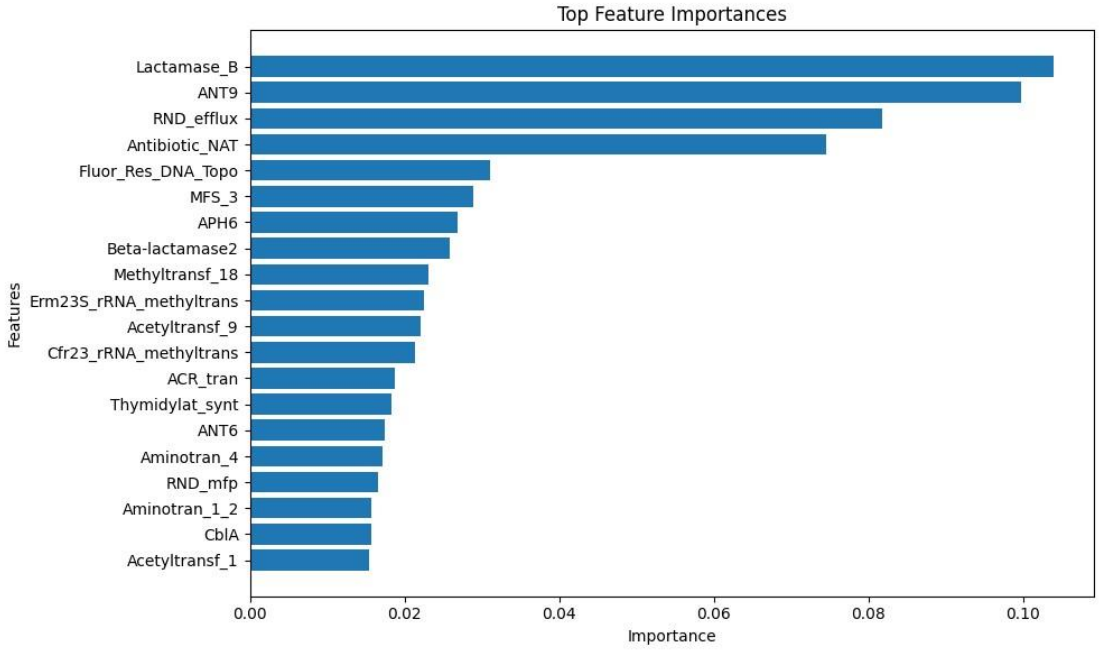

Figure 7: Top 20 Important Features

Discussion

Since their inception in the early 20th century, antibiotics have revolutionized medicine, enabling lifesaving treatments and complex medical procedures. However, their widespread and often inappropriate use has led to a dramatic increase in antibiotic resistance, posing a global health threat. This resistance is exacerbated by the horizontal gene transfer of resistance genes from environmental microbiota to pathogens. Metagenomic studies have been crucial in understanding microbial communities and their resistance mechanisms, leading to advancements in antibiotic research and AMR surveillance[13]. In this study, a machine learning model, specifically a Random Forest algorithm, was employed to predict the environmental origins of metagenomic samples based on resistance patterns. This approach has shown that the most significant features for prediction are related to bacterial targets rather than drug classes or mechanisms. The study's findings highlight the importance of environmental factors in the spread of antibiotic resistance and inform strategies for public health interventions.

Data Distribution

Analysis of the distribution of gene families across different environments showed that ABC Transporter and CPT are the most abundant genes families (Figure 1 a). CPT is a chloramphenicol phosphotransferase that inactivates chloramphenicol by modifying its hydroxyl group[14]. Chloramphenicol (Cm), an antibiotic produced by the soil bacterium *Streptomyces venezuelae*, is phosphorylated by CPT, resulting in its reversible inactivation[15]. Considering the fact that bacteria can perform Horizontal Gene Transfer (HGT), it is possible that the gene providing resistance to chloramphenicol is transferred to other bacteria as a result of the mentioned selective pressure. ABC Transporter is a family of effluxes that are found in membranes of bacteria. With ATP-binding and hydrolysis, ABC-transporters move various substrates across cellular membranes, including harmful materials such as antibiotics. Aside from the fact that ABC-transporters can be found in all domains of life[16], the fact that they can bind to a variety of molecules and ions also serves as an explanation to their high frequency.

It is possible that various antibiotics can bind to these efflux pumps and be exported from the prokaryotic cell, thereby conferring resistance to multiple antibiotics. This mechanism is prevalent among bacteria, as evidenced by the fact that other efflux pumps rank second in the distribution of mechanism classification, following ABC transporters (Figure 1 b). This can be conceded to the fact that efflux pumps can expel a wide variety of substances out of the bacterial cell, different classes of antibiotics included. Moreover, selective pressure may be applied, as only bacteria that carry the AMR genes may survive. Naturally, HGT also plays a role in this mechanism's prevalence.

A study of the distribution of classes of antibiotic drugs indicated that Glycylcycline is the most frequently developed drug class that bacteria have developed resistance to (Figure 1 c). Glycylcycline antibiotics block protein synthesis in bacterium, thus preventing its reproduction. Glycylcyclines are often used as last resort antibiotics for infections caused by multi-drug resistant bacteria[17]. It is possible that the heavy reliance on them leads to stronger selective pressure for resistance.

The UMAP projection visualizes the distinct clusters identified in the data (Figure 2). This suggests that the antibiotic resistance gene profiles for these environments are sufficiently different to be distinguishable from one another when reduced to two

dimensions. It indicates that each environment has a unique composition of resistance genes.

Model Results

Evaluation of the model illustrated the model's superior capability in accurately predicting the Human Microbiome and Soil environments (Figure 3, Figure 4). This heightened performance may be attributed to the substantial volume of samples from these categories, with the former possessing double the quantity of the latter, which itself significantly exceeds the third largest class. Moreover, it is likely that the human microbiome is studied more extensively than other metagenomic environments due to its critical role in medicine and human health, resulting in the generation of high-quality samples. Additionally, organisms in the human microbiome are exposed to high concentrations of antibiotics, which can lead to the emergence of more AMR as a result of natural selection. The abundance of data within these groups likely contributes to enhanced model precision and robustness. In the realm of machine learning, it is generally observed that expansive datasets correlate with improved model efficacy, a trend that appears to be reflected in this instance.

Conversely, the categories of Animal Microbiome and Sewage are characterized by a notably smaller sample size. This limitation in data quantity may account for their diminished performance within the model. Nevertheless, the ROC curves reveal a comparatively high mean ROC-AUC score, indicating the model's proficiency in correctly identifying true positive results. However, this is concurrently accompanied by a substantial rate of false positives. Such a phenomenon could stem from potential overfitting of the model to the limited samples available or its reliance on predicting these labels based on features that are not exclusively characteristic of these specific environments.

Testing the model on the test set after training on the train set revealed that the model demonstrates moderate discernment with a ROC-AUC score of 68% (Figure 5). While it can differentiate between animal microbiome samples and those of other environments, a notable degree of misclassification persists. The limited number of animal microbiome samples in the test set, amounting to just 28, possess a significant obstacle in enabling the model to deliver confident and accurate predictions for this class. Combining with the fact this class consists samples of animals from different phyla, which may affect the distribution of each sample individually, these two explanations contribute to the observed moderate ROC-AUC score.

Given the distribution of the test set, the model's performance on the animal microbiome class is particularly noteworthy. Despite a moderate ROC-AUC score of 68%, the small number of samples in the test set means that the ROC curve is based on a limited number of data points. This can lead to a less stable estimate of the model's performance for this class, as each misclassification has a significant impact on the AUC value.

For the sewage class, the ROC-AUC score of 88% suggests that the model is quite capable of identifying true positives, but given the moderate sample size (229 samples), it is clear that there are false positives affecting the performance. This may be due to overlapping features with other classes in the sewage samples, as a lot of microorganisms from the human microbiome end up in the sewage, or it could be related to the inherent diversity within the sewage environment itself.

The test performance points out to suboptimal outcomes on the PRC for the animal microbiome class (Figure 6). This suboptimal performance is likely attributable to the limited sample size in both the training and testing datasets. The scarcity of examples restricts the model's ability to generalize and accurately predict new samples from this class. Furthermore, the methodology of feature consolidation, where some features were merged into larger families while others remained isolated, may have contributed to feature sparsity. This inconsistency in feature representation could adversely affect the learning process of the model.

Although the sewage class possesses a reasonable sample volume in the testing set, the model's performance, as indicated by a PR-AUC score of 72%, is not on par with the human microbiome and soil categories. This may be reflective of the sewage environment's intrinsic complexity, potentially comprising diverse microbial communities derived from both human and animal sources, which could dilute the distinctiveness of AMR gene profiles and complicate the classification process.

Feature Importance

Ordering the top 20 features in terms of importance elucidates that four features markedly outweigh others in significance: Lactamase B, ANT9, RND Efflux, and Antibiotic NAT (Figure 7).

The enzyme β-lactamase, denoted here as Lactamase B, is synthesized by certain bacteria, endowing them with resistance to various β-lactam antibiotics such as penicillins and cephalosporins[18]. The pronounced relevance of this feature intimates that genes encoding for β-lactamases are potent indicators of a sample's environmental provenance, underscoring the presence of antibiotic resistance mechanisms in the environment.

ANT9 symbolizes a class of aminoglycoside nucleotidyltransferases, specific in modifying the 9-hydroxyl group of aminoglycoside antibiotics[19], thus conferring resistance. The prominence of ANT9 as a feature could be reflective of its predominance in environments where aminoglycosides are extensively utilized.

The term RND Efflux denotes the Resistance-Nodulation-Division efflux systems, integral in multi-drug resistance by extruding a variety of substances, including antibiotics, from the cell[20]. The notable importance of this feature in the analysis may indicate a widespread presence of RND efflux genes in the sampled environments, hinting at an environmental imprint left by diverse antibiotic exposures.

Lastly, NAT stands for Aminoglycoside 3-N-acetyltransferase[21]. Aminoglycoside 3-N-acetyltransferase is an enzyme that confers bacterial resistance to aminoglycoside antibiotics by acetylating the 3-amino group of the aminoglycoside structure, thereby inhibiting the antibiotic's ability to bind effectively to its bacterial target[22].

Overall, there appears to be a positive correlation between the feature importance results and the data distribution, with more frequently occurring mechanisms or gene families also being more significant in determining the environment.

This research, through the use of a Random Forest machine learning model, sheds light on the mechanisms of antibiotic resistance spread in various environments. By analyzing metagenomic samples, it identifies key features (like Lactamase B, ANT9, RND Efflux, and Antibiotic NAT) that are indicators of environmental origins. This is

crucial as it helps in understanding how different environments contribute to the development and dissemination of antibiotic resistance. The risks associated with antibiotic resistance include the potential for increased prevalence of drug-resistant infections and the challenges they pose to public health. The findings highlight the need for targeted monitoring and intervention strategies in various environments to manage and mitigate the spread of antimicrobial resistance. This research contributes significantly to the field of AMR surveillance, emphasizing the importance of environmental factors in the evolution and spread of antibiotic resistance.

Future Work

The predictive model's performance points to important directions for further investigation.

- Remove duplicands: Remove samples that are similar to each other in terms of the AMR genes distribution, thus resulting in similar feature vectors.
- Isolate factors: train the model on each of the factors: Gene IDs, gene families, mechanisms, and drug classes, individually. This way it is possible to assess the impact of each on the model's predictive accuracy.
- Combine similar features: Investigate uniting features describing similar AMR genes to create more general features. Some features that could be merged weren’t due to limit of biological knowledge. Combining the features in a more efficient way could reduce the sparsity of the data, enhance the model's generalizability, and mitigate overfitting issues.
- Data augmentation: Enriching the dataset with a greater number of samples is essential, with particular emphasis on underrepresented environments like sewage and animal microbiome. This expansion will aid in bolstering the model’s robustness and predictive precision.
- Compare the work done here to similar projects held in different places in order to see strengths and weaknesses of each strategy.

References:

1. Crofts, T. S., Gasparrini, A. J. & Dantas, G. Next-generation approaches to understand and combat the antibiotic resistome. *Nat Rev Microbiol* **15**, 422–434 (2017).
2. Boeckel, T. P. V. *et al.* Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. *The Lancet Infectious Diseases* **14**, 742–750 (2014).
3. O'Neill, J. *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations*. https://apo.org.au/node/63983 (2016).
4. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol* **13**, 42–51 (2015).
5. Fernández-Arrojo, L., Guazzaroni, M.-E., López-Cortés, N., Beloqui, A. & Ferrer, M. Metagenomic era for biocatalyst identification. *Current Opinion in Biotechnology* **21**, 725–733 (2010).
6. Garza, D. R., van Verk, M. C., Huynen, M. A. & Dutilh, B. E. Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat Microbiol* **3**, 456–460 (2018).
7. Sharma, A. Random Forest vs Decision Tree | Which Is Right for You? *Analytics Vidhya* https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ (2020).
8. Whole Genome Shotgun Submissions. https://www.ncbi.nlm.nih.gov/genbank/wgs/.
9. Gurbich, T. A. *et al.* MGnify Genomes: A Resource for Biome-specific Microbial Genome Catalogues. *Journal of Molecular Biology* **435**, 168016 (2023).
10. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* **9**, 207–216 (2015).
11. The Comprehensive Antibiotic Resistance Database. https://card.mcmaster.ca/.
12. scikit-learn: machine learning in Python — scikit-learn 1.4.0 documentation. https://scikit-learn.org/stable/.
13. de Abreu, V. A. C., Perdigão, J. & Almeida, S. Metagenomic Approaches to Analyze Antimicrobial Resistance: An Overview. *Front Genet* **11**, 575592 (2021).
14. Chloramphenicol Phosphotransferase (The Comprehensive Antibiotic Resistance Database). https://card.mcmaster.ca/ontology/36388.
15. Izard, T. & Ellis, J. The crystal structures of chloramphenicol phosphotransferase reveal a novel inactivation mechanism. *EMBO J* **19**, 2690–2700 (2000).
16. Akhtar, A. A. & Turner, D. PJ. The role of bacterial ATP-binding cassette (ABC) transporters in pathogenesis and virulence: Therapeutic and vaccine potential. *Microbial Pathogenesis* **171**, 105734 (2022).
17. Yaghoubi, S. *et al.* Tigecycline antibacterial activity, clinical effectiveness, and mechanisms and epidemiology of resistance: narrative review. *Eur J Clin Microbiol Infect Dis* **41**, 1003–1022 (2022).
18. Eiamphungporn, W., Schaduangrat, N., Malik, A. A. & Nantasenamat, C. Tackling the Antibiotic Resistance Caused by Class A β-Lactamases through the Use of β-Lactamase Inhibitory Protein. *Int J Mol Sci* **19**, 2222 (2018).
19. ANT(9) (The Comprehensive Antibiotic Resistance Database). https://card.mcmaster.ca/ontology/36367.

20. Resistance Nodulation cell Division (RND) Antibiotic Efflux Pump (The Comprehensive Antibiotic Resistance Database). https://card.mcmaster.ca/ontology/36005.
21. AAC(3) (The Comprehensive Antibiotic Resistance Database). https://card.mcmaster.ca/ontology/36461.
22. Magalhaes, M. L. B. & Blanchard, J. S. The Kinetic Mechanism of AAC(3)-IV Aminoglycoside Acetyltransferase from Escherichia coli. *Biochemistry* **44**, 16275–16283 (2005).

Appendix

The code and datasets used in this project can be found in the following directories:

- WGS dataset:
  /davidb/bio_db/NCBI/WGS/Metagenomes/
- Mgnify dataset:
  /davidb/assemblies/preassembled/Mgnify/MgyAssemblies/contigs.min2500/
- Denovo dataset:
  /davidb/assemblies/denovo
- The HMM file of Resfams' AMR families:
  /davidb/nirborger/Resfams-full.hmm
- Metadata for the Resfams HMM profiles:
  /davidb/nirborger/180102_resfams_metadata_updated_v1.2.2_with_CARD_v2.xlsx
- MergeByAccession.json file path:
  /davidb/nirborger/MergesByAccession.json
  This is a dictionary containing instructions how to merge the Query_Accession critertion
- MergeByID,json file path:
  /davidb/nirborger/MergesByID.json
  This is a dictionary containing instructions how to merge the Query_ID critertion
- Full code, including all necessary files:
  https://bitbucket.org/BursteinLab/env_prediction/src/master/