A deep learning approach for predicting antimicrobial resistance across various bacterial species using whole genome sequence

by
Prateeksha Muddemmanavar

Under the supervision of
Dr. Debarka Sengupta

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, in Computational Biology

Department of Computational Biology Indraprastha Institute of Information Technology - Delhi
May, 2024

Certificate

This is to certify that the thesis titled *"A deep learning approach for predicting antimicrobial resistance across various bacterial species using whole genome sequence"* being submitted by **Prateeksha Muddemmanavar** to the Indraprastha Institute of Information Technology, Delhi, for the award of the Master of Technology in Department of Computational Biology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May,2024

Dr Debarka Sengupta

Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Acknowledgements

I extend my heartfelt appreciation to Dr. Debarka Sengupta for his invaluable guidance and support during my M.Tech thesis. His expertise, encouragement, and constructive feedback have been crucial in shaping my work, inspiring me to strive for excellence and overcome challenges. Dr. Sengupta's commitment for creating a supportive research atmosphere and his confidence in my capabilities have been incredibly inspiring.. It has been a privilege to work under his guidance, and I am grateful for his significant time and contributions to my academic journey.

I am also grateful to Abhishek Halder, a PhD student under Dr. Debarka Sengupta, whose unwavering support and invaluable assistance played a pivotal role in completing my thesis. His expertise, guidance, and encouragement were essential throughout this process, and I am truly grateful for his input. Abhishek was a steady source of motivation for my structured research efforts during my M.Tech thesis journey, and I gained valuable insights from him that will be invaluable assets for my future endeavors.

Furthermore, I extend my thanks to my friends for their continuous encouragement and support. Additionally, I acknowledge the contributions of the broader research community, whose insights shared through publications, conferences, and online platforms have been indispensable in shaping my research.

Lastly, I am profoundly thankful to my family for their unwavering love and support, which has been a constant source of strength throughout.

Abstract

Antimicrobial resistance (AMR) presents an immediate threat to public health as microorganisms evolve to resist antimicrobial drugs, leading to challenging or untreatable infections. AMR-related costs could exceed 1 trillion dollars globally by 2050, surpassing major causes of death. In 2019, AMR directly caused 1.27 million deaths worldwide, surpassing mortality rates of HIV/AIDS and malaria. In particular, tuberculosis (TB) claimed 1.3 million lives in 2022, ranking as the second most prevalent infectious disease globally. This study introduces in silico approach utilizing deep learning to analyze the entire genome sequence of top pathogens such as *Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumonia*, *Streptococcus pneumoniae*, *Pseudomonas aeruginosa*, and *Mycobacterium tuberculosis*. The goal is to swiftly and accurately predict a pathogen's resistance to specific drugs, eliminating the necessity for complex laboratory experiments. In addition to employing the standard label encoding technique, this study introduces three novel encoding methods for mutational data: transition-transversion encoding, codon frequency encoding, and gene based codon gain. These novel techniques are the major contribution of this study. Prediction of the AMR profile of the patient's pathogen against various drugs is crucial before prescribing treatment. This approach helps in identifying the most appropriate drug that will be effective in treating the patient at the early stage, thereby reducing the likelihood of treatment failure caused by prescribing a drug to which the pathogen is resistant. While the model's performance varies across different drugs and pathogens, it demonstrates the potential for application in antimicrobial resistance prediction.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

Antimicrobial resistance (AMR) denotes microorganisms' ability, including bacteria, viruses, fungi, and parasites, to withstand the effects of antimicrobial drugs [19, 2]. This resistance poses a significant threat to public health, leading to infections that are challenging or impossible to treat, resulting in increased morbidity, mortality, and healthcare costs [14, 15, 16]. If left unaddressed, the AMR could lead to an expense of US$1 trillion in healthcare expenses by 2050, and annual losses of US$1 trillion to US$3.4 trillion in gross domestic product (GDP) by 2030, according to a report published in 2023 by the World Health Organization (WHO) [23, 21]. In 2019, approximately 1.27 million deaths were directly caused by AMR across the globe[21, 6]. In particular, drug-resistant infections claimed more lives than HIV/AIDS (864,000 deaths) or malaria (643,000 deaths). A 2016 review [10] on AMR estimates that by 2050, up to ten million people could die annually from causes related to AMR. The analysis identified the chest, bloodstream, and abdomen as the three primary sites of bacterial AMR infections, accounting for 78.8% of deaths directly related to AMR. The six most lethal bacterial pathogens contributed to nearly three-quarters of all resistance-related deaths. For example, antibiotic-resistant Escherichia coli alone claimed the lives of approximately 200,000 people in 2019 [22]. The study identified *Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Streptococcus pneumoniae*, and *Pseudomonas aeruginosa* as few of the top primary pathogens accountable for deaths associated with antimicrobial resistance [20]. A recent report from the World Health Organization (WHO) highlights that globally Tuberculosis (TB) claimed the lives of 1.3 million individuals in 2022, with an additional 167,000 deaths among individuals co-infected with HIV. TB stands as the second most prevalent infectious disease following COVID-19, surpassing HIV and AIDS. There is an inadequate research in the face of rising levels of resistance and an urgent need for additional measures to ensure equitable access to new and existing vaccines, diagnostics, and medicines [23]. Traditional antimicrobial susceptibility testing (AST) is a method that's slow

and not very efficient, mainly suitable for bacteria that can be grown in a lab [13]. While it's still commonly used for analyzing antimicrobial resistance (AMR) in medical settings, it requires specialized facilities and skilled personnel [3]. Our study aimed to comprehensively anticipate the antimicrobial resistance patterns of leading pathogens against various standard antibiotics. Furthermore, our analysis included TB bacteria, which is the second most contagious disease. Notably, COVID-19 was not factored into the detection of antimicrobial resistance profiles due to the unavailability of corresponding data.

Bacteria develop AMR through mechanisms such as drug misuse [25, 9], modification of drug targets, and activation of efflux pumps, which allows survival in the presence of antimicrobial agents [5]. The prevalence of bacteria resistant to antibiotics is increasing due to widespread antibiotic use. If left unaddressed, infections, once easily treatable with antibiotics, may become untreatable [20].

Chapter 2

Motivation

The urgent global challenge of AMR poses a significant threat to public health, risking human lives worldwide. Beyond its immediate health impacts, AMR also carries substantial economic consequences, affecting individuals and nations. Innovative strategies are needed to confront the rising resistance of pathogens, including faster AMR profile diagnostics and treatments and advancements in deep learning methodologies for more effective interventions. However, the prevalent deep learning approach predominantly relies on label encoding, which lacks semantic meaning. Recognizing the limitation, this study focuses on developing a novel encoding technique for problems of Single Nucleotide Polymorphism (SNP) data, converting each feature with meaningful numerical representations. Through the proposed multiple encoding techniques derived from biological insights, our findings suggest that encoding with biological significance holds promise for achieving superior results compared to traditional label encoding methods.

Chapter 3

Related Work

Yang et al. [24] used different machine learning techniques, like support vector machine (SVM), logistic regression (LR), and random forest (RF), trained on genetic data, to predict AMR accurately. Also, deep learning methods, as shown by studies like Arango et al. [1] and Stokes et al. [17], have shown promise in predicting new antibiotics, AMR genes, and AMR peptides. In a recent study [13], various machine learning algorithms, including logistic regression, support vector machine, random forest, and convolutional neural network (CNN), were tested to predict antimicrobial resistance using genetic data. They found that random forests and CNNs outperformed logistic regression and support vector machine, achieving impressive scores of up to 0.96 for predicting antimicrobial resistance [13]. Another recent study [8] developed 24 classifiers to predict drug resistance in Mycobacterium tuberculosis (MTB) for eight medications. These classifiers used logistic regression, random forest, and 1D CNN algorithms and achieved better results than existing methods for several medications. Green et al. [7] proposed two deep CNNs for predicting antibiotic resistance in MTB: a multi-drug CNN (MD-CNN) and 13 single-drug CNNs (SD-CNNs). The MD-CNN predicted resistance to 13 antibiotics with high accuracy, and each SD-CNN focused on a specific antibiotic, achieving excellent performance. In a study [4], a wide-and-deep neural network showed superior performance in predicting antimicrobial resistance to 10 antibiotics using M. tuberculosis genetic data compared to previous methods.

Chapter 4

Dataset

The genome sequences of the six pathogens mentioned previously, meeting the criteria of "Good" Genome Quality, "WGS" (whole genome sequence) or "Complete" in Genome Status, and associated with a human host group within the PATRIC (Pathosystems Resource Integration Center) database, were selected for analysis. The dataset consists of FASTA files containing genome sequences in contig format. Genomes displaying laboratory-validated AMR phenotypes labeled as either "susceptible" or "resistant" were then considered for further examination. This resulted in a total of 3202 *Escherichia coli*, 3500 *Staphylococcus aureus*, 3176 *Klebsiella pneumonia*, 1600 *Streptococcus pneumoniae*, 1320 *Pseudomonas aeruginosa*, and 12667 *Mycobacterium tuberculosis* samples being included. We considered only the top few first and second-line drugs with a sample size exceeding 250 in the analysis. Figure 1 illustrates the distribution of phenotypes for each of these drugs across the various pathogens. In summary, a total of 48 combinations of pathogen and drug pair resistance profiles were analyzed in this study.

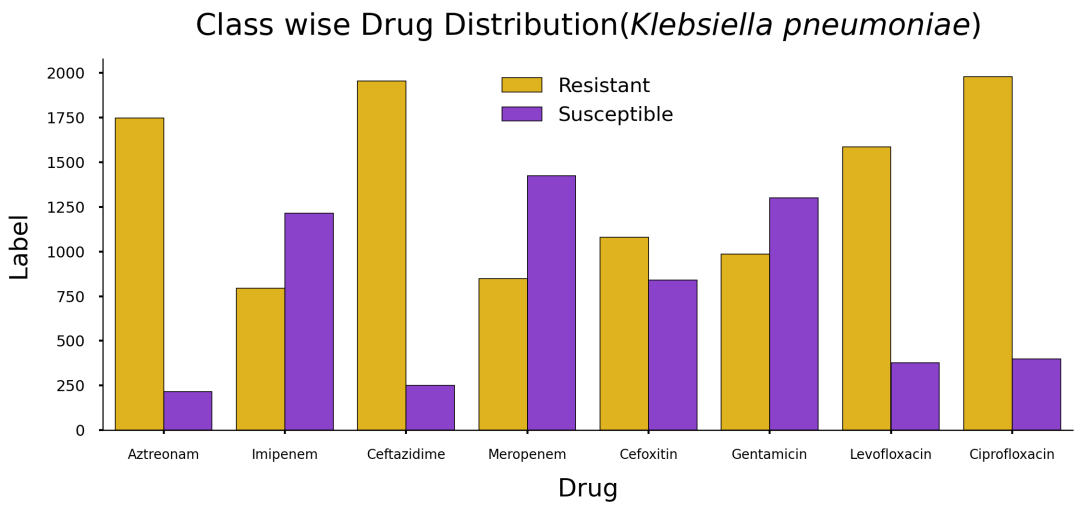

Figure 4.1: Class wise drug distribution of *Klebsiella pneumonia*

The figure 4.1 illustrates the distribution of samples by class for *Klebsiella pneumoniae*. Ciprofloxacin has the most samples, totaling 2380, whereas Cefoxitin has the lowest, with 1920 samples.

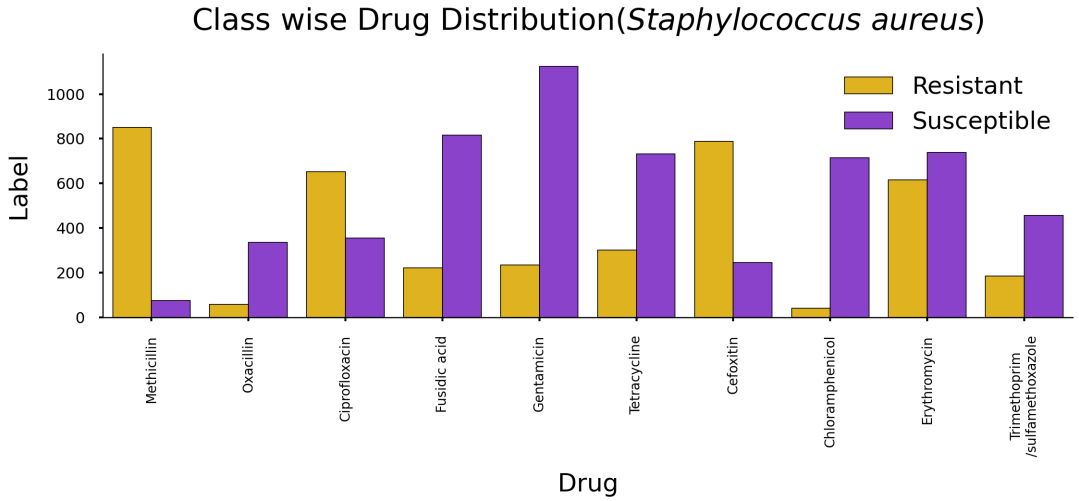

Figure 4.2: Class wise drug distribution of *Staphylococcus aureus*

The figure 4.2 shows the distribution of *Staphylococcus aureus* samples by class. Gentamicin has the highest number of samples at 1,358, while Oxacillin has the fewest with 394 samples. The imbalance ratio between resistant and susceptible samples ranges from 0.05 to 11.5.

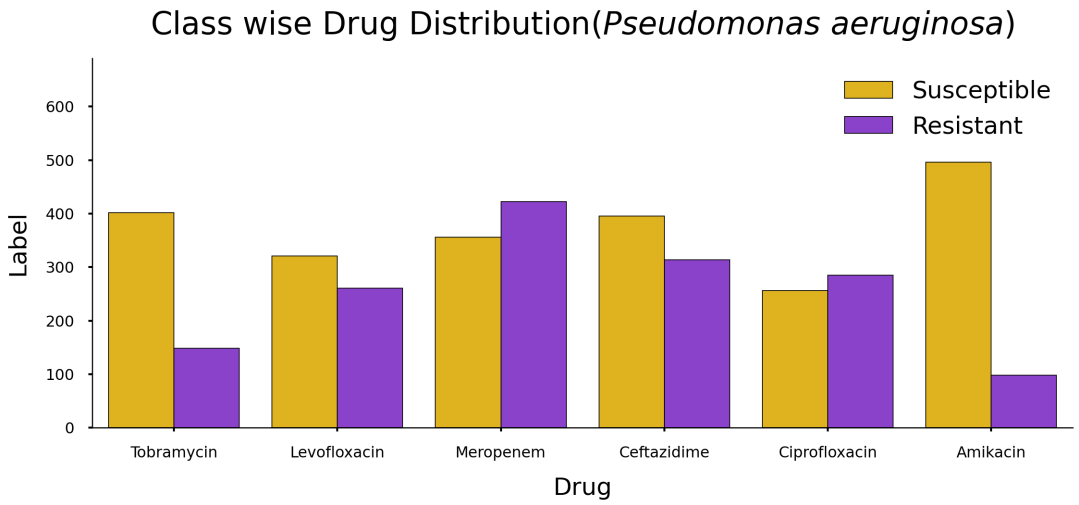

Figure 4.3: Class wise drug distribution of *Staphylococcus aureus*

The figure 4.3 depicts the distribution of *Pseudomonas aeruginosa* samples by class. Meropenem has the most samples, totaling 778, whereas Ciprofloxacin has the fewest, with 541 samples. The imbalance ratio between resistant and susceptible samples varies from 0.19 to 1.18.

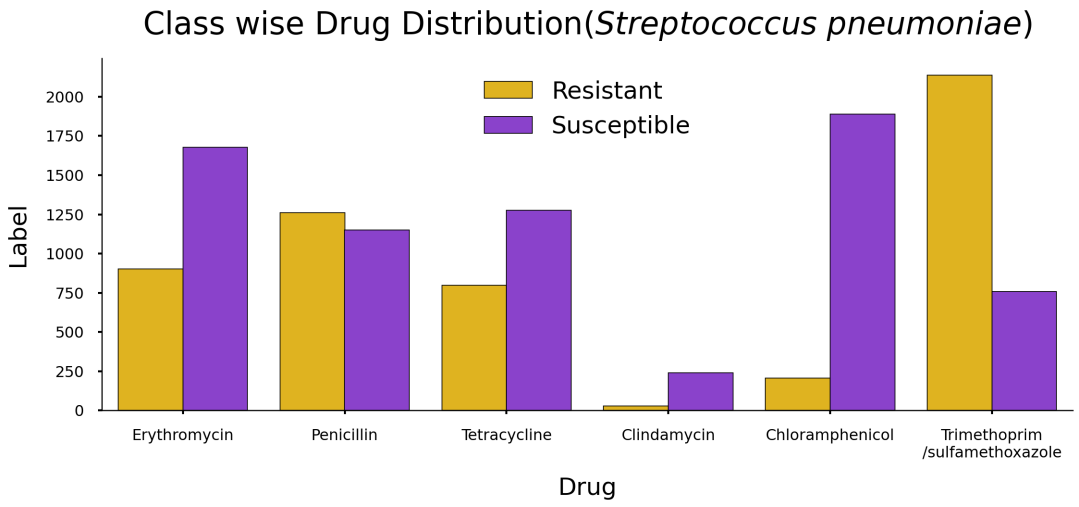

Figure 4.4: Class wise drug distribution of *Streptococcus pneumoniae*

The figure 4.4 shows the distribution of *Streptococcus pneumoniae* samples by class. Trimethoprim/Sulfamethoxazole has the highest number of samples at 2896, while Clindamycin has the lowest with 268 samples. The imbalance ratio between resistant and susceptible samples ranges from 0.11 to 2.82.

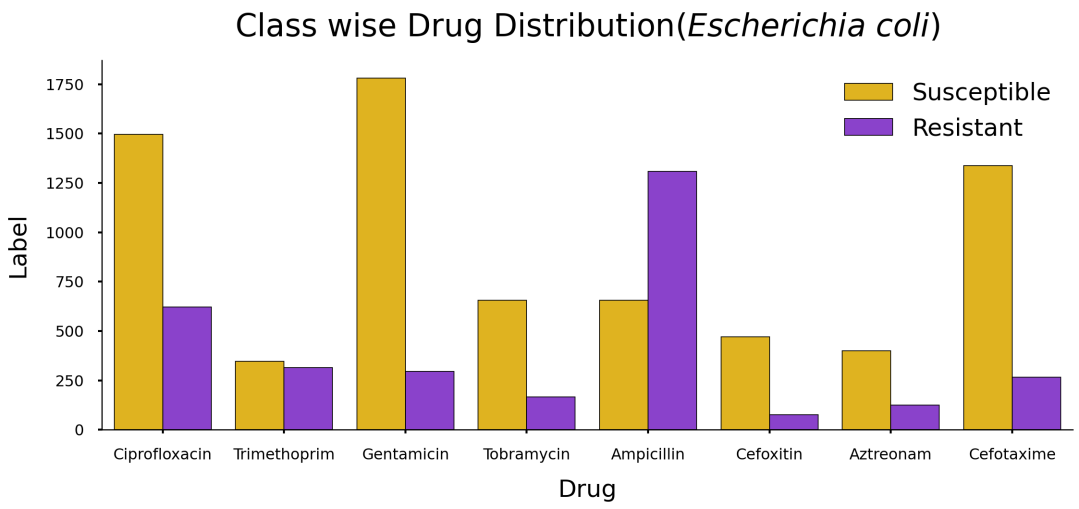

Figure 4.5: Class wise drug distribution of *Escherichia coli*

The figure 4.5 illustrates the distribution of *Escherichia coli* samples by class. Ciprofloxacin has the most samples, numbering 2,199, while Aztreonam has the fewest, with 524 samples. The imbalance ratio between resistant and susceptible samples ranges from 0.16 to 1.99.

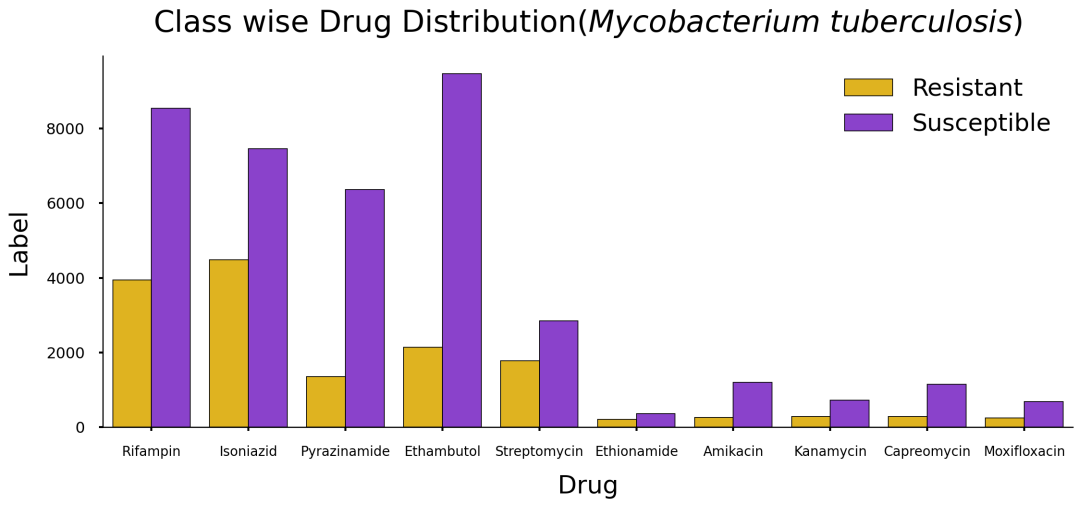

Figure 4.6: Class wise drug distribution of *Mycobacterium tuberculosis*

The figure 4.6 shows the distribution of *Mycobacterium tuberculosis* samples by class. Rifampin has the highest number of samples at 12484, while Ethionamide has the fewest with 567 samples. The imbalance ratio between resistant and susceptible samples ranges from 0.21 to 0.62.

Chapter 5

Methodology

5.1 Variant calling

Snippy version 4.6.0 was employed for variant calling using its robust default settings, which included a stringent minimum coverage depth of 10, a high minimum Variant Call Format (VCF) quality threshold of 100, and an automatic minimum fraction threshold. 5.1 describes the detailed Variant Calling pipeline.

The reference genome used for the Streptococcus pneumonia, Klebsiella pneumonia, Mycobacterium tuberculosis, Staphylococcus aureus, Escherichia coli, Pseudomonas aeruginosa are *Streptococcus pneumoniae* Hu17, *Klebsiella pneumoniae subsp. pneumoniae* HS11286, *Mycobacterium tuberculosis* H37Rv, *Staphylococcus aureus* subsp. aureus NCTC 8325, *Escherichia coli* O157:H7 str. Sakai, *Pseudomonas aeruginosa* PAO1 respectively.

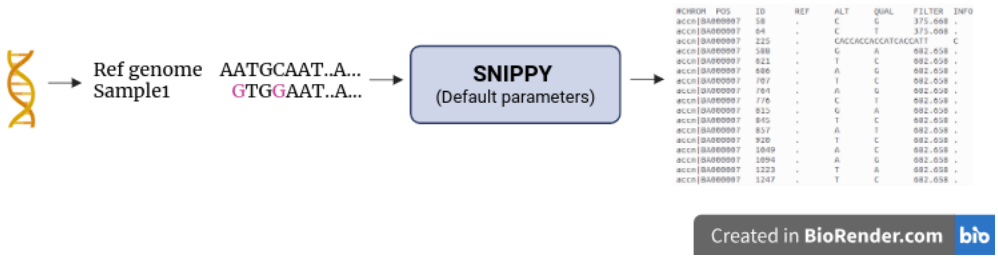

Figure 5.1: Pipeline for Generating VCF files via alignment to a pathogen's reference genome

5.2 Data encoding

It is essential in machine learning to convert non-numerical input into numerical inputs, as most algorithms operate on numerical data only. Converting categorical variables into numerical format enables algorithms to process and analyze the data effectively. In the context of SNP data analysis, encoding nucleotide bases into numerical values is vital for computational analysis, as it allows algorithms to identify genetic patterns and variations across the genome.

5.2.1 Label encoding

Label encoding is a technique used to convert categorical variables into numerical form by assigning a unique integer to each category. In SNP data analysis, the variant nucleotide bases adenine (A), guanine (G), cytosine (C), and thymine (T) are assigned integers 1, 2, 3, and 4, respectively, through the process of label encoding. If no mutation occurs at a particular position, it is represented by the integer 0. This approach represents the complete SNP mutation dataset using five integers, ranging from one to five, facilitating computational analysis while preserving the categorical structure of the data. Nevertheless, it's crucial to acknowledge that label encoding might unintentionally suggest ordinal relationships between categorical variables, which may not be appropriate. It is the main bottleneck of the label encoding method.

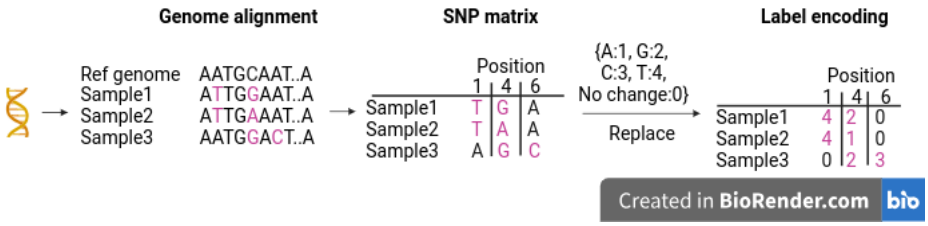

Figure 5.2: Block diagram of label encoding

5.2.2 Transition-transversion encoding

We introduce a novel encoding method called transition-transversion encoding, which provides a biologically informed approach to representing nucleotide variations based on their chemical structure. Unlike traditional label encoding methods that mainly focus on identifying variant nucleotides at mutation sites without considering the reference nucleotide alleles, this approach takes into account both the variant nucleotide allele (SNPs) and the specific

nature of the nucleotide change in relation to the reference nucleotide at a particular genomic position. Recent studies in antimicrobial resistance and microbial evolutionary dynamics [18] underscore the importance of transitions [11, 12] compared to transversions. Transition mutations involve the substitution of a purine base with another purine, or a pyrimidine with another pyrimidine, whereas transversion mutations entail the substitution of a purine with a pyrimidine or vice versa. Purines (A, G) consist of a double-ring structure, while pyrimidines (C, T) possess a single six-membered ring. To highlight the significance of transitions over transversions, this encoding method assigns higher integer values to transition groups. Transversions from purine to pyrimidine bases (A→T, A→C, G→C, G→T) are labeled as 1, while those from pyrimidine to purine bases (T→A, T→G, C→A, C→G) are labeled as 2. Transitions within purine bases (A↔G) are denoted as 3, and transitions within pyrimidine bases (T↔C) are denoted as 4. Unchanged bases are assigned a value of 0. This encoding methodology is grounded in the understanding that transitions play a more substantial role in antimicrobial resistance mechanisms and microbial evolution [11].

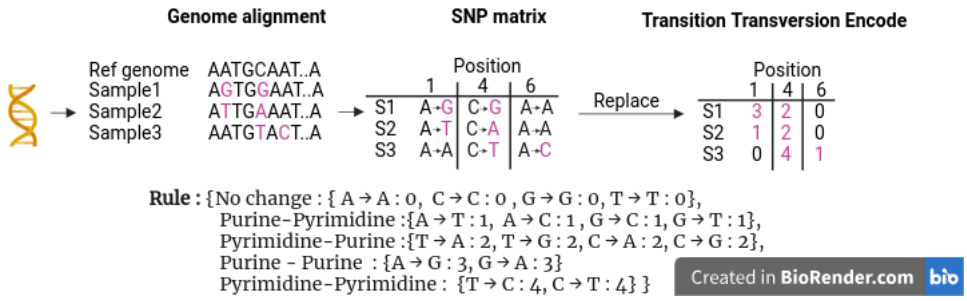

Figure 5.3: Block diagram of transition-transversion encoding

5.2.3 Codon frequency encoding

Codon bias, the non-random usage of synonymous codons in genomes, arises from mutation biases, selective pressures favoring certain codons for efficient translation, and variations in GC content across genomic regions. Encoding mutational data using a codon bias approach preserves biological context by considering the impact of mutations on specific codons, the fundamental units of genetic translation, thereby enabling a detailed analysis of mutational effects and capturing subtle variations in gene expression and protein function. Codon bias significantly influences codon frequency by determining which codons are used more frequently, impacting codon frequency-based encoding. Algorithm 1 proposes converting SNP data into numerical form using codon frequency, accounting for selection pressure among codons and reflecting the evolutionary and functional constraints that shape genetic variation

and protein expression.

Algorithm 1 Codon frequency encoding algorithm

Identify all relative positions in the gene where non synonymous mutations have occurred in at least one of the studied strains.
Generate a dataframe (df) to represent codon bias, with strains as rows and relative mutation positions obtained in step 1 as columns, labeled as gene_position. **for** $i = 1, 2, \ldots, number\_of\_strains$ **do**
    for $j = 1, 2, \ldots, number\_of\_unique\_amino\_acid\_mutation\_positions$ **do**
        mut_pos = extract position from j.
        mut_gene = extract gene name from j.
        df.loc[i, j] = frequency of the altered codon present at the mut_pos in the entire genome of the i^{th} strain.
df contains the output of codon frequency encoding from SNP data.

5.2.4 Gene based codon gain

Gene-based codon gain, derived from codon bias, extends the concept of codon frequency encoding by incorporating the intricate variations in codon usage among genes, influenced by factors such as gene function, expression level, and evolutionary history. Analyzing gene-based codon frequency offers a comprehensive understanding of codon bias within individual genes, capturing specific variations that directly impact protein expression and function, thus underscoring the importance of gene-level analysis. In this methodology, each genetic mutation is replaced by the log-likelihood ratio of the altered codon's frequency in the strain compared to that of the reference codon frequency within the reference gene. Algorithm 2 outlines the process for gene-based codon gain, offering a sophisticated framework for this analysis.

Algorithm 2 Gene based codon gain encoding algorithm

Identify all relative positions in the gene where non synonymous mutations have occurred in at least one of the studied strains.

Generate a dataframe (df) to represent codon bias, with strains as rows and relative mutation positions obtained in step 1 as columns, labeled as gene_position. **for** $i = 1, 2, \ldots, number\_of\_strains$ **do**

  for $j = 1, 2, \ldots, number\_of\_unique\_amino\_acid\_mutation\_positions$ **do**

    mut_pos = extract position from j.

    mut_gene = extract gene name from j.

    df.loc[i, j] = log(frequency of the altered codon at the mut_pos inside the mut_gene for the i^{th} strain/ frequency of the reference codon at the mut_pos inside the mut_gene for the reference genome).

5.3 Model architecture

The model architecture outlined in the study consists of a fully connected deep neural network (DNN) with optional dropout layers and activation functions. Specifically, the architecture involves initializing a DNN with 2 to 3 layers, each comprising nodes ranging from 32 to 512. Dropout layers may be optionally added with dropout rates between 0.1 and 0.5 to mitigate overfitting. The ReLU activation function is applied in each node, except for the output layer, where the sigmoid activation function is employed. Additionally, L1/L2 regularization techniques may be optionally applied to prevent overfitting and encourage generalization. The complete algorithm of the present study is given in algorithm 3.

Algorithm 3 Algorithm for AMR Phenotype Classification from numeric SNP data

1: **Input:** Numeric SNP data containing variant alleles at mutation positions for a given pathogen strains.
2: **Output:** Predicted AMR phenotype of pathogens.
3: **Step 1:** Split the data into training (64%), validation (16%), and test sets (20%).
4: **Step 2:** Encode the data using one of the specified encoding strategies.
5: **Step 3:** Initialize a fully connected deep neural network (DNN) with 2 to 4 layers, each having nodes between 32 to 512.
6: **Step 4:** Optionally add dropout layers with dropout rates ranging between 0.1 and 0.5 to prevent overfitting.
7: **Step 5:** Utilize ReLU activation function in each node except the output layer to introduce non-linearity. Employ the sigmoid activation function in the output layer.
8: **Step 6:** Optionally apply L1/L2 regularization techniques as necessary to avoid overfitting and encourage generalization.
9: **Step 7:** Use binary cross-entropy as the loss function to penalize model predictions that diverge from actual labels.
10: **Step 8:** Train the model on the training data specific to the organism and drug.
11: **Step 9:** Fine-tune model parameters based on validation data to minimize loss and achieve optimal performance.
12: **Step 10:** Evaluate the model's performance on the test data.
13: **Step 11:** Repeat steps 1-10 for each organism and drug pair independently.

Figure 5.4 shows the complete end-to-end block diagram of AMR phenotype classification using different encoding techniques.

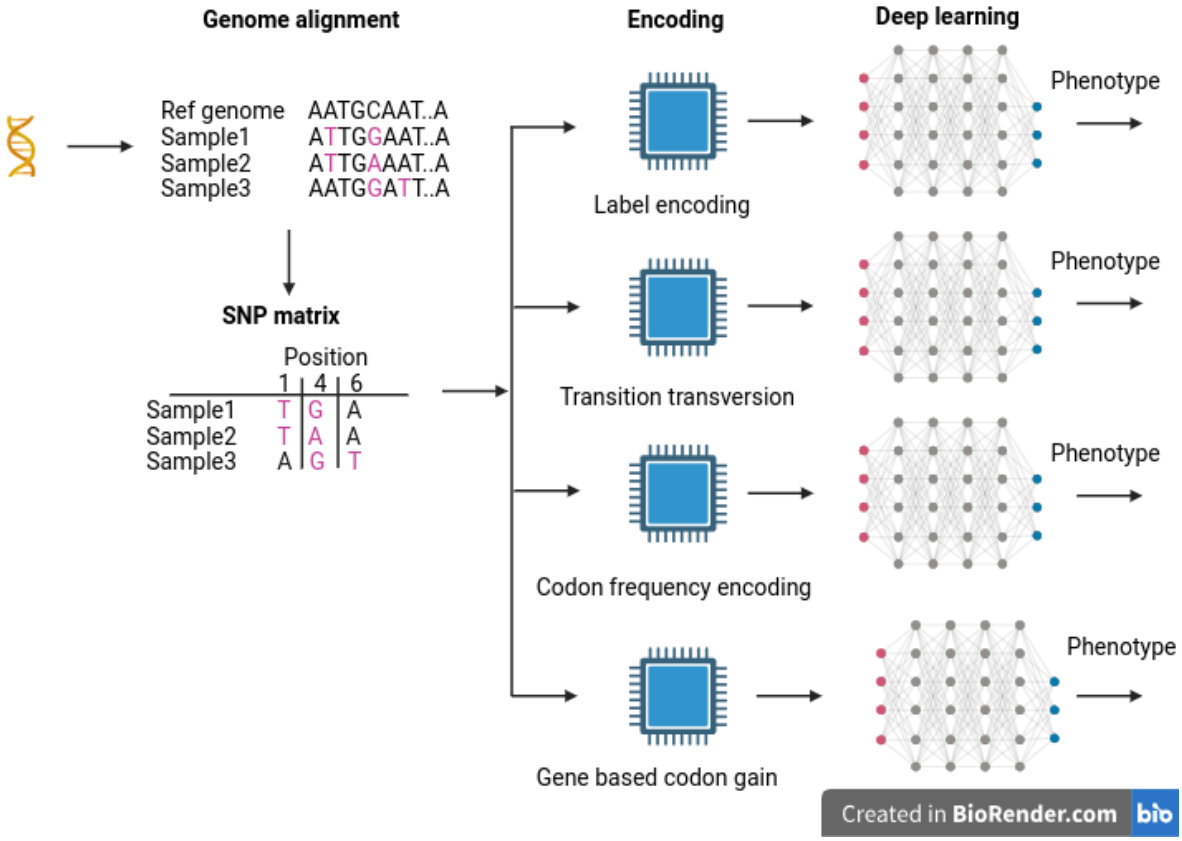

Figure 5.4: Complete block diagram of AMR phenotype classification using whole genome sequence

Chapter 6

Data Splitting Strategy and Metric Selection

6.1 Data splitting

The complete dataset was split into three parts: 64% for training, 16% for validation, and the remaining 20% for testing. During the training phase, the model learns various patterns from the data. The validation set helps in fine-tuning the model's hyperparameters, simultaneously minimizing loss and preventing overfitting by continually monitoring its loss. The test set is crucial as it comprises completely unseen data, offering a reliable assessment of the model's generalization abilities. A model was considered trained when both the training and validation losses converged to lower values without overfitting. The final outcome was reported based on the unseen test data to ensure generalization.

6.2 Evaluation metric

Accuracy assesses the ratio of correctly classified instances to the total instances, though it's not ideal for imbalanced datasets. Precision measures the proportion of true positive predictions among all positive predictions generated by the model, while recall measures the ratio of true positive predictions to all actual positive instances. The F1 score is the harmonic mean of precision and recall. It provides a single score that balances both precision and recall. ROC AUC quantifies classifier performance by computing the area under the ROC curve, illustrating the trade-off between true positive and false positive rates at varying thresholds. A higher ROC AUC score signifies superior classifier performance. Matthews Correlation Coefficient (MCC) serves as a robust measure for binary classifications, factoring

in both true and false positives and negatives, making it a suitable metric even in the presence of class imbalance. Specificity measures the proportion of true negatives out of all actual negative instances. In this study, we have presented all the aforementioned metrics. The following equation represents these metrics:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}} \tag{6.1}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{6.2}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6.3}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.4}$$

$$\text{ROC AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))dt \tag{6.5}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{6.6}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{6.7}$$

Chapter 7

Results

Table 7.1: Drug vs metric for *Mycobacterium tuberculosis*

Antibiotic	Method	Accuracy	Precision	Recall	F1 Score	Kappa	MCC	ROC	Specificity
Rifampin	Label Encode	0.962	0.95	0.93	0.94	0.913	0.913	0.981	0.977
	Transition-transversion	0.958	0.939	0.929	0.934	0.904	0.904	0.98	0.972
	Codon frequency encode	0.949	0.928	0.911	0.92	0.883	0.883	0.967	0.967
	Gene based codon gain	0.945	0.945	0.877	0.909	0.87	0.871	0.973	0.976
Isoniazid	Label Encode	0.948	0.92	0.945	0.932	0.891	0.891	0.982	0.951
	Transition-transversion	0.95	0.927	0.941	0.934	0.894	0.895	0.981	0.955
	Codon frequency encode	0.945	0.965	0.886	0.924	0.881	0.883	0.967	0.981
	Gene based codon gain	0.943	0.97	0.877	0.921	0.878	0.88	0.968	0.983
Ethambutol	Label Encode	0.915	0.73	0.86	0.79	0.737	0.741	0.952	0.928
	Transition-transversion	0.915	0.73	0.86	0.79	0.737	0.741	0.952	0.928
	Codon frequency encode	0.884	0.632	0.895	0.741	0.669	0.686	0.938	0.882
	Gene based codon gain	0.895	0.667	0.867	0.754	0.689	0.699	0.945	0.902
Pyrazinamide	Label Encode	0.882	0.622	0.833	0.712	0.64	0.651	0.934	0.892
	Transition-transversion	0.891	0.656	0.804	0.723	0.656	0.661	0.922	0.91
	Codon frequency encode	0.895	0.701	0.701	0.701	0.637	0.637	0.906	0.936
	Gene based codon gain	0.878	0.622	0.789	0.695	0.621	0.628	0.916	0.897
Streptomycin	Label Encode	0.904	0.883	0.868	0.875	0.798	0.798	0.948	0.927
	Transition-transversion	0.903	0.864	0.89	0.877	0.798	0.798	0.951	0.912
	Codon frequency encode	0.881	0.857	0.829	0.843	0.747	0.747	0.916	0.913
	Gene based codon gain	0.886	0.846	0.862	0.854	0.761	0.761	0.927	0.901
Amikacin	Label Encode	0.921	0.788	0.773	0.78	0.732	0.732	0.917	0.953
	Transition-transversion	0.921	0.777	0.792	0.785	0.736	0.736	0.912	0.949
	Codon frequency encode	0.88	0.689	0.714	0.701	0.627	0.627	0.883	0.921
	Gene based codon gain	0.886	0.656	0.734	0.693	0.623	0.625	0.912	0.918
Kanamycin	Label Encode	0.862	0.734	0.81	0.77	0.672	0.674	0.92	0.883
	Transition-transversion	0.901	0.827	0.827	0.827	0.759	0.759	0.94	0.931
	Codon frequency encode	0.882	0.826	0.741	0.781	0.701	0.703	0.914	0.938
	Gene based codon gain	0.857	0.773	0.706	0.738	0.641	0.642	0.89	0.917
Moxifloxacin	Label Encode	0.802	0.607	0.693	0.647	0.51	0.513	0.798	0.84
	Transition-transversion	0.828	0.742	0.53	0.61	0.512	0.524	0.784	0.934
	Codon frequency encode	0.839	0.771	0.551	0.642	0.543	0.555	0.775	0.942
	Gene based codon gain	0.873	0.648	0.66	0.654	0.576	0.576	0.871	0.92
Ethionamide	Label Encode	0.789	0.725	0.69	0.707	0.543	0.543	0.855	0.847
	Transition-transversion	0.807	0.75	0.714	0.731	0.581	0.581	0.833	0.861
	Codon frequency encode	0.771	0.7	0.666	0.682	0.505	0.505	0.768	0.833
	Gene based codon gain	0.763	0.659	0.738	0.696	0.503	0.505	0.781	0.777
Capreomycin	Label Encode	0.884	0.725	0.66	0.691	0.62	0.621	0.88	0.938
	Transition-transversion	0.901	0.868	0.589	0.702	0.649	0.663	0.879	0.978
	Codon frequency encode	0.859	0.785	0.392	0.523	0.452	0.489	0.789	0.973
	Gene based codon gain	0.884	0.767	0.589	0.666	0.598	0.605	0.815	0.956

The table 7.1 represents all the metrics for the pathogen *Mycobacterium tuberculosis*.

Table 7.2: Drug vs metric for *Staphylococcus aureus*

Antibiotic	Method	Accuracy	Precision	Recall	F1 Score	Kappa	MCC	ROC	Specificity
Oxacillin	Label Encode	0.962	0.846	0.916	0.88	0.857	0.858	0.993	0.97
	Transition-transversion	0.962	0.846	0.916	0.88	0.857	0.858	0.993	0.97
	Codon frequency encode	0.974	0.916	0.916	0.916	0.901	0.901	0.996	0.985
	Gene based codon gain	0.974	0.857	1	0.923	0.908	0.911	0.992	0.97
Methicillin	Label Encode	0.962	0.976	0.982	0.979	0.738	0.738	0.93	0.733
	Transition-transversion	0.967	0.976	0.988	0.982	0.768	0.77	0.941	0.733
	Codon frequency encode	0.972	0.976	0.994	0.985	0.8	0.806	0.95	0.73
	Gene based codon gain	0.956	0.976	0.976	0.976	0.709	0.709	0.932	0.733
Cefoxitin	Label Encode	0.98	0.987	0.987	0.987	0.946	0.946	0.997	0.959
	Transition-transversion	0.98	0.99	0.981	0.987	0.947	0.947	0.997	0.979
	Codon frequency encode	0.971	0.987	0.974	0.98	0.92	0.921	0.992	0.959
	Gene based codon gain	0.975	0.993	0.974	0.984	0.934	0.935	0.992	0.979
Ciprofloxacin	Label Encode	0.94	0.976	0.931	0.953	0.872	0.873	0.978	0.957
	Transition-transversion	0.94	0.968	0.938	0.953	0.872	0.872	0.984	0.943
	Codon frequency encode	0.94	0.976	0.931	0.953	0.872	0.873	0.984	0.957
	Gene based codon gain	0.985	0.992	0.984	0.988	0.967	0.967	0.996	0.985
Trimethoprim/sulfamethoxazole	Label Encode	0.945	0.941	0.864	0.901	0.864	0.865	0.94	0.978
	Transition-transversion	0.945	0.916	0.891	0.904	0.866	0.866	0.947	0.967
	Codon frequency encode	0.953	0.942	0.891	0.916	0.884	0.885	0.949	0.978
	Gene based codon gain	0.937	0.891	0.891	0.891	0.848	0.848	0.95	0.956
Chloramphenicol	Label Encode	0.986	1	0.75	0.857	0.85	0.86	0.993	1
	Transition-transversion	0.986	1	0.75	0.857	0.85	0.86	0.982	1
	Codon frequency encode	0.986	1	0.75	0.857	0.85	0.86	0.972	1
	Gene based codon gain	0.986	1	0.75	0.857	0.85	0.86	0.972	1
Tetracycline	Label Encode	0.888	0.893	0.7	0.785	0.711	0.721	0.898	0.965
	Transition-transversion	0.888	0.911	0.683	0.78	0.709	0.721	0.886	0.972
	Codon frequency encode	0.913	0.875	0.816	0.844	0.784	0.785	0.928	0.952
	Gene based codon gain	0.898	0.897	0.733	0.807	0.739	0.746	0.917	0.965
Gentamicin	Label Encode	0.977	1	0.872	0.931	0.918	0.921	0.943	1
	Transition-transversion	0.981	1	0.893	0.943	0.932	0.934	0.9444	1
	Codon frequency encode	0.981	1	0.893	0.943	0.932	0.934	0.95	1
	Gene based codon gain	0.977	1	0.872	0.931	0.918	0.921	0.941	1
Erythromycin	Label Encode	0.819	0.74	0.926	0.823	0.642	0.659	0.895	0.729
	Transition-transversion	0.826	0.753	0.918	0.827	0.656	0.669	0.891	0.75
	Codon frequency encode	0.822	0.809	0.796	0.803	0.642	0.642	0.902	0.844
	Gene based codon gain	0.811	0.736	0.91	0.814	0.627	0.642	0.892	0.729
Fusidic acid	Label Encode	0.932	0.875	0.795	0.833	0.791	0.792	0.97	0.969
	Transition-transversion	0.942	0.88	0.84	0.86	0.824	0.824	0.973	0.969
	Codon frequency encode	0.961	0.891	0.931	0.911	0.886	0.886	0.984	0.969
	Gene based codon gain	0.937	0.86	0.84	0.85	0.811	0.811	0.974	0.963
Clindamycin	Label Encode	0.945	0.96	0.906	0.932	0.886	0.887	0.952	0.973
	Transition-transversion	0.945	0.96	0.905	0.932	0.886	0.887	0.957	0.973
	Codon frequency encode	0.937	0.941	0.905	0.923	0.871	0.871	0.959	0.96
	Gene based codon gain	0.945	0.96	0.905	0.932	0.886	0.887	0.963	0.973

The table 7.2 represents all the metrics for the pathogen *Staphylococcus aureus*.

Table 7.3: Drug vs metric for *Streptococcus pneumoniae*

Antibiotic	Method	Accuracy	Precision	Recall	F1 Score	Kappa	MCC	ROC	Specificity
Erythromycin	Label Encode	0.938	0.92	0.9	0.91	0.863	0.863	0.968	0.958
	Transition-transversion	0.938	0.902	0.922	0.912	0.864	0.864	0.946	0.972
	Codon frequency encode	0.938	0.92	0.9	0.91	0.868	0.863	0.968	0.958
	Gene based codon gain	0.941	0.912	0.922	0.917	0.872	0.872	0.969	0.952
Penicillin	Label Encode	0.946	0.948	0.948	0.948	0.892	0.892	0.943	0.979
	Transition-transversion	0.946	0.955	0.94	0.948	0.892	0.892	0.975	0.952
	Codon frequency encode	0.946	0.952	0.944	0.948	0.892	0.892	0.975	0.947
	Gene based codon gain	0.946	0.941	0.956	0.949	0.892	0.892	0.973	0.934
Tetracycline	Label Encode	0.966	0.956	0.956	0.956	0.928	0.928	0.98	0.972
	Transition-transversion	0.963	0.955	0.95	0.952	0.923	0.923	0.979	0.972
	Codon frequency encode	0.966	0.956	0.956	0.956	0.928	0.928	0.975	0.972
	Gene based codon gain	0.963	0.944	0.962	0.953	0.923	0.924	0.986	0.964
Trimethoprim/sulfamethoxazole	Label Encode	0.943	0.966	0.955	0.961	0.854	0.854	0.967	0.907
	Transition-transversion	0.946	0.971	0.955	0.963	0.863	0.864	0.965	0.921
	Codon frequency encode	0.943	0.962	0.962	0.961	0.853	0.853	0.963	0.894
	Gene based codon gain	0.95	0.969	0.962	0.966	0.871	0.871	0.975	0.914
Chloramphenicol	Label Encode	0.969	0.853	0.833	0.843	0.826	0.826	0.925	0.984
	Transition-transversion	0.971	0.916	0.785	0.846	0.83	0.833	0.939	0.992
	Codon frequency encode	0.966	0.888	0.761	0.82	0.802	0.805	0.93	0.989
	Gene based codon gain	0.966	0.868	0.785	0.825	0.806	0.806	0.924	0.986
Clindamycin	Label Encode	0.962	1	0.66	0.8	0.78	0.8	0.899	1
	Transition-transversion	0.962	1	0.66	0.8	0.78	0.8	0.899	1
	Codon frequency encode	0.962	1	0.66	0.8	0.78	0.8	0.899	1
	Gene based codon gain	0.962	1	0.66	0.8	0.78	0.8	0.899	1

The table 7.3 represents all the metrics for the pathogen *Streptococcus pneumoniae*.

Table 7.4: Drug vs metric for *Pseudomonas aeruginosa*

Antibiotic	Method	Accuracy	Precision	Recall	F1 Score	Kappa	MCC	ROC	Specificity
Ceftazidime	Label Encode	0.676	0.619	0.698	0.656	0.352	0.354	0.681	0.658
	Transition-transversion	0.697	0.647	0.698	0.671	0.391	0.392	0.689	0.696
	Codon frequency encode	0.661	0.615	0.634	0.625	0.317	0.317	0.67	0.683
	Gene based codon gain	0.64	0.603	0.555	0.578	0.266	0.267	0.672	0.708
Tobramycin	Label Encode	0.9	0.827	0.8	0.813	0.745	0.745	0.912	0.937
	Transition-transversion	0.89	0.781	0.833	0.806	0.73	0.731	0.907	0.912
	Codon frequency encode	0.89	0.8	0.8	0.8	0.72	0.72	0.901	0.925
	Gene based codon gain	0.9	0.827	0.8	0.813	0.745	0.745	0.912	0.937
Ciprofloxacin	Label Encode	0.844	0.87	0.824	0.846	0.688	0.689	0.906	0.865
	Transition-transversion	0.834	0.882	0.789	0.833	0.67	0.674	0.896	0.884
	Codon frequency encode	0.871	0.921	0.824	0.87	0.743	0.748	0.905	0.923
	Gene based codon gain	0.825	0.851	0.807	0.828	0.651	0.652	0.909	0.846
Amikacin	Label Encode	0.831	0.5	0.35	0.411	0.317	0.324	0.75	0.929
	Transition-transversion	0.831	0.5	0.35	0.411	0.317	0.324	0.748	0.929
	Codon frequency encode	0.831	0.5	0.2	0.285	0.209	0.238	0.754	0.959
	Gene based codon gain	0.815	0.428	0.3	0.352	0.248	0.254	0.772	0.919
Levofloxacin	Label Encode	0.683	0.641	0.653	0.647	0.36	0.36	0.712	0.707
	Transition-transversion	0.675	0.625	0.673	0.648	0.347	0.348	0.706	0.676
	Codon frequency encode	0.735	0.677	0.769	0.72	0.47	0.473	0.73	0.707
	Gene based codon gain	0.641	0.586	0.653	0.618	0.281	0.282	0.711	0.63
Meropenem	Label Encode	0.602	0.671	0.529	0.592	0.215	0.221	0.631	0.69
	Transition-transversion	0.608	0.65	0.611	0.63	0.216	0.216	0.62	0.605
	Codon frequency encode	0.615	0.676	0.564	0.615	0.236	0.24	0.634	0.676
	Gene based codon gain	0.653	0.666	0.729	0.696	0.295	0.297	0.693	0.567

The table 7.4 represents all the metrics for the pathogen *Pseudomonas aeruginosa*.

Table 7.5: Drug vs metric for *Escherichia coli*

Antibiotic	Method	Accuracy	Precision	Recall	F1 Score	Kappa	MCC	ROC	Specificity
Ciprofloxacin	Label Encode	0.964	0.973	0.903	0.937	0.912	0.913	0.986	0.99
	Transition-transversion	0.959	0.928	0.935	0.931	0.903	0.903	0.982	0.97
	Codon frequency encode	0.957	0.927	0.927	0.927	0.897	0.897	0.984	0.97
	Gene based codon gain	0.959	0.949	0.911	0.93	0.901	0.902	0.983	0.98
Cefotaxime	Label Encode	0.884	0.629	0.735	0.678	0.608	0.611	0.893	0.914
	Transition-transversion	0.887	0.644	0.716	0.678	0.61	0.612	0.886	0.921
	Codon frequency encode	0.89	0.66	0.698	0.678	0.613	0.613	0.87	0.929
	Gene based codon gain	0.862	0.567	0.716	0.633	0.55	0.556	0.871	0.891
Cefoxitin	Label Encode	0.89	0.636	0.466	0.538	0.478	0.485	0.879	0.957
	Transition-transversion	0.89	0.666	0.4	0.5	0.443	0.461	0.869	0.968
	Codon frequency encode	0.845	0.444	0.533	0.484	0.394	0.394	0.745	0.894
	Gene based codon gain	0.881	0.583	0.466	0.518	0.452	0.455	0.856	0.947
Trimethoprim	Label Encode	0.772	0.779	0.73	0.754	0.543	0.544	0.804	0.811
	Transition-transversion	0.757	0.746	0.746	0.746	0.514	0.514	0.815	0.768
	Codon frequency encode	0.734	0.741	0.682	0.71	0.466	0.468	0.785	0.782
	Gene based codon gain	0.734	0.705	0.761	0.732	0.47	0.471	0.771	0.71
Gentamicin	Label Encode	0.872	0.544	0.627	0.582	0.507	0.509	0.882	0.913
	Transition-transversion	0.831	0.443	0.728	0.551	0.455	0.476	0.873	0.848
	Codon frequency encode	0.86	0.5	0.677	0.579	0.498	0.505	0.866	0.89
	Gene based codon gain	0.872	0.568	0.423	0.485	0.414	0.42	0.855	0.946
Tobramycin	Label Encode	0.865	0.761	0.484	0.592	0.517	0.535	0.874	0.961
	Transition-transversion	0.871	0.833	0.454	0.588	0.52	0.553	0.878	0.977
	Codon frequency encode	0.841	0.577	0.787	0.666	0.565	0.577	0.858	0.854
	Gene based codon gain	0.829	0.553	0.787	0.65	0.54	0.556	0.851	0.839
Ampicillin	Label Encode	0.73	0.777	0.839	0.805	0.366	0.369	0.783	0.511
	Transition-transversion	0.73	0.77	0.847	0.807	0.361	0.365	0.782	0.77
	Codon frequency encode	0.712	0.811	0.74	0.774	0.38	0.383	0.754	0.656
	Gene based codon gain	0.73	0.812	0.774	0.792	0.406	0.407	0.782	0.641
Aztreonam	Label Encode	0.828	0.629	0.68	0.653	0.54	0.54	0.913	0.87
	Transition-transversion	0.857	0.708	0.68	0.693	0.6	0.6	0.933	0.912
	Codon frequency encode	0.847	0.695	0.64	0.666	0.568	0.568	0.925	0.912
	Gene based codon gain	0.866	0.761	0.64	0.695	0.611	0.614	0.93	0.937

The table 7.5 represents all the metrics for the pathogen *Escherichia coli*.

Table 7.6: Drug vs metric for *Klebsiella pneumoniae*

Antibiotic	Method	Accuracy	Precision	Recall	F1 Score	Kappa	MCC	ROC	Specificity
Ceftazidime	Label Encode	0.852	0.968	0.861	0.912	0.466	0.498	0.888	0.78
	Transition-transversion	0.902	0.946	0.943	0.944	0.519	0.519	0.854	0.58
	Codon frequency encode	0.909	0.948	0.948	0.948	0.545	0.548	0.86	0.6
	Gene based codon gain	0.9	0.943	0.943	0.943	0.503	0.503	0.855	0.56
Gentamicin	Label Encode	0.803	0.746	0.822	0.782	0.604	0.606	0.865	0.789
	Transition-transversion	0.803	0.753	0.807	0.779	0.602	0.603	0.854	0.8
	Codon frequency encode	0.814	0.747	0.857	0.799	0.628	0.633	0.863	0.781
	Gene based codon gain	0.825	0.777	0.832	0.803	0.646	0.648	0.87	0.819
Imipenem	Label Encode	0.888	0.865	0.849	0.857	0.765	0.765	0.935	0.913
	Transition-transversion	0.883	0.845	0.861	0.853	0.756	0.756	0.926	0.897
	Codon frequency encode	0.878	0.848	0.842	0.845	0.744	0.744	0.904	0.901
	Gene based codon gain	0.878	0.866	0.817	0.841	0.742	0.743	0.918	0.917
Meropenem	Label Encode	0.874	0.786	0.911	0.844	0.74	0.746	0.928	0.852
	Transition-transversion	0.885	0.792	0.941	0.86	0.764	0.772	0.934	0.852
	Codon frequency encode	0.872	0.845	0.805	0.825	0.725	0.725	0.923	0.912
	Gene based codon gain	0.876	0.835	0.835	0.835	0.737	0.737	0.934	0.901
Ciprofloxacin	Label Encode	0.924	0.978	0.929	0.953	0.754	0.761	0.957	0.9
	Transition-transversion	0.909	0.988	0.901	0.943	0.724	0.743	0.955	0.95
	Codon frequency encode	0.922	0.981	0.924	0.951	0.75	0.759	0.965	0.912
	Gene based codon gain	0.934	0.969	0.952	0.96	0.774	0.775	0.949	0.85
Cefoxitin	Label Encode	0.817	0.888	0.773	0.826	0.636	0.643	0.875	0.875
	Transition-transversion	0.812	0.891	0.759	0.82	0.626	0.635	0.862	0.88
	Codon frequency encode	0.846	0.879	0.842	0.86	0.689	0.69	0.881	0.851
	Gene based codon gain	0.828	0.875	0.81	0.841	0.654	0.656	0.873	0.851
Aztreonam	Label Encode	0.893	0.9111	0.974	0.941	0.273	0.301	0.693	0.232
	Transition-transversion	0.885	0.911	0.965	0.937	0.252	0.269	0.714	0.232
	Codon frequency encode	0.882	0.91	0.962	0.936	0.245	0.259	0.585	0.232
	Gene based codon gain	0.872	0.914	0.945	0.929	0.255	0.256	0.704	0.279
Levofloxacin	Label Encode	0.933	0.986	0.93	0.957	0.805	0.812	0.973	0.947
	Transition-transversion	0.941	0.983	0.943	0.962	0.823	0.827	0.967	0.934
	Codon frequency encode	0.938	0.98	0.943	0.961	0.815	0.818	0.971	0.921
	Gene based codon gain	0.938	0.986	0.936	0.961	0.818	0.824	0.974	0.947

The table 7.6 represents all the metrics for the pathogen *Klebsiella pneumoniae*.

7.0.1 MCC and F1 score plots vs Drugs

Separate bar plots have been plotted to represent the F1 scores and MCC for all six organisms across different antibiotics. Each organism is represented with two plots: one illustrating the F1 scores and the other showing the MCC values. Each antibiotic's performance is evaluated using four encoding techniques explained. In the plot, every drug is represented by a cluster of four bars, with each bar corresponding to a specific encoding technique. The height of the bars indicates F1 score or MCC, allowing for the direct comparison of the efficacy of each encoding technique.

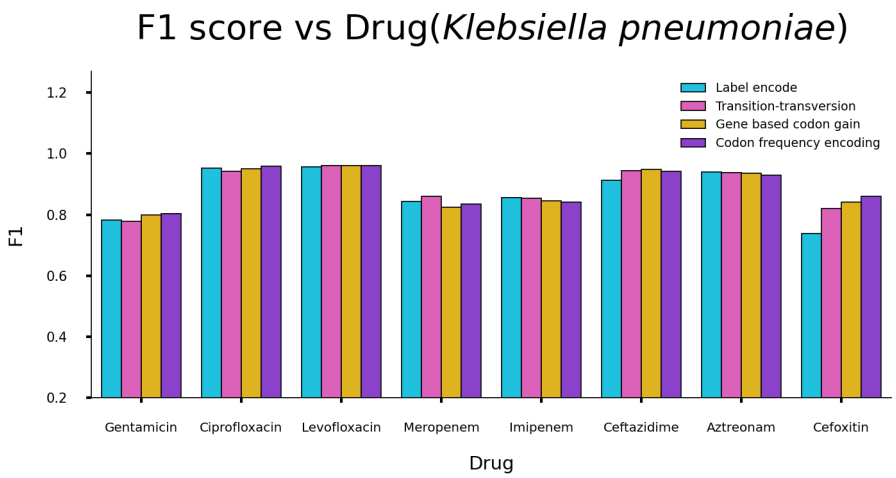

Figure 7.1: F1 vs drug for *Klebsiella pneumonia*

The figure 7.1 displays F1 score plot for *Klebsiella pneumoniae* for 8 different drugs, Levofloxacin showing higest F1 score. For all the drugs except Aztreonam one of the biologically intuitive techniques have performed better than label encode, but the enhancement compared to label encoding is not consistently observed with any one method.

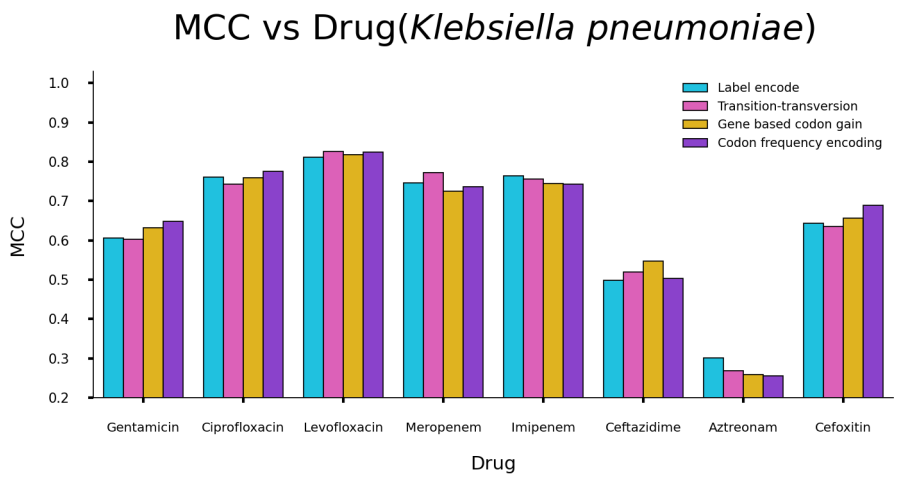

Figure 7.2: MCC vs drug for *Klebsiella pneumonia*

MCC values for *Klebsiella pneumoniae* for 8 different drugs is plotted in figure 7.2, Levofloxacin showing higest and Aztreonam showing lowest MCC values. Label encode results have suppressed the results of all other methods in Imipenem and Aztreonam, But any of of the proposed methods have outperformed label encode in other antibiotics.

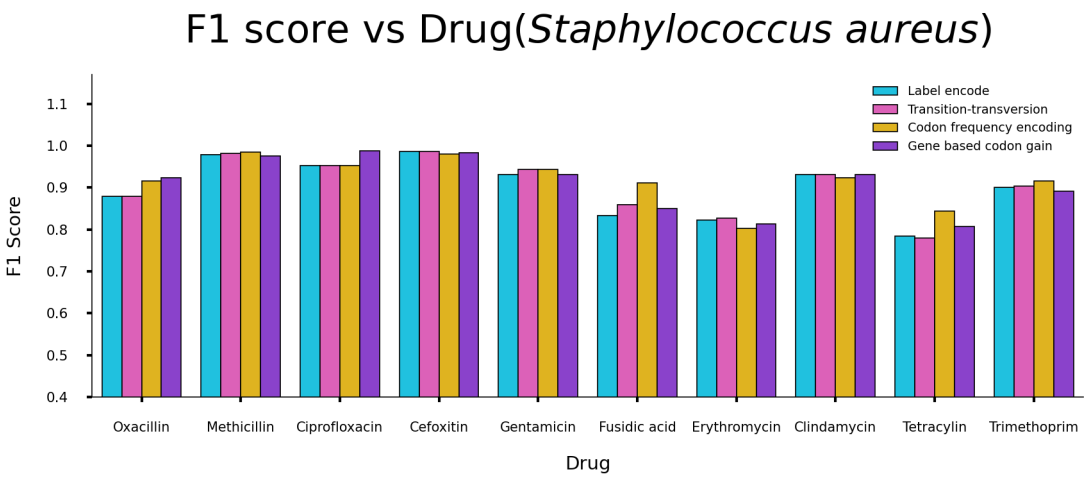

Figure 7.3: F1 vs drug for *Staphylococcus aureus*

The F1 score plot for *Staphylococcus aureus* for 10 different antibiotics is shown in figure 7.3. One of the biologically intuitive encoding has exhibited improved/at par label encoding for all drugs, with Cefoxitin having over all highest F1 score for all the methods.

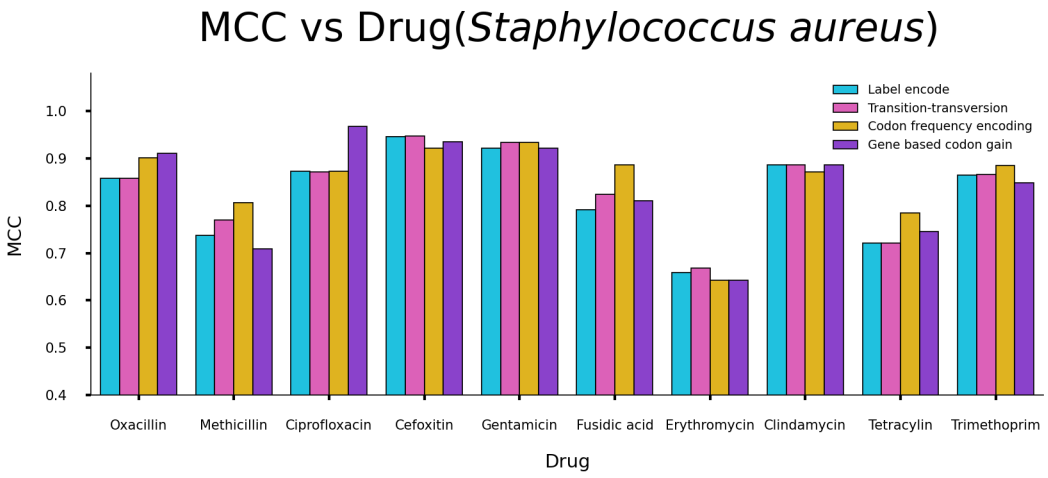

Figure 7.4: MCC vs drug for *Staphylococcus aureus*

In MCC plot, 7.4 for *Staphylococcus aureus*, label encoding values have been suppressed in all the cases. Among which codon frequency encoding have performed better in 6 drugs.

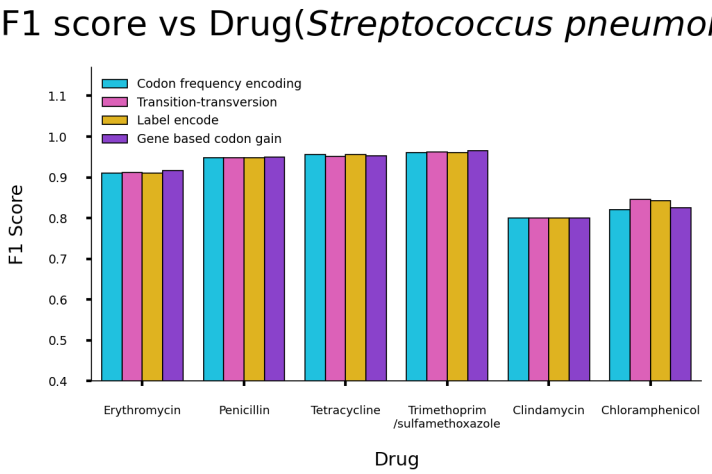

Figure 7.5: F1 vs drug for *Streptococcus pneumonia*

The F1 score 7.5 for *Streptococcus pneumonia* for 6 drug shows that all the encoding techniques have performed almost equal in predicting the phenotype, shows the high effectiveness in predictive performance and the possibly recognition of similar patterns by all the techniques.

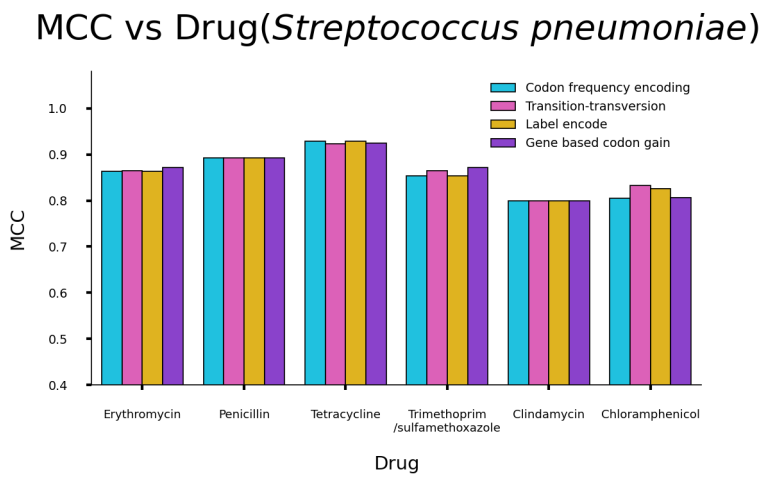

Figure 7.6: MCC vs drug for *Streptococcus pneumonia*

The figure shows 7.6 MCC values for *Streptococcus pneumonia* for 6 different drugs. Codon frequency and gene codon gain techniques have consistently performed better/ equivalent to label encode.

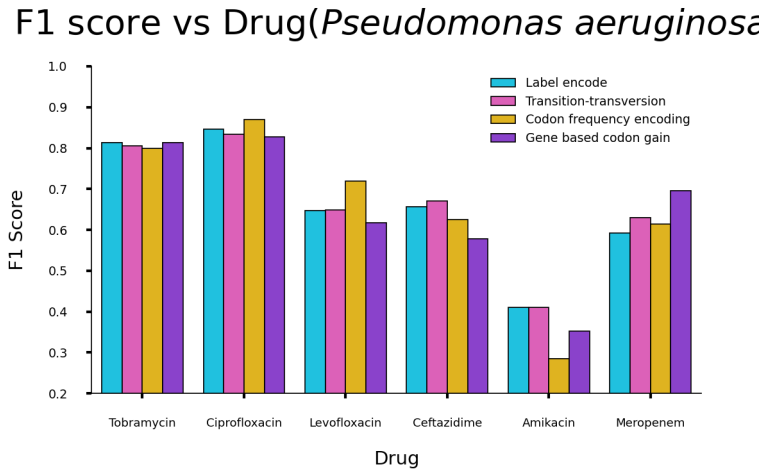

Figure 7.7: F1 vs drug for *Pseudomonas aeruginosa*

The figure 7.7 illustrate F1 score plot for *Pseudomonas aeruginosa* for 6 different drugs, Ciprofloxacin showing highest F1 score. For all the drugs one of the biologically intuitive techniques have performed better than label encode, Amikacin showed lowest F1 score with all the encoding techniques.

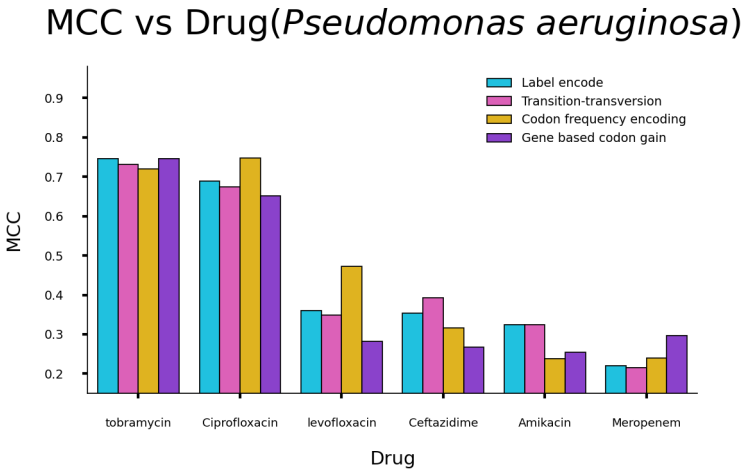

Figure 7.8: MCC vs drug for *Pseudomonas aeruginosa*

The MCC plot 7.8 for *Pseudomonas aeruginosa* shows, Ciprofloxacin have highest F1 score. For all the drugs one of the biologically intuitive techniques have performed better than label encode, despite having better F1 score, Meropenem showed lowest MCC score with all the encoding techniques.

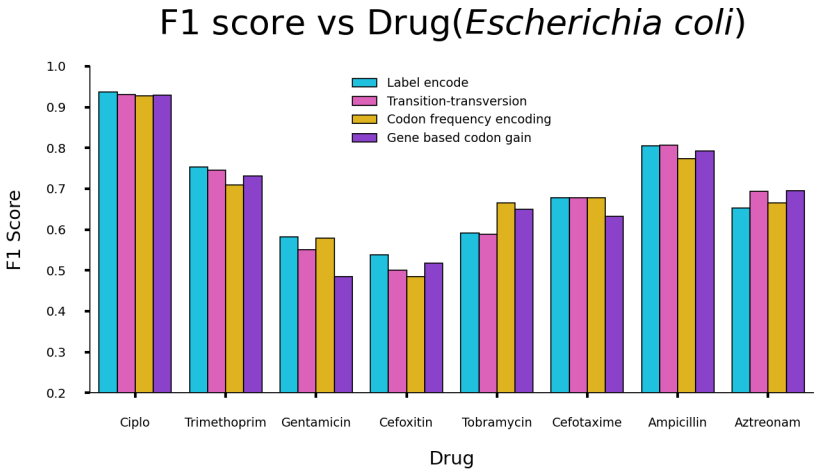

Figure 7.9: F1 vs drug for *Escherichia coli*

In 7.9 F1 score plot for *Escherichia coli* label encoding have outperformd in 4 out 8 antibiotics having highest F1 for Ciprofloxacin. Any one of the proposed encoding have exhibited improved/at par performance in other antibiotics. However, there is no one method consistently showing improvement over label encoding.

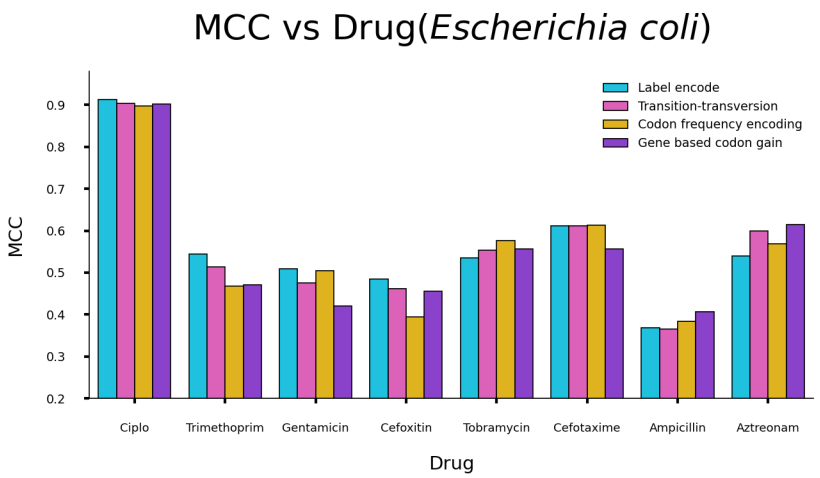

Figure 7.10: MCC vs drug for *Escherichia coli*

For *Escherichia coli* , MCC values have been shown in plot 7.10. Similar to its F1 plot, label encoding have outperformd in 4 drugs. Ampicillin having lowest and Ciprofloxacin having highest MCC score. One of the proposed encoding have exhibited improved/at par performance in other antibiotics.

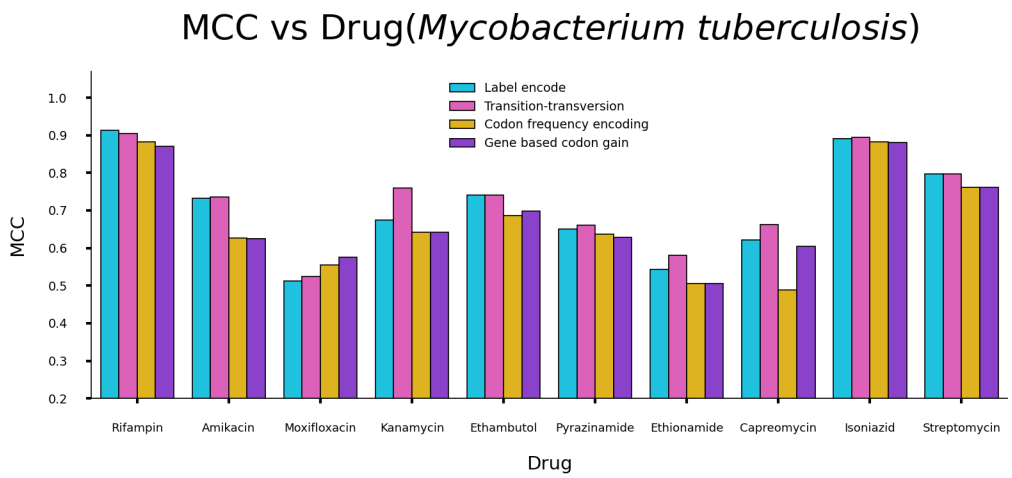

Figure 7.11: F1 vs drug for *Mycobacterium tuberculosis*

MCC values for *Mycobacterium tuberculosis* for 10 different drugs is plotted in figure 7.11. transition-transversion have performed better compared to all other methods in most of the drugs. 9 out of 10 drugs have highest F1 for any one of the biologically intuitive encoding.

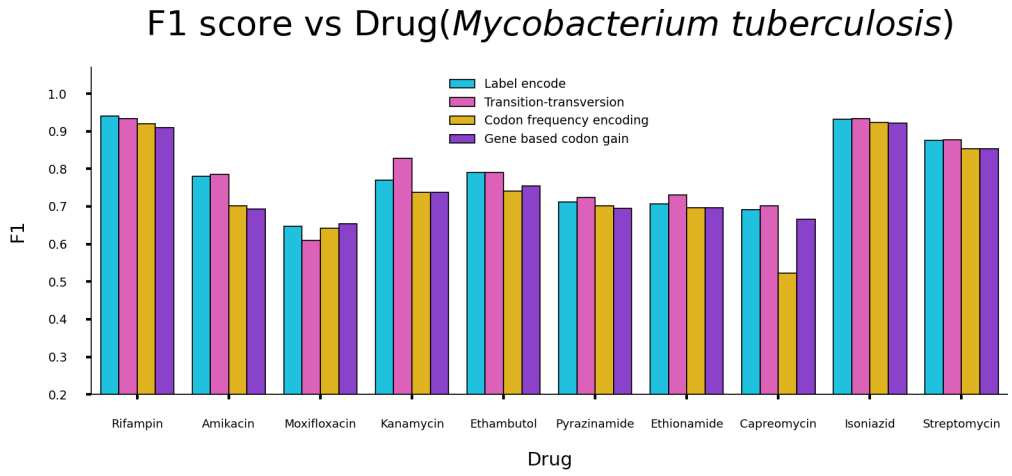

Figure 7.12: MCC vs drug for *Mycobacterium tuberculosis*

The F1 score plot for *Mycobacterium tuberculosis* for 10 different antibiotics is shown in figure 7.12. One of the biologically intuitive encoding has exhibited improved/at par label encoding for all drugs.

No single encoding method have performed consistently better. However, the result analysis shows that the proposed encoding technique holds promise in capturing specific biological patterns.

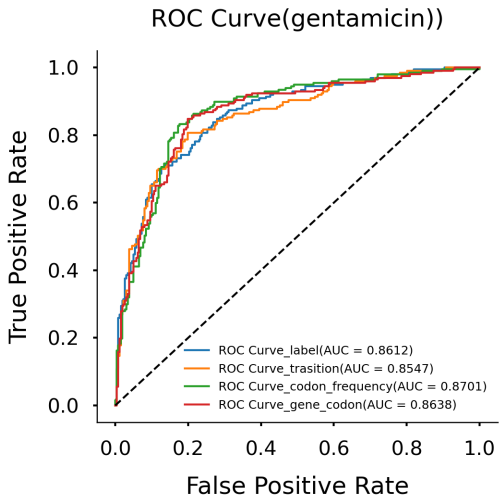

Figure 7.13: ROC Curve for Gentamicin in *Klebsiella pneumoniae*

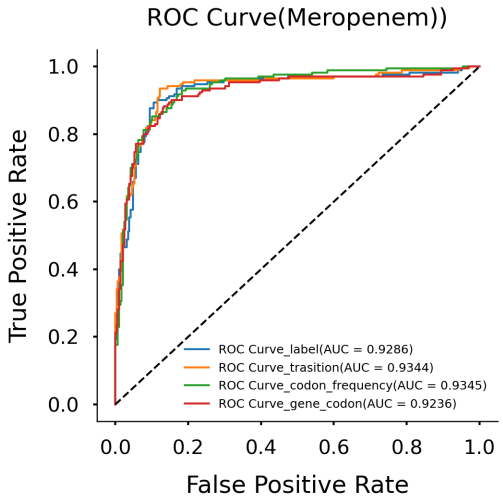

Figure 7.14: ROC Curve for Meropenem in *Klebsiella pneumoniae*

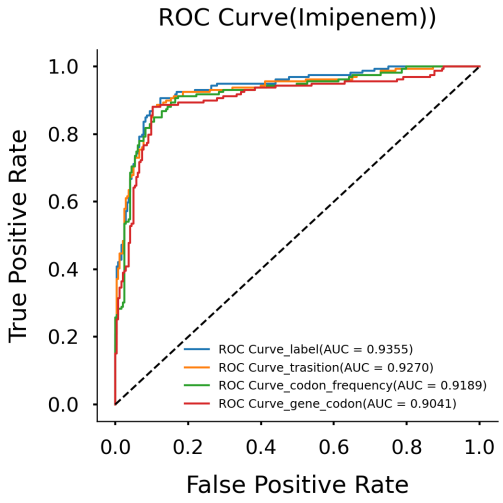

Figure 7.15: ROC Curve for Imipenem in *Klebsiella pneumoniae*

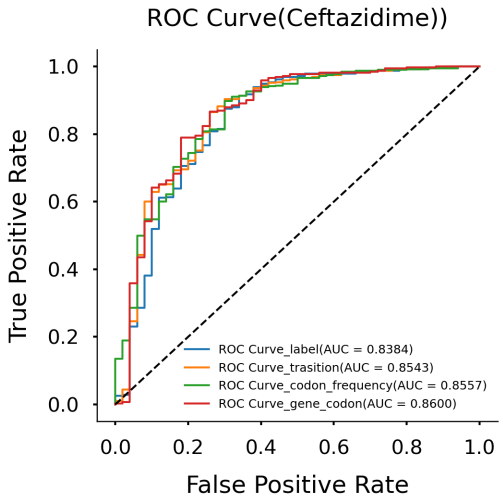

Figure 7.16: ROC Curve for Ceftazidime in *Klebsiella pneumoniae*

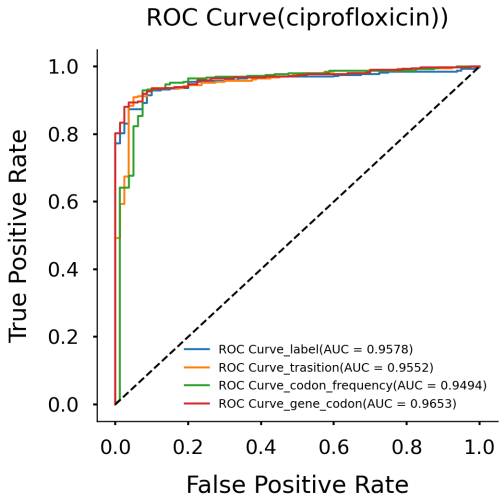

Figure 7.17: ROC Curve for Ciprofloxicin in *Klebsiella pneumoniae*

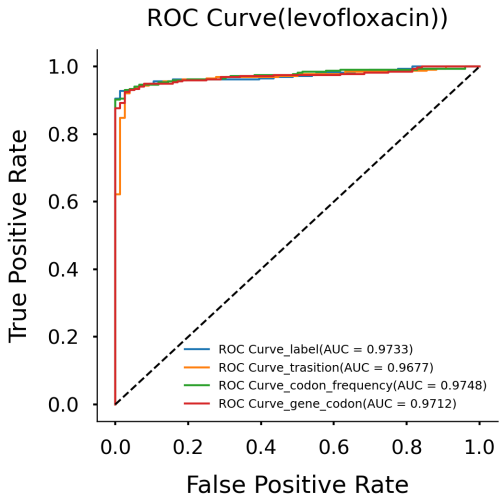

Figure 7.18: ROC Curve for Levofloxicin in *Klebsiella pneumoniae*

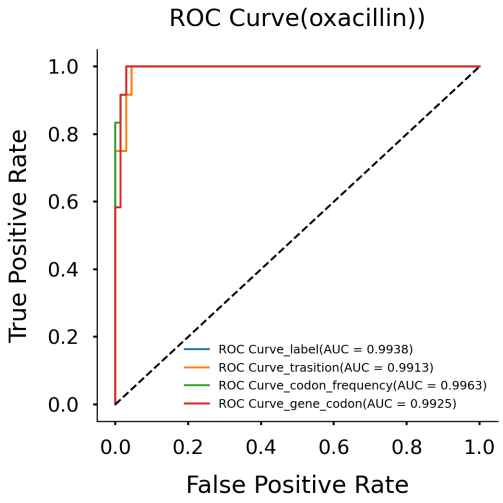

Figure 7.19: ROC Curve for Oxacillin in *Staphylococcus aureus*

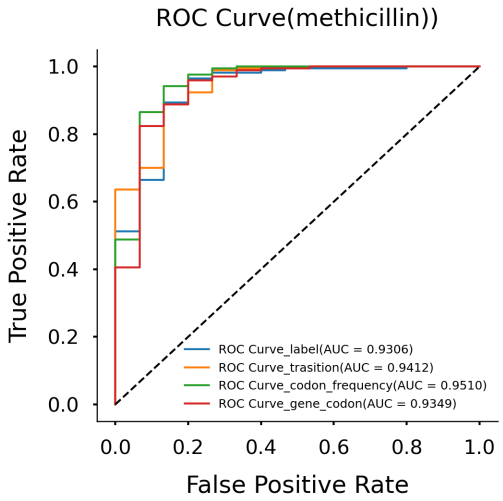

Figure 7.20: ROC Curve for Methicillin in *Staphylococcus aureus*

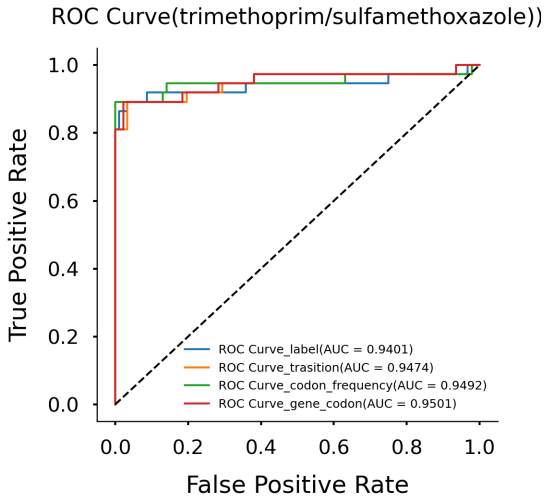

Figure 7.21: ROC Curve for Trimethoprim/sulfamethoxazole in *Staphylococcus aureus*

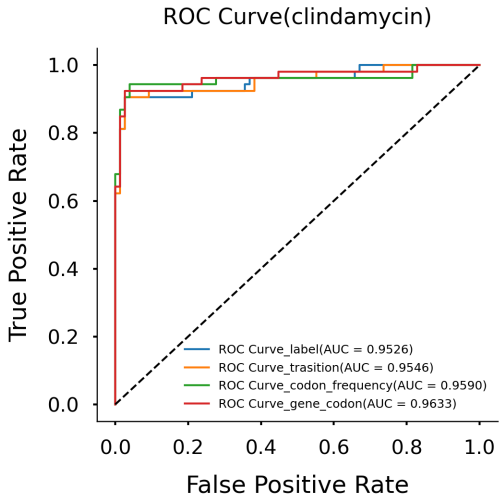

Figure 7.22: ROC Curve for Clindamycin in *Staphylococcus aureus*

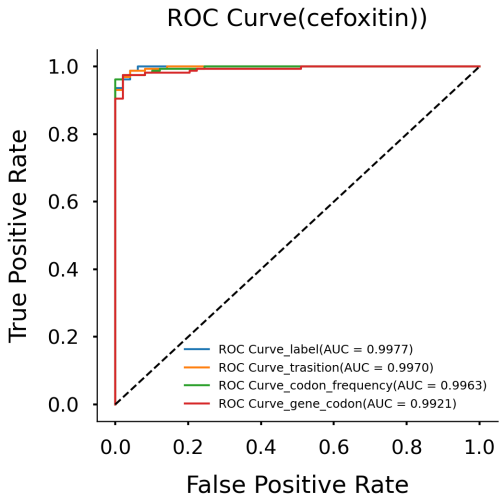

Figure 7.23: ROC Curve for Cefoxitin in *Staphylococcus aureus*

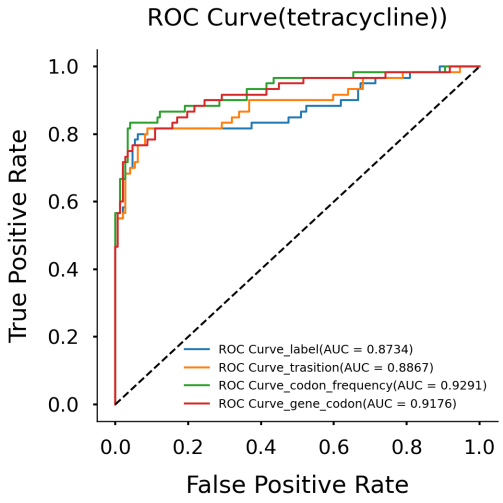

Figure 7.24: ROC Curve for tetracycline in *Staphylococcus aureus*

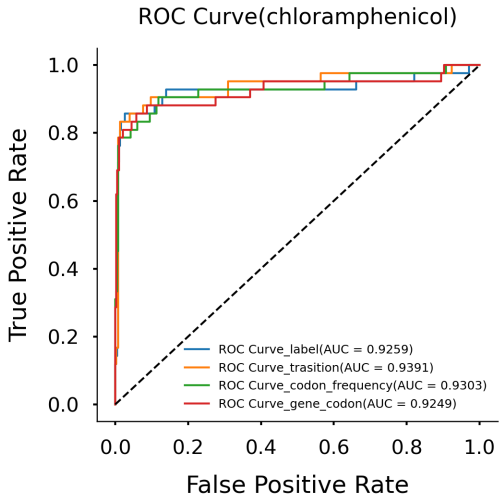

Figure 7.25: ROC Curve for Chloramphenicol in *Streptococcus pneumoniae*

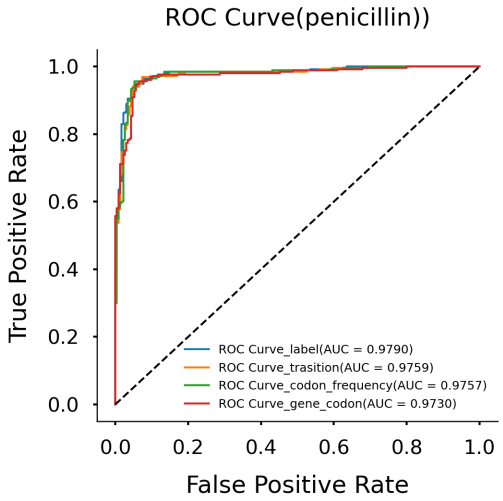

Figure 7.26: ROC Curve for Penicillin in *Streptococcus pneumoniae*

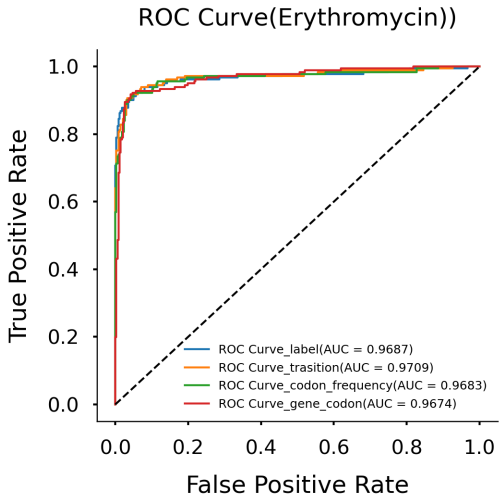

Figure 7.27: ROC Curve for Erythromycin in *Streptococcus pneumoniae*

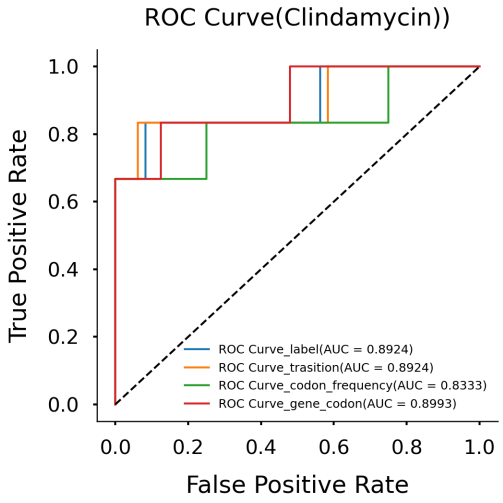

Figure 7.28: ROC Curve for Clindamycin in *Streptococcus pneumoniae*

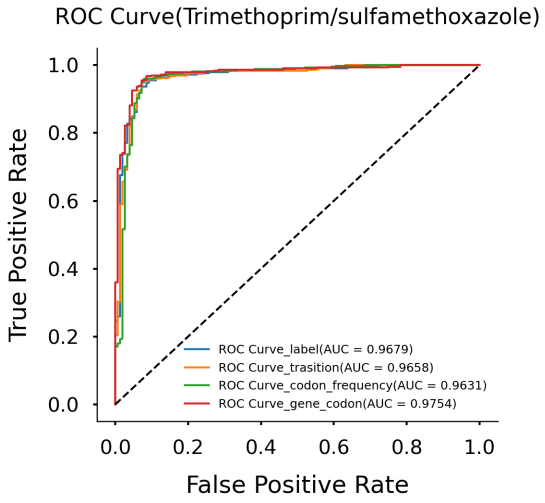

Figure 7.29: ROC Curve for Trimethoprim/sulfamethoxazole in *Streptococcus pneumoniae*

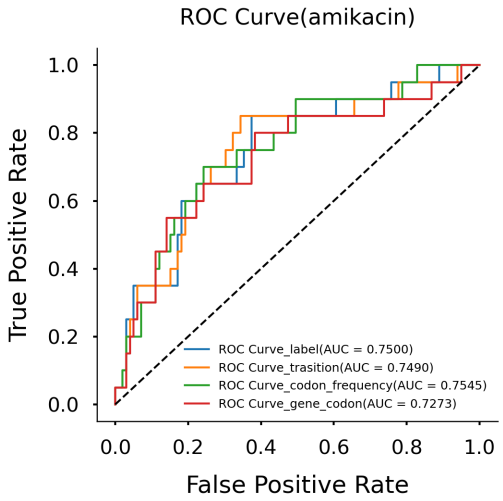

Figure 7.30: ROC Curve for Amikacin in *Pseudomonas aeruginosa*

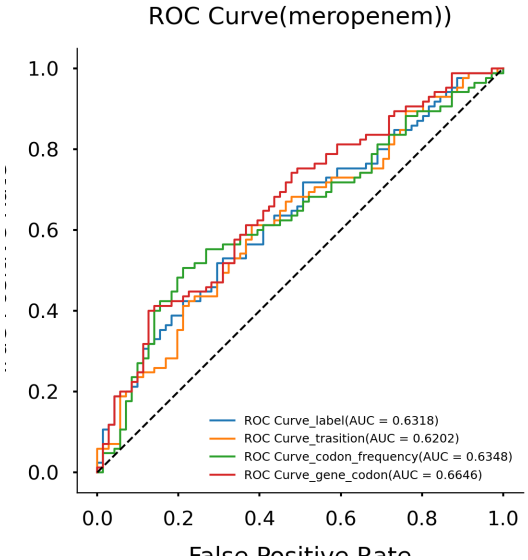

Figure 7.31: ROC Curve for Meropenem in *Pseudomonas aeruginosa*

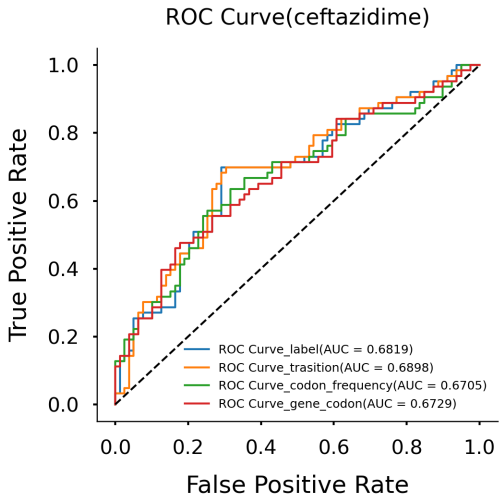

Figure 7.32: ROC Curve for Ceftazidime in *Pseudomonas aeruginosa*

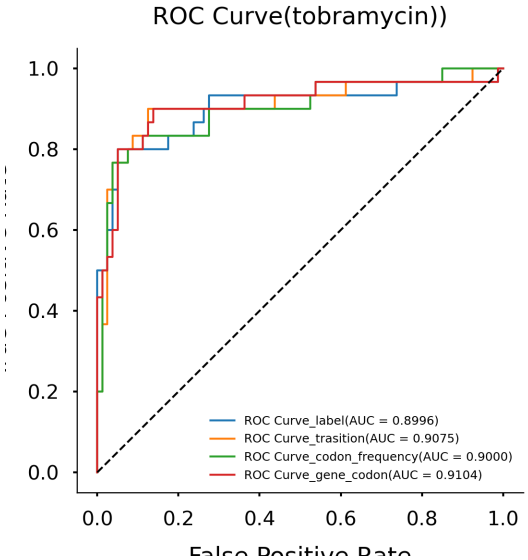

Figure 7.33: ROC Curve for Tobramycin in *Pseudomonas aeruginosa*

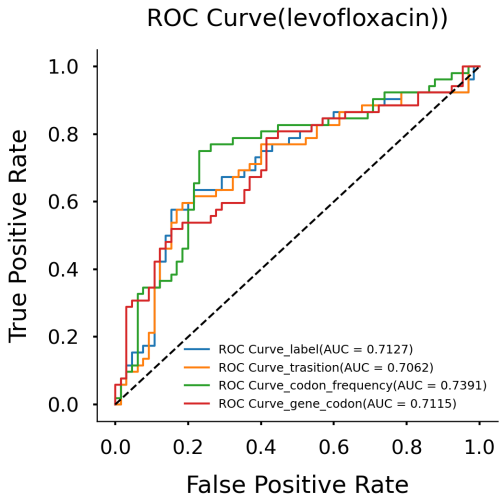

Figure 7.34: ROC Curve for Levofloxacin in *Pseudomonas aeruginosa*

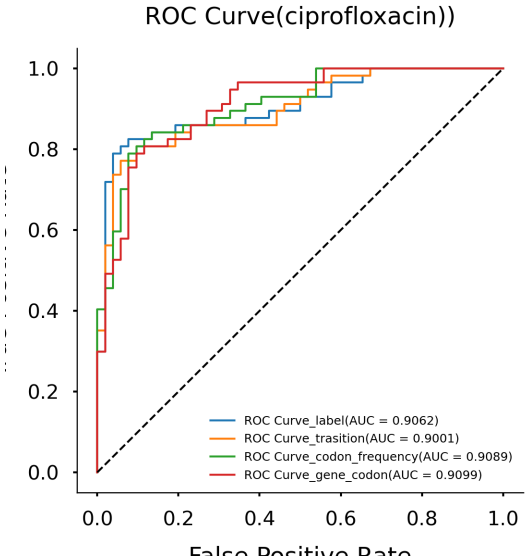

Figure 7.35: ROC Curve for Ciprofloxacin in *Pseudomonas aeruginosa*

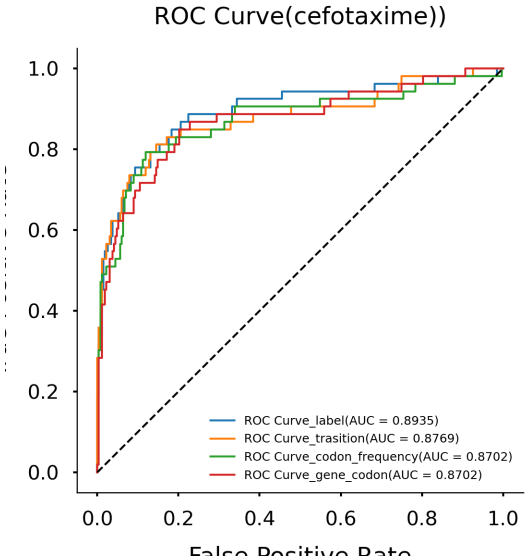

Figure 7.36: ROC Curve for Cefotaxime in *Escherichia coli*

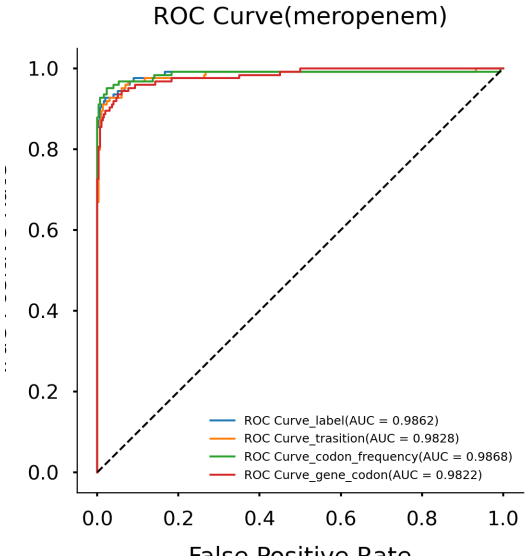

Figure 7.37: ROC Curve for Ciprofloxacin in *Escherichia coli*

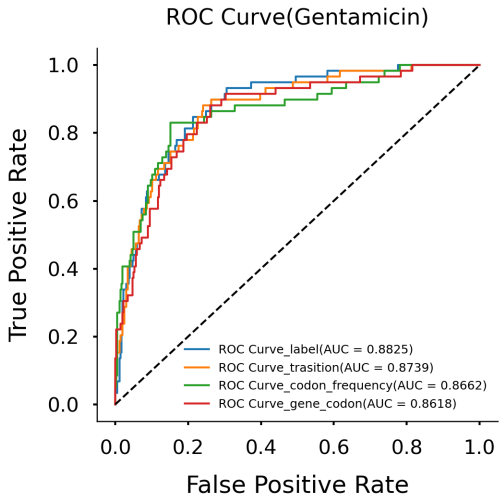

Figure 7.38: ROC Curve for Gentamicin in *Escherichia coli*

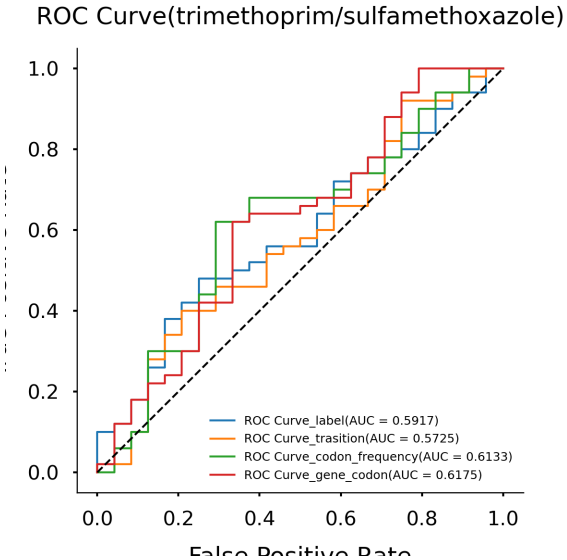

Figure 7.39: ROC Curve for Trimethoprim in *Escherichia coli*

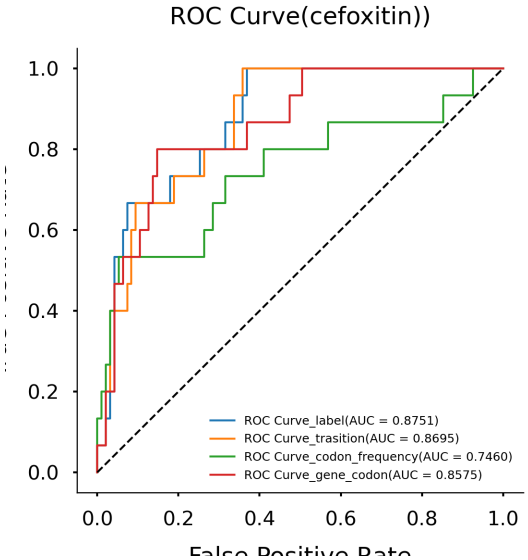

Figure 7.40: ROC Curve for Cefoxitin in *Escherichia coli*

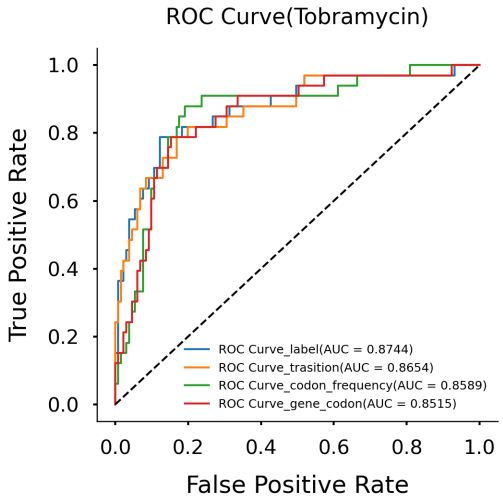

Figure 7.41: ROC Curve for Tobramycin in *Escherichia coli*

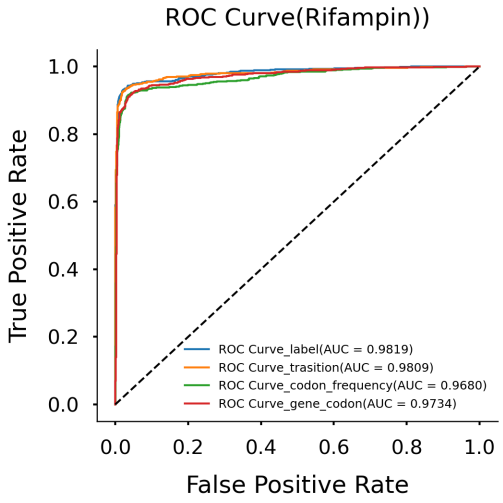

Figure 7.42: ROC Curve for Rifampin in *Mycobacterium tuberculosis*

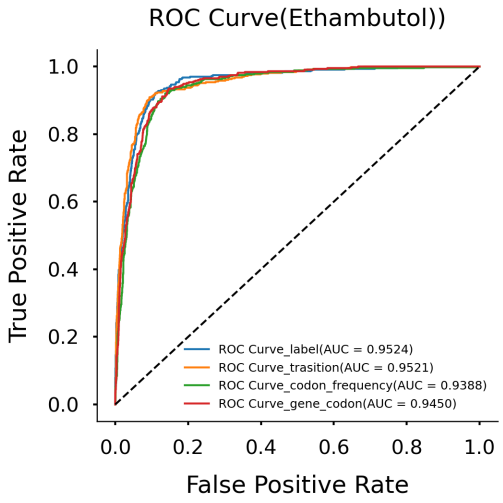

Figure 7.43: ROC Curve for Ethambutol in *Mycobacterium tuberculosis*

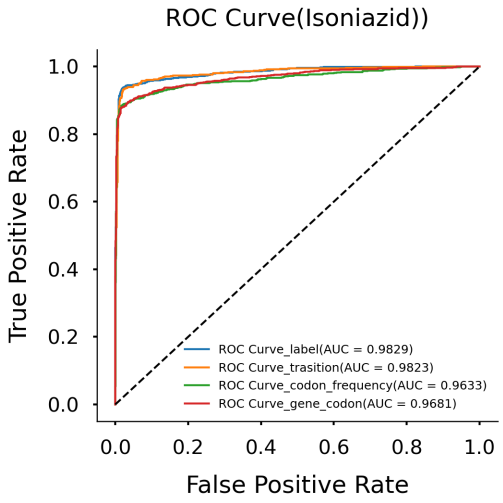

Figure 7.44: ROC Curve for Isoniazid in *Mycobacterium tuberculosis*

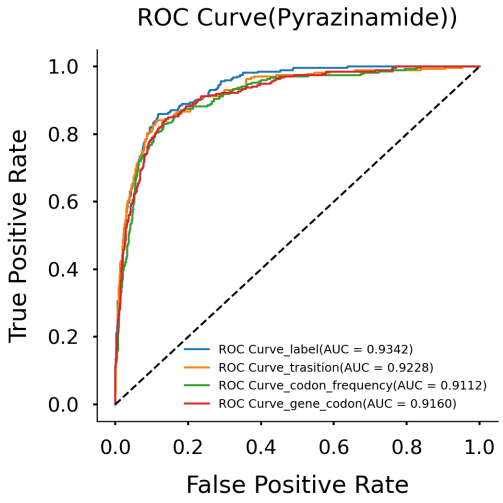

Figure 7.45: ROC Curve for Pyrazinamide in *Mycobacterium tuberculosis*

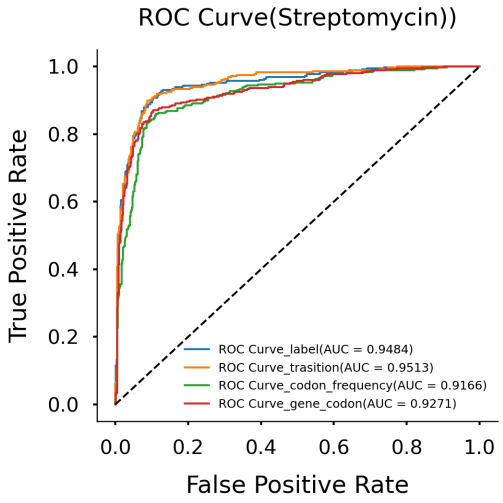

Figure 7.46: ROC Curve for Streptomycin in *Mycobacterium tuberculosis*

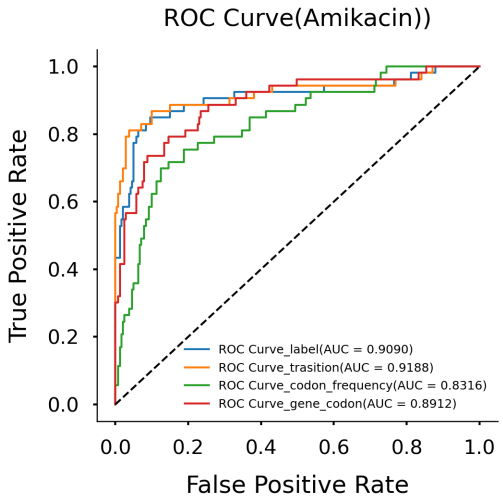

Figure 7.47: ROC Curve for Amikacin in *Mycobacterium tuberculosis*

The ROC-AUC plots above demonstrate the potential of biologically intuitive encoding for predicting AMR profiles across different pathogens. Although no method consistently

surpassed the ROC-AUC value achieved by label encoding—indeed, label encoding sometimes outperformed the other methods—there is significant room for improvement in classification results, depending on the specific drug-pathogen pair being analyzed.

Chapter 8

Discussion

In this study, several new encoding techniques for SNP data are proposed, which are biologically intuitive. These include transition-transversion encoding, codon frequency encoding, and gene-based codon gain. It's a well-established principle in machine learning that no single model or algorithm is universally suitable for all purposes or all types of data. However, the proposed encoding technique holds promise in capturing specific biological patterns derived from molecular mechanisms, thereby enabling more meaningful and improved classification. The results indicate that in many cases, one of these biologically intuitive encodings provides better results than simple label encoding. However, there are still several biological contexts that could be addressed. For instance, challenges such as horizontal gene transfer make it difficult to capture AMR resistance. Effectively capturing cross-resistance is also challenging. Additionally, important mutations in non-coding and regulatory regions are not addressed by codon frequency and gene-codon frequency encoding, making classification difficult. Furthermore, plasmid-based resistance is not accounted for in the current techniques.

Chapter 9

Future Scope

The present study introduces new encoding methods for the SNP matrix. However, thorough investigations are necessary to enhance explainability of the model and increase confidence in its prediction outcomes. In the future, it's also important to validate the impact of domain shifting.

Bibliography

[1] Gustavo Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S Heath, Peter Vikesland, and Liqing Zhang. Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6:1–15, 2018.

[2] Pranav Bhaskar and Bijayani Sahu. Antimicrobial resistance: global concern and the critical need for new antibiotics. *The Applied Biology & Chemistry Journal (TABCJ)*, 4(1):1–3, 2023.

[3] Manish Boolchandani, Alaric W D'Souza, and Gautam Dantas. Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics*, 20(6):356–370, 2019.

[4] Michael L Chen, Akshith Doddi, Jimmy Royer, Luca Freschi, Marco Schito, Matthew Ezewudo, Isaac S Kohane, Andrew Beam, and Maha Farhat. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in mycobacterium tuberculosis resistance prediction. *EBioMedicine*, 43:356–369, 2019.

[5] Alessandra da Silva Dantas. Antimicrobial resistance, 2022.

[6] Guillermo Fernandez. Turning the juggernaut, 2022.

[7] Anna G Green, Chang Ho Yoon, Michael L Chen, Yasha Ektefaie, Mack Fina, Luca Freschi, Matthias I Gröschel, Isaac Kohane, Andrew Beam, and Maha Farhat. A convolutional neural network highlights mutations relevant to antimicrobial resistance in mycobacterium tuberculosis. *Nature communications*, 13(1):3817, 2022.

[8] Xingyan Kuang, Fan Wang, Kyle M Hernandez, Zhenyu Zhang, and Robert L Grossman. Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and cnn. *Scientific reports*, 12(1):2427, 2022.

[9] VA Minimol, Abhay Kumar, and Mukteswar Prasad Mothadaka. Evolution and milestones in the development of amr in bacteria. In *Handbook on Antimicrobial Resistance: Current Status, Trends in Detection and Mitigation Measures*, pages 289–302. Springer, 2023.

[10] Jim O'Neill. Tackling drug-resistant infections globally: final report and recommendations. 2016.

[11] Joshua L Payne, Fabrizio Menardo, Andrej Trauner, Sonia Borrell, Sebastian M Gygli, Chloe Loiseau, Sebastien Gagneux, and Alex R Hall. Transition bias influences the evolution of antibiotic resistance in mycobacterium tuberculosis. *PLoS biology*, 17(5):e3000265, 2019.

[12] Omolara Dorcas Popoola, Roseline Tolulope Feyisola, Titilayo Oyeronke Adesetan, Omowunmi Abosede Banjo, Taiwo Adeolu Dele-Osibanjo, Oluwafemi Daniel Amusa, Kehinde Bolarinwa, Sunday Ebenezer Popoola, Benjamin Thoha Thomas, and Moses Olusola Efuntoye. Transition mutation bias is crucial to adaptive extended spectrum beta lactamase (esbl) resistance evolution. *Scientific African*, 24:e02132, 2024.

[13] Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Anne-Christin Hauschild, Oliver Schwengers, and Dominik Heider. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, 38(2):325–334, 2022.

[14] Md Abdus Salam, Md Yusuf Al-Amin, Moushumi Tabassoom Salam, Jogendra Singh Pawar, Naseem Akhter, Ali A Rabaan, and Mohammed AA Alqumber. Antimicrobial resistance: a growing serious threat for global public health. In *Healthcare*, volume 11, page 1946. MDPI, 2023.

[15] Shafqat Ali Shah et al. Antimicrobial resistance. *Journal of Khyber College of Dentistry*, 12(04), 2022.

[16] Sharan Shyam, Prafull Mohan, and Sharmila Sinha. Antibiotics-access or excess? reviewing alternative targets for antibacterial activity.

[17] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[18] Arlin Stoltzfus and Ryan W Norris. On the causes of evolutionary transition: transversion bias. *Molecular biology and evolution*, 33(3):595–602, 2016.

[19] Ka Wah Kelly Tang, Beverley C Millar, and John E Moore. Antimicrobial resistance (amr). *British Journal of Biomedical Science*, 80:11387, 2023.

[20] Tosin Thompson. The staggering death toll of drug-resistant bacteria. *Nature*, 2022.

[21] Masood ul Haq and Hafiz Muhammad Rizwan. Antimicrobial resistance (amr) in post covid era–'second pandemic'slowly moving across the world. *Journal of Sheikh Zayed Medical College (JSZMC)*, 13(3):01–01, 2022.

[22] Theo Vos, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, Foad Abd-Allah, Ahmed Abdelalim, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The lancet*, 396(10258):1204–1222, 2020.

[23] World Health Organization. Antimicrobial resistance. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance, 2023.

[24] Yang Yang, Katherine E Niehaus, Timothy M Walker, Zamin Iqbal, A Sarah Walker, Daniel J Wilson, Tim EA Peto, Derrick W Crook, E Grace Smith, Tingting Zhu, et al. Machine learning for classifying tuberculosis drug-resistance from dna sequencing data. *Bioinformatics*, 34(10):1666–1671, 2018.

[25] Ying Zhu, Wei E Huang, and Qiwen Yang. Clinical perspective of antimicrobial resistance in bacteria. *Infection and drug resistance*, pages 735–746, 2022.