

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Prediction of Antibiotic Resistance in *Staphylococcus aureus*
Using the Machine Learning Pipeline ALPAR

verfasst von | submitted by

Viktoria Maria Berndorfer BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 066 863

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Masterstudium Biologische Chemie

Betreut von | Supervisor:

Univ.-Prof. David Berry Privatdoz. PhD

Acknowledges

To begin with, I would like to express my sincere gratitude to Professor David Berry for his invaluable supervision, guidance, and support throughout the course of my Master's thesis. Thank you for this great opportunity and for trusting me to pursue my own ideas. I am also deeply grateful to Jan-Philipp Wittlinger for generously providing me with the isolate collection, which was essential for the experimental validation conducted in this study. I particularly value and appreciate his constant support, insightful guidance, and motivating encouragement throughout every step of this thesis. Finally, I want to thank the Gut group for their valuable inspiration and constructive feedback.

Abstract

Antimicrobial resistances (AMR) are an emerging threat to humans and has a great impact in various fields like health, agriculture, animal farming, and the environment. Pathogens like *Staphylococcus aureus* often exhibit multi-drug resistance and are responsible for many hospital-acquired infections worldwide. Advances in artificial intelligence and increasingly available genomic data have motivated researchers to employ machine learning in the prediction of antimicrobial resistance. This thesis aims to establish the novel pipeline ALPAR (Automated Learning Pipeline for Antimicrobial Resistance) for the high priority pathogen *S. aureus*. Publicly available whole genome sequence data is utilized to train machine learning models, which are then evaluated for their performance. Antimicrobial susceptibility testing by Minimum Inhibitory Concentration (MIC) assays is conducted for a *S. aureus* isolate collection to experimentally validate the predictions made by the pipeline. Finally, the predicted resistance phenotypes are compared to the predictions of the benchmark tool ABRicate. In the analysis of the most important features in the prediction process, known resistance genes as well as promising new candidate AMR determinants were identified. Although all models demonstrated high Matthews Correlation Coefficient (MCC) values in validation, their generalization varied significantly. This work emphasizes the importance of experimentally verifying the performance of machine learning models and demonstrates the transferability of ALPAR to other bacterial species.

Kurzfassung

Steigende Antibiotikaresistenzen werden zunehmend mehr zur Gefahr für den Menschen. Von den Auswirkungen sind mehrere Bereiche wie das Gesundheitswesen, die Landwirtschaft, die Viehzucht und die Umwelt betroffen. Krankheitserreger wie *Staphylococcus aureus* sind weltweit für viele Krankenhausinfektionen verantwortlich und weisen zunehmend Multiresistenzen auf, was den Behandlungserfolg schwierig gestaltet. Der technische Fortschritt im Bereich künstliche Intelligenz und die Verfügbarkeit von größeren Datenmengen haben dazu geführt, dass WissenschaftlerInnen maschinelles Lernen im Kampf gegen Antibiotikaresistenz einsetzen. Ziel dieser Arbeit ist es, die neu publizierte Pipeline ALPAR (Automated Learning Pipeline for Antimicrobial Resistance) für *S. aureus* zu etablieren. Für das Trainieren der Modelle, werden öffentlich verfügbare Genomsequenzdaten herangezogen. Anschließend wird deren Leistung anhand verschiedener Bewertungsmaße evaluiert und experimentell validiert. Hierfür werden Minimale Hemmkonzentration (MHK) Tests an einer Kollektion externer *S. aureus* Isolate durchgeführt. Die experimentellen Ergebnisse werden schließlich mit den Vorhersagen der ALPAR Pipeline und dem Referenztool ABRicate verglichen. Eine Analyse der wichtigsten Merkmale im Vorhersage-Prozess konnte viele bereits bekannte Resistenzgene nachweisen und potentiell unbekannte Kandidaten identifizieren. In der Validierung schnitten alle Modelle mit hoher Leistung ab, allerdings konnten starke Unterschiede in der Generalisierbarkeit festgestellt werden. Diese Arbeit betont die Bedeutung der experimentellen Validierung im maschinellen Lernen und zeigt die Übertragbarkeit der ALPAR Pipeline auf andere Bakterienspezies.

Contents

1	Introduction	9
1.1	Public Health Relevance	9
1.1.1	Antibiotic Resistance (AMR) as a Global Threat	9
1.1.2	The Clinical Importance of the ESKAPE Pathogen	10
1.2	Mechanisms of Antibiotic Actions and Resistances	10
1.2.1	β -lactams	11
1.2.2	Aminoglycosides	12
1.2.3	Tetracyclines	12
1.2.4	Macrolides and Lincosamides	12
1.2.5	Fusidic Acid	12
1.2.6	Fluoroquinolones	12
1.2.7	Sulfonamides and Trimethoprim	13
1.3	Biomedical Applications of Artificial Intelligence (AI) and Foundations of Machine Learning	13
1.3.1	AI in Clinical Diagnostics, AMR Surveillance and Drug Discovery	13
1.3.2	Overview on Machine Learning	14
1.4	Architecture and Functionality of the Automated Learning Pipeline for Antimicrobial Resistance (ALPAR)	15
2	Aim of this thesis	17
3	Materials and Methods	19
3.1	Data Collection	19
3.2	Automated Learning Pipeline for Antimicrobial Resistance (ALPAR)	19
3.3	Training and Testing of the Models	19
3.4	Antimicrobial Resistance Prediction for an Isolate Collection	21
3.5	Antibiotic Susceptibility Testing	22
3.6	Comparison of the Prediction Performance for ALPAR and ABRicate	23
4	Results and Discussion	25
4.1	Collection of Training Data for the Machine Learning Models	25
4.2	Phylogenetic Relationships within Training Data	26
4.3	Selection of informative Features via Phylogenetic Scoring	28
4.4	Model Training Procedure and Performance Validation	29
4.5	Assessment of Feature Importance in Random Forest Models	36
4.5.1	Penicillin	37

4.5.2	Other β -lactams (methicillin, ceftiofur, oxacillin)	38
4.5.3	Tetracycline	42
4.5.4	Gentamicin	44
4.5.5	Ciprofloxacin	46
4.5.6	Erythromycin and clindamycin	47
4.5.7	Co-trimoxazole	49
4.5.8	Fusidic acid	51
4.5.9	Vancomycin	53
4.6	Genome Wide Association Study (GWAS) based Identification of Resistance-Associated Variants	55
4.7	Antimicrobial Resistance Predictions for a <i>Staphylococcus aureus</i> Isolate Collection	59
4.8	Experimental validation of Antibiotic Resistance Predictions for Erythromycin and Penicillin	61
5	Conclusion and outlook	67
	Statement of Assistive Technologies	69
	References	80
	Supplementary Material	81

1 | Introduction

1.1 Public Health Relevance

1.1.1 Antibiotic Resistance (AMR) as a Global Threat

Antibiotic resistance is a critical and growing global health concern that undermines the effectiveness of treatments for many infectious diseases. The emergence and spread of antibiotic resistance results from a variety of clinical, environmental and societal factors, emphasizing the urgent need for further research and coordinated countermeasures.

According to the review on antimicrobial resistance by O'Neill, in 2050 the world will face 10 million deaths per year related to antimicrobial resistance. This goes hand in hand with a reduction in the GDP and costs of 100 trillion USD. Antimicrobial resistance will obstruct the treatment of infections caused by microbes. In addition, it will have a negative impact on common infectious diseases such as urinary tract infections, on transplantation procedures and chemotherapy, and even on standard operations like dental surgeries or cesarean sections [1, 2]. It is a silent pandemic with a multisectoral impact that affects not only humans, but also animals and the environment. In order to combat antimicrobial resistance, coordination among stakeholders is essential, and collaborations must be established across national borders. Surveillance must be strengthened at national and global levels through the monitoring of pathogens and the reporting of new emerging resistances. Collective policies for antibiotic treatment procedures must be implemented, and antibiotic stewardship should be improved. A ban on over-the-counter administration and prevention of antibiotic misuse is very important and public education must be encouraged. The paucity in the development of novel antibiotic drugs needs to be addressed. Pharmaceutical companies increasingly stop their research and development of new antibiotics and shift their focus to drug discovery for chronic diseases. Innovative antibiotic drugs rarely reach the market and new derivatives of existing antibiotics quickly develop resistance [2].

Various factors drive the rapid dissemination of antibiotic resistance, including environmental aspects, drug handling, patient behavior, and irrational prescription. A correlation was observed between overuse and misuse of antibiotics and the accelerating development of drug resistance. They are used to treat viral infections, are overprescribed, are applied in incorrect doses and duration times, or have been massively misused in farming and livestock feeding to prevent animals from disease and to boost growth and feed conversion rates [2].

When exposed, bacteria develop resistance to antibiotics in response to selection pressure. Resistance can arise from genetic mutations in antibiotic target genes, transporters, or regulatory genes. Alternatively, resistance determinants from commensals and environmental microbiota can be taken up through horizontal gene transfer (HGT). Bacteria can be intrinsically resistant to certain antibiotics, acquire resistance through mutations or HGT, and adapt transiently to lower

concentrations of antibiotics. The dissemination of resistance occurs primarily through human contact, but also via animal and environmental sources. Other transmission routes comprise the consumption of animals that were previously fed antibiotics or contact with fecal-contaminated water from sewage treatment plants or natural fertilizers [2].

1.1.2 The Clinical Importance of the ESKAPE Pathogen

Multiresistant superbugs emerged as healthcare-associated infections, rendering standard antibiotics ineffective. Among those are the ESKAPE pathogens, comprising of the six pathogens *Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Actinobacter baumannii*, *Pseudomonas aeruginosa* and *Streptococcus pneumoniae*. These are often resistant to multiple antibiotics due to a prolonged treatment history and include prevalent superbugs such as MRSA. They are associated with the highest number of deaths related to resistance. According to the highly cited Lancet report on the global burden of antimicrobial resistance published in 2022, the ESKAPE pathogens were accountable for 3.57 million deaths associated with resistance and 929000 deaths attributable to resistance in 2019 [2, 3].

1.2 Mechanisms of Antibiotic Actions and Resistances

Antibiotics are one of the most important medical discoveries of the 20th century. They increase the average human lifespan by serving as treatments for infectious diseases and are administered prophylactically in medical procedures to reduce the risk of infection. The first antibiotic used was salvarsan, an arsenic-based drug that was synthesized by Paul Ehrlich to treat syphilis. It was already known that bacteria could kill other microbes through secretable compounds. This led Selman Waksman to a systemic study of Actinomycetes, soil microbes that produce a range of secondary metabolites active against other microbes. Among those is streptomycin, the first antibiotic used to treat tuberculosis. The golden age of antibiotic discovery began during which most of the antibiotic classes still used today were first identified. After that, the discovery of new antibiotic classes and novel drugs gradually declined, while antimicrobial resistance increased [2, 4].

Antibiotics are assigned to different classes depending on their mode of action and their chemical structure. In general, there are four mechanisms of action: arrest of cell wall biosynthesis, protein biosynthesis, DNA replication, or interference with folic acid metabolism. As already mentioned in the previous section, bacteria develop resistances against antibiotics either through mutations in genes associated with the target of the compound or by acquisition of foreign genetic material through horizontal gene transfer. Mutations can lead to decreased uptake of the drug, increased efflux, changes in metabolism, or modification of the antibiotic target. They are often maintained only during exposure, because they place a significant burden on cellular homeostasis. Environmental antimicrobial compounds are the origin of many antibiotics used today. Microbes coexisting with these substances have developed an intrinsic resistance. It is believed to be the foundation of many resistances in clinical pathogens that were spread via HGT. Bacteria can receive foreign DNA via HGT through transformation (DNA fragments), transduction (by bacteriophages) and conjugation (physical contact between cells) [5]. In the dissemination of resistance genes, the exchange of mobile genetic elements (MGE) plays an important role. These include plasmids shared between bacteria and moving DNA elements like transposons, insertion

sequences, and gene cassettes [6].

A core set of resistance mechanisms is commonly observed in bacteria. These strategies reoccur across different types of antibiotics and include the regulation of drug uptake through porins or efflux pumps, modification of the antibiotic, target protection or bypass, and inactivation of the antibiotic [7].

The cell envelope is a natural barrier for antibiotics, as they need to enter the cell to cause action. Gram-negative bacteria have an additional outer membrane that provides them with increased protection against antibiotic attack. It contains porins, which are channels in the outer membrane that allow the influx of small hydrophilic molecules into the cell. Many antibiotics are capable of passing through these channels. In response, bacteria can develop resistance to antibiotics by changing the permeability of the cell envelope. This is mediated by alterations in the composition of the cytoplasmic membrane or by mutations in genes that encode porins. For instance, a mutation can affect the electric charge of amino acids in the channel, thus reducing the permeability of the drug [5, 7].

Another widely used strategy is efflux-mediated antibiotic export to reduce intracellular concentrations. This is an active transport process that requires energy. Efflux pumps serve a multifaceted purpose in bacteria, contributing not only to the export of toxic substances but also to quorum sensing, adaption to environmental stress, and regulation of virulence. Currently, there exist five to seven different families of efflux pumps. Among the most important ones are the resistance-nodulation-division family (RND), the ATP-binding cassette family (ABC), the major facilitator superfamily (MFS), the multidrug and toxic compound extrusion family (MATE) and the small multidrug resistance family (SMR). In *S. aureus*, the MFS efflux pump NorA represents the most well-characterized transporter [5, 7, 8].

1.2.1 β -lactams

Bacteria possess a cell wall for protection and shape, consisting of the peptidoglycan. This polymer is composed of alternating units of N-acetyl muramic acid (NAM) and N-acetyl glucosamine (NAG) that are cross-linked by amino acids. The main enzymes involved in the cross-linking of the peptidoglycan are transpeptidase and carboxypeptidase [7, 9]. These enzymes are targeted by classical β -lactams such as penicillin and cephalosporin, which exhibit a chemical structure similar to that of the enzyme substrate. β -lactam antibiotics then form a covalent bond with the enzyme, rendering it ineffective. Therefore, these enzymes are also called penicillin-binding proteins [10]. Shortly after Alexander Fleming discovered penicillin, bacteria acquired resistance by expression of β -lactamases that hydrolyze the β -lactam ring of penicillin. The inactivation of the antibiotic by degradation or enzymatic drug modification is a favorable mechanism in terms of fitness cost [2, 4, 7]. In order to target bacteria producing β -lactamase, the antibiotic methicillin was introduced. But again, methicillin-resistant *S. aureus* (MRSA) was reported, making it one of the most dominant multi-resistant pathogens [11]. It acquired resistance by the *mecA* gene that encodes a penicillin-binding protein (PBP) with weak association to penicillins. This strategy is a prominent example of target bypass through evolving new molecules with similar functions but with a lower affinity to the antibiotic [5, 7]. The glycopeptide vancomycin was initially introduced to treat MRSA, but in 2002 the first vancomycin-resistant *S. aureus* (VISA) was detected. Glycopeptides like vancomycin also target cell wall synthesis, but through a different mechanism of action. They inhibit cross-linking of NAM units in the peptidoglycan precursor

by binding to the D-alanyl-D-alanine motif in the peptide chain. Destruction of components of the cell wall eventually leads to cell lysis [7, 9].

1.2.2 Aminoglycosides

Antibiotics can arrest protein biosynthesis by stalling the ribosome. In translation, the 50S and 30S ribosomal subunits comprise the ribosome and continue to synthesize mRNA until it reaches a stop signal. Aminoglycosides such as gentamicin or streptomycin, primarily target the 16S rRNA of the 30S subunit of the bacterial ribosome. Bacteria can develop resistance against aminoglycosides through aminoglycoside-modifying enzymes that render the antibiotic ineffective. But also other strategies such as efflux-mediated export, decreased uptake and target site modification of the ribosome are employed [5, 7, 12]

1.2.3 Tetracyclines

Tetracyclines also bind to the 16S rRNA, which prevents tRNA from binding to the aminoacyl-tRNA recognition site (A site) of the ribosome. Thereby, protein synthesis is stopped, leaving cells unable to further grow and maintain cellular homeostasis [13]. Resistance is conferred by efflux pumps, decreased permeability, ribosomal mutations or protection proteins and target destruction by tetracycline inactivating enzymes [7, 13].

1.2.4 Macrolides and Lincosamides

The 23S rRNA in the 50S subunit of the ribosome can also be targeted by antibiotics. This is done by macrolides, lincosamides and streptogramins, that arrest protein synthesis through blocking the exit tunnel in the ribosome, leading to truncated polypeptides and malfunctioning cells [7, 14]. In addition to efflux-mediated resistance, ribosomal protection proteins, and mutations in the ribosome, bacteria can acquire resistance to macrolides and lincosamides through enzymatic modification. This is typically achieved by enzymatic addition of moieties, placed on the target through phosphotransferases or on the 23S rRNA by methyltransferases that catalyze the methylation of ribosomal rRNA [7].

1.2.5 Fusidic Acid

Another antibiotic that targets protein synthesis is fusidic acid, which is often used to treat staphylococcal infections. It binds to the elongation factor G (EF-G) in the ribosome and inhibits the translocation process, thereby stalling translation. The mechanism of resistance to fusidic acid demonstrates a classic example of target protection, which allows the protein to function either by mediating dissociation of the antibiotic, blocking the binding site, or in the presence of the antibiotic. The resistant determinant FusB, typically acquired by *S. aureus*, induces dissociation of the drug and rescues protein synthesis [5, 7, 15].

1.2.6 Fluoroquinolones

Quinolones and fluoroquinolones terminate replication by binding to DNA gyrase and topoisomerase IV, thereby inhibiting nucleic acid synthesis. These enzymes are involved in the structural

conversion of DNA, essential for DNA synthesis and transcription [16]. The resistance mechanism to fluoroquinolones typically involves target alteration, which is conferred through point mutations in the DNA gyrase and topoisomerase IV genes. Modification or protection of the target itself reduces drug efficacy and enables bacteria to escape the action of antibiotics [5, 7].

1.2.7 Sulfonamides and Trimethoprim

Sulfonamides and trimethoprim affect the folic acid pathway that is essential for the biosynthesis of key compounds like purines and thymidylate in DNA replication. They are often administered in combination, but inhibit different parts of this pathway. Sulfonamides target dihydropteroate synthetase, while trimethoprim inhibits the subsequent enzyme dihydrofolate reductase [7, 9, 17].

1.3 Biomedical Applications of Artificial Intelligence (AI) and Foundations of Machine Learning

1.3.1 AI in Clinical Diagnostics, AMR Surveillance and Drug Discovery

Artificial intelligence is a powerful tool with considerable potential in combating the emerging antibiotic resistance crisis. Increasing computational power and growing data sets have inspired scientists to incorporate these tools in clinical diagnostics, AMR surveillance, and antibiotic discovery [18].

Bacterial infections are typically diagnosed by antimicrobial susceptibility testing (AST), polymerase chain reaction (PCR), or mass spectrometry. Bacterial cultures take a considerable amount of time, which is a major limitation to rapid diagnostics. PCR is restricted to primers that target known resistance determinants, and mass spectrometry relies on databases that often insufficiently represent some bacterial species. Early detection is often crucial for serious bacterial infections such as sepsis, meningitis, and toxic shock syndrome. AI can help improve the diagnosis of sepsis by distinguishing it from mimicking diseases, accelerate the detection and identification of species through interpretation of spectral fingerprints and improve patient treatment outcomes through AMR prediction. This complies with antibiotic stewardship, which advocates that infections should be treated with appropriate antibiotics while ensuring sustainable use of last resort antibiotics [18]. Electronic health records (EHR) were introduced to take advantage of clinical and laboratory data that was previously overlooked in healthcare systems. However, using EHR for AI applications remains challenging, as data is often unstructured and inconsistent [18, 19].

Monitoring AMR determinants in healthcare, agriculture, and environmental settings is another important factor in the global fight against AMR. This is part of the One Health approach, because resistance emergence is complex and is not limited to healthcare only. AMR surveillance can help identify new emerging AMR determinants and expose transmission networks. The establishment of early warning systems can contribute to effective resource management and intentional reactions. Efforts in global AMR surveillance have led to increasingly available whole genome sequencing (WGS) data with assigned AST phenotypes [18–20]. Classic AMR prediction based on genomic data comprises rule-based approaches, such as the ABRicate tool, where the phenotype is classified by the presence or absence of known resistance determinants referenced from curated databases [21, 22]. AI enhances AMR surveillance primarily by identifying novel

AMR determinants through the analysis of important features on which the model bases its decisions [20].

Although antibiotic resistance in pathogens continues to increase, the pace of antibiotic discovery remains slow. The use of AI to support antimicrobial drug discovery has great potential. It is applied in AI-based virtual screenings to predict antimicrobial function from molecular structure or in generative models that explore the chemical space for new compounds with targeted characteristics. Natural products such as secondary metabolites are still a valuable reservoir that can be investigated using artificial intelligence. Metagenomic and genomic data are also used to predict antimicrobial peptides that may be helpful for drug discovery. Furthermore, the combination and repurposing of existing drugs offer opportunities for new therapeutic options [18, 19].

1.3.2 Overview on Machine Learning

Machine learning can be defined as the use of models for the prediction of an outcome variable that learned on data. In general, supervised learning, where data has labels, is distinguished from unsupervised learning, where data points come without labels. Supervised models rely on a certain ground truth that is associated with the labels, while unsupervised models aim to find patterns in the data without prior information. Classification refers to problems where the output variable belongs to discrete categories, while in regression a continuous set of variables is predicted. Clustering typically involves finding patterns and groups in the data. In order to capture the underlying relationship, a model type is selected and then subsequently parameterized in the learning process. The data is split into a data set for training and a validation data set for performance evaluation. Ideally, the final model is judged by an independent test set. Choosing a suitable model can be difficult and lead to common pitfalls such as underfitting or overfitting of the model. To avoid these extremes, it is important to manage the bias-variance trade-off. The bias describes the assumptions a model relies on to explain a particular problem, while the variance explains the variation between models trained on different randomly sampled training data sets. A model that does not fit the training data well and performs poorly is called underfitting. This corresponds to high bias and low variance. Overfitting is more dangerous as it simulates good performance in the training process but generalizes poorly to new unseen data. An overfit model had learned to identify structures in the data that are not predictive of the intended outcome variable. The model exhibits high variance and low bias [23, 24].

Features are the input that is fed to the model in training. The amount of features also has an impact on underfitting or overfitting. Too little features will result in underfitting, while too many features can lead to overfitting. Depending on the application, they require manual creation or can be extracted from data. Experience and expertise are needed to choose and reduce the features to a reasonable amount [23].

The selection of the machine learning framework depends on the application. In principle, traditional machine learning methods are distinguished from deep learning techniques. The architecture of deep learning is based on neural networks that pass data through multiple connected nodes and layers [18]. Although powerful, deep learning requires large amounts of data with clear relationships and many features. Furthermore, training is often longer and computationally more expensive for deep learning models. It is also not trivial to identify key features, which is crucial for clinical usage. Therefore, traditional machine learning methods like support vector machines

(SVM), gradient boosting or random forests are often preferred over deep learning [24]. SVMs require labeled input data and can be applied to classification or regression problems. They are a good choice when the underlying relationship of the data is nonlinear, while still being applicable for linear problems as well. They function by finding a hyperplane in high-dimensional space, that separates the data. In order to select the best classifying hyperplane, the margin separating the hyperplane from the nearest data point is maximized. Outliers are usually treated by defining a soft margin, that allows some data points to violate the hyperplane. A kernel function is a mathematical technique used to transform data from a low-dimensional into high-dimensional space, making separation of the data easier [24, 25]. Other methods frequently used are k nearest neighbors and ensemble-based decision trees like gradient boost and random forest [24]. In principle, decision trees are composed of a tree-like hierarchy, with nodes containing questions about features and child nodes with respective answers. A data point is classified by following the path from the root node to the leaf node, depending on the answers that apply. Ensembles of decision trees combine multiple single decision trees, providing increased accuracy. Random forest techniques reduce chances of misclassifications by a collection of trees, that were built randomly by replacement sampling out of the training set. The classification results are based on the most frequently predicted class. [26] In gradient boosting, multiple weak decision trees are combined into a strong classifier by repeatedly updating weights to minimize the error [27].

1.4 Architecture and Functionality of the Automated Learning Pipeline for Antimicrobial Resistance (ALPAR)

In recent years, machine learning approaches have become popular for resistance prediction, as they do not require prior knowledge of AMR genes and can help to find new resistance determinants. Most machine learning frameworks predict resistance by extracting features from whole genome sequencing data such as genes, single nucleotide polymorphisms (SNPs), nucleotide k-mers, amino acid k-mers and combinations of genes and SNPs [22, 28]. Yurtseven et al. demonstrated the importance of controlling for population effects in the machine learning process. The authors published a novel pipeline called ALPAR for *Mycobacterium tuberculosis*, that predicts antibiotic resistance by including a phylogenetic ranking of features. They were able to show an increase in the performance of AMR prediction and also discovered a set of promising resistance-associated mutations [29]. Although there are already reports on successful AMR prediction in *S. aureus* published, the ALPAR pipeline has promising aspects in feature selection and performance [30]. It can use a combination of gene content and SNPs as input features and accounts for population structure via a phylogeny-related parallelism score (PRPS) score [29].

Figure 1.1 shows an overview of the structure of ALPAR, adapted to the workflow of this thesis. The input of the pipeline requires genomic FASTA files arranged in a hierarchical folder path [31]. From these files, binary tables are created. A phylogenetic binary table is constructed that captures the resistance phenotypes for all input genomes. Prokka is used to annotate genomic sequence data, followed by a gene presence and absence analysis using the clustering tools CD-Hit or panaroo [32–34]. A custom protein database can be specified optionally. SNPs are extracted with snippy from each genome by comparison with a reference genome [35]. Genes and SNPs are then combined in a genotype binary table, and a threshold is applied to eliminate features that are only present in a few genomes. Phylogenetic trees are created using MashTree

or PanACoTA [36, 37]. For the calculation of the PRPS score, a matrix is created with the pairwise distances of the nodes in the phylogenetic tree. In the following step, the clades are collapsed whenever a feature is present in all the leaves. For every variant, the sum of pairwise distances between all branch points with this feature is calculated, and its logarithm is taken. This procedure is repeated for all features. In the machine learning training process, only the top-ranked variants with the highest PRPS score are chosen as input for the model. Although this percentage can be adjusted by the user, Yurtseven et al. determined that a threshold of 30 % is optimal. The framework includes three different machine learning algorithms: SVM, random forest and gradient boosting. Gini importance can be conducted for random forest and gradient boosting models, while permutation importance can be used for SVMs. The final outcome of the ALPAR pipeline includes trained models that can be used for prediction and a genome wide association study (GWAS) analysis conducted with pyseer to detect important AMR genes and mutations in the data [29, 31, 38].

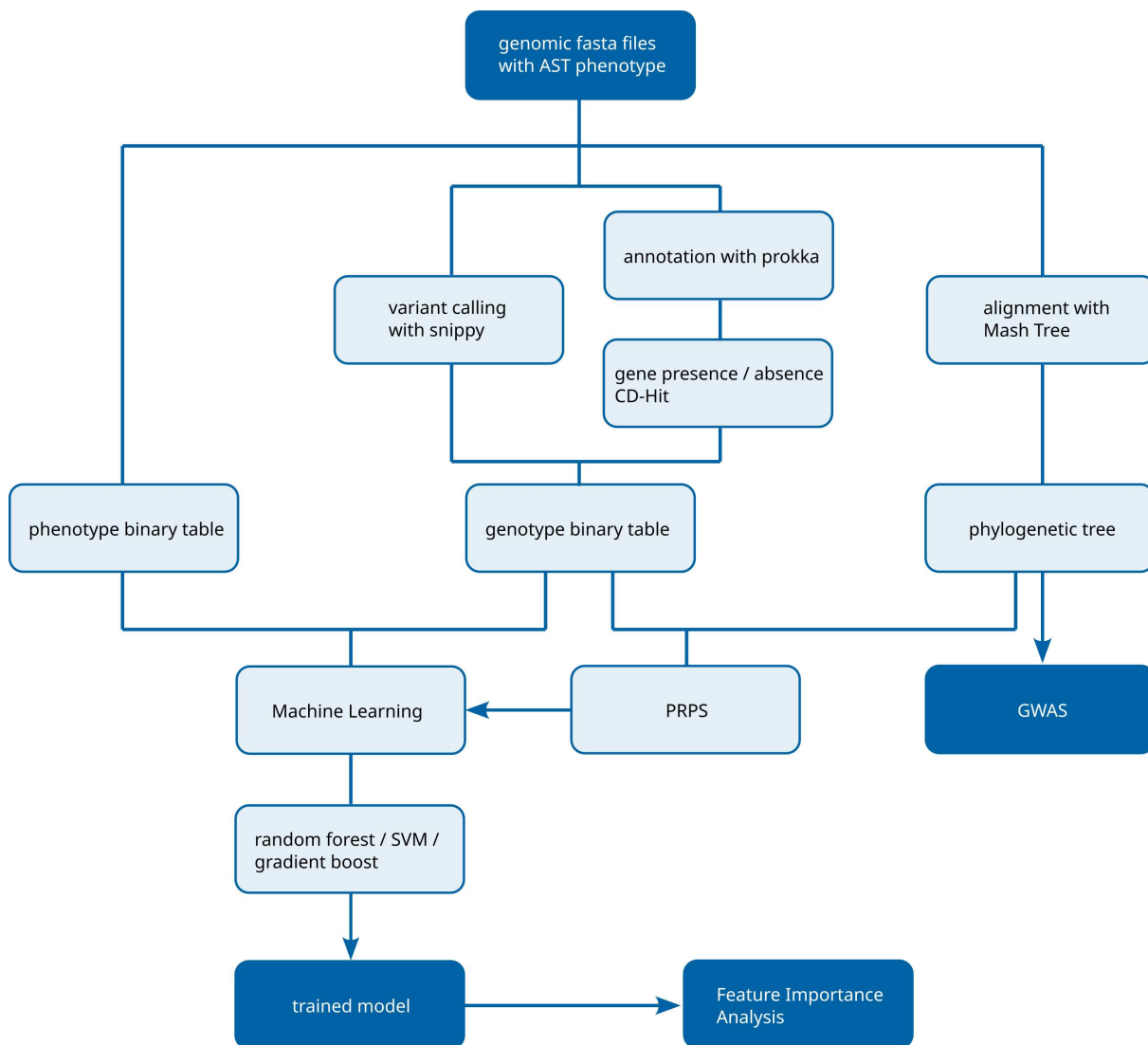


Figure 1.1: Overview on the ALPAR pipeline adapted to the workflow of this work. The figure was inspired by the flow chart on the github page of ALPAR [31].

2 | Aim of this thesis

This thesis aims to establish the newly published tool ALPAR (Automated Learning Pipeline for Antimicrobial Resistance) to predict antimicrobial resistance (AMR) for the high priority pathogen *Staphylococcus aureus*. The purpose is to demonstrate the transferability of this machine learning framework to another species. Although there are already reports on successful AMR prediction in *S. aureus* published, the ALPAR pipeline has promising aspects in feature selection and performance [30]. It can use a combination of gene content and SNPs as input features and accounts for population structure via a phylogeny-related parallelism score (PRPS) score [29]. This work will involve the selection of appropriate data despite limited availability, the training of machine learning models, and the evaluation of their performance. *S. aureus* isolates from atopic dermatitis patients will serve as a testing dataset for experimental validation. Key variants will be detected through a feature importance analysis to make reliable and generalizable predictions. Finally, an evaluation is conducted to determine whether ALPAR outperforms the benchmark tool ABRicate, along with contextualizing the intended use of the pipeline.

3 | Materials and Methods

This chapter presents a detailed explanation of the materials and methods used throughout this work.

3.1 Data Collection

Input data sets for the training and testing of machine learning models were obtained from the Biological and Viral Bioinformatics Resource Center (BV-BRC) [39]. The database was filtered for unique genomes of good quality (genome quality analysis from BV-BRC database [40]), whole-genome sequences or complete genomes and laboratory-based evidence on their respective antimicrobial resistance phenotype. Antibiotics were selected based on the availability of at least 100 genomic sequences available after filtering. FASTA files were downloaded on November 16, 2024. Corrupted genomes in the database were excluded from the data set and can be found in the table 6.1 in the supplementary section.

The reference genome ASM1342v1 (GCA_000013425.1) for *Staphylococcus aureus* (strain *Staphylococcus aureus* subsp. *aureus* NCTC 8325) was obtained from the National Center for Biotechnology Information (NCBI) on November 18, 2024 [41]. As a reference proteome, *Staphylococcus aureus* (strain NCTC 8325 / PS 47) was chosen and accessed on November 18, 2024 [42, 43].

3.2 Automated Learning Pipeline for Antimicrobial Resistance (ALPAR)

The pipeline featured in this work is the Automated Learning Pipeline for Antimicrobial Resistance (ALPAR) published by Yurtseven et al. in 2023 [29]. Version 0.7.2 was installed following the instructions on GitHub [31].

3.3 Training and Testing of the Models

The automatic pipeline was run with a provided reference file and a custom protein database (see section Data Collection). The options `--no_datasail` and `mashtree --fast` were used [36].

ALPAR annotates genomic sequences with prokka and compares them with the reference genome in variant calling using snippy [32, 35]. Genes, single nucleotide polymorphisms (SNPs), DNA insertions and deletions (INDELS) that were not identified in the reference genome were collected as features in a genotype binary table. Features that were present in less than 0.2% of all genomes were filtered out. To further reduce the number of features and to select for features that are not linked to the population structure, a phylogeny-related parallelism score (PRPS) is calculated as described in Yurtseven et al. Then 30% of the features with the highest PRPS

score are selected for training.

For each antibiotic, models were trained using three different algorithms: Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (GB). The default values of all parameters that were kept are listed in table 3.1.

parameter	value
<code>prps_percentage</code>	30%
<code>test_train_split</code>	0.20 (80% training, 20% testing)
<code>ml_algorithm</code>	Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB)
<code>random_state</code> (for train test split)	42
<code>n_estimators</code> (RF)	100
<code>min_samples_split</code> (RF)	2
<code>max_depth</code> (RF, GB)	10
<code>min_samples_leaf</code> (RF, GB)	1
<code>resampling_strategy</code>	holdout (RF, GB), cross-validation (SVM)
<code>cv</code> (number of splits for cross-validation)	4
<code>scoring</code> (for cross-validation)	Matthews Correlation Coefficient (MCC)
<code>feature_importance_analysis_strategy</code>	gini (RF, GB), permutation importance (SVM)
<code>svm_kernel</code>	linear
<code>C parameter</code>	optimized (for [1, 2]) using GridSearchCV from sklearn [44]

Table 3.1: Parameters used for training the models.

The pipeline lists a range of different validation metrics to assess the performance of the models. The following metrics were chosen for the evaluation:

- Matthews Correlation Coefficient (MCC)
- Precision
- ROC curve and ROC AUC Score
- Sensitivity (recall)
- Specificity

The specificity was calculated from the confusion matrix using the following equation.

$$Specificity = \frac{True\ Negatives}{True\ Negatives + True\ Positives}$$

Learning curves were made with sklearn for all antibiotic models with random forest, gradient boosting and support vector machine [44]. The parameters from table 3.1 were used. The parameter C of support vector machine models was optimized using gridSearchCV from sklearn [44]. For the generation of the learning curves, a 4-fold cross validation with Matthews correlation coefficient scoring was used. Training size was split into 10 parts.

A gini importance analysis was conducted for the random forest models. Nine additional models with different randomization seeds (0, 1, 9, 18, 42, 100, 222, 2000, 3497, 12345) were trained using random forest with a holdout resampling strategy. Then, it was counted how often a feature was among the top ten of the most important features. A mean importance score was calculated from the results of the ten models.

The ALPAR pipeline includes a genome wide association study (GWAS) using pyseer, that takes population structure into account [38]. Multiple testing was controlled by a Bonferroni correction. Significant variants were identified by applying a p-value threshold of 0.05 that was divided by the total number of SNPs.

All resulting features of the feature importance analysis and the GWAS study were blasted against the Comprehensive Antibiotic Resistance Database (CARD) to identify already known resistance determinants and hypothetical proteins were characterized using BLAST [45, 46].

Phylogenetic trees of the training data were created using iTOL [47]. Annotations include genome IDs and a color expressing the genome’s phenotype (resistant = red, susceptible = blue).

3.4 Antimicrobial Resistance Prediction for an Isolate Collection

The models were tested on independent data from a *Staphylococcus aureus* isolate collection consisting of 87 isolates. *S. aureus* was isolated from skin swap samples, collected from seven different patients suffering from atopic dermatitis of the Medical University of Vienna and sequenced by the Joint Microbiome Facility (JMF) using Nano Pore sequencing.

The original pipeline bases its predictions solely on single nucleotide polymorphisms and omits presence or absence of genes. In order to include this, equalization of binary mutation tables was done manually. In the first step, SNPs in the binary mutation tables of the model and the isolates were filtered to extract the names of the genes. Basic Local Alignment Search Tool (BLAST) was used to build a library of genes in the binary mutation table of the model [45]. Genes from the binary mutation table of the isolates are blasted against the model’s database to assign the respective names. In the last step, column names in the binary mutation table of the isolates are renamed after the highest hit of the BLAST output.

The prediction performance of the models for erythromycin and penicillin was evaluated in comparison to the benchmarked tool ABRicate. The isolates chosen for each phenotype subgroup were selected from different patients if possible. Five isolates per subgroup were tested, when there were enough isolates with respective phenotypes. An overview of the isolates can be obtained from table 6.2.

isolate ID	patient ID	antibiotic tested
iso01.06	p022	ERY
iso01.09	p022	ERY, PEN
iso01.23	p022	ERY
iso02.09	p022	ERY
iso05.05	p023	ERY, PEN
iso05.20	p023	PEN
iso06.07	p009	ERY, PEN
iso06.11	p009	PEN
iso07.02	p009	PEN
iso07.07	p009	ERY, PEN
iso11.12	p025	PEN
iso12.08	p026	ERY, PEN
iso17.02	p021	PEN
iso17.14	p021	ERY
iso18.01	p021	ERY, PEN
iso19.13	p029	ERY, PEN
iso20.08	p029	ERY

Table 3.2: Overview on isolates experimentally tested. (ERY = erythromycin, PEN = penicillin, R = resistant, S = susceptible)

3.5 Antibiotic Susceptibility Testing

Antibiotic susceptibility testing was performed by a Minimum Inhibitory Concentration (MIC) assay. The clinical breakpoints of $11 \text{ mg} \cdot \text{L}^{-1}$ for erythromycin and $0.1251 \text{ mg} \cdot \text{L}^{-1}$ for penicillin were based on the official reference breakpoints of the European Society of Clinical Microbiology and Infectious Diseases (EUCAST) (version 15.0) [48]. Two biological replicas were performed with triplicates in each assay.

Overnight cultures were prepared in Mueller Hinton broth (Sigma-Aldrich) and incubated overnight at 37°C with agitation. Penicillin (penicillin G sodium salt, Sigma-Aldrich) was dissolved in water and erythromycin (erythromycin, Sigma-Aldrich) in ethanol (Ethyl alcohol, pure, $\geq 99.8\%$ (GC), for molecular biology, Merck) resulting in a final concentration of 1 % v/v ethanol. The sterile antibiotic stock solutions listed in table 3.3 were prepared.

concentration / $mg \cdot L^{-1}$	
erythromycin	penicillin
88.0	1.0008
44.0	0.5004
22.0	0.2502
11.0	0.1251
5.50	0.0625
2.75	0.0313

Table 3.3: Concentrations of antibiotic stock solutions.

The optical density (OD) of the overnight cultures was measured at 600 nm and the inoculum was adjusted to an OD of 0.002. The antibiotic stock solutions were diluted 1:1 with inoculum in 96 well plates (Greiner 96 Flat Bottom Transparent Polystyrene). For each isolate, a growth control was prepared in a 1:1 dilution of inoculum and medium. A vehicle control of a 1:1 dilution of 1 % v/v ethanol in water and inoculum was used to test solvent effects. For the sterile control a 1:1 dilution of the highest antibiotic stock solution and medium was made. The 96 well plates were then covered and incubated at 37 °C for 22 hours. A plate reader (infinite 200Pro) was used to measure the optical density at 600 nm with 25 flashes and no shaking.

3.6 Comparison of the Prediction Performance for ALPAR and ABRicate

The prediction performance of all 12 models trained by ALPAR was compared to the benchmarked AMR prediction tool ABRicate (version 1.0.1) using the database AMRFinderPlus from NCBI (5386 sequences, January 14, 2025) [21, 49]. Predictions were made for all 87 isolates of the collection and then compared to ALPAR’s predictions by grouping together the results of all isolates per patient. Due to the fact that all isolates were predicted resistant to tetracycline, the reference genome was also used as an input for ABRicate with a similar outcome.

4 | Results and Discussion

4.1 Collection of Training Data for the Machine Learning Models

The amount of data and its quality are an important factor when it comes to building a training and a testing data set for machine learning. Polluted, faulty or incomplete input data can result in poor model performance and unreliable predictions [28, 50]. The ALPAR pipeline requires genomic sequence data labeled with a binary antimicrobial resistance phenotype of "resistant" or "susceptible". It is challenging to find large amounts of publicly available whole genome sequences that come with an antimicrobial resistance phenotype derived from laboratory antimicrobial susceptibility testing (AST) [51, 52]. To address this data availability problem, the BV-BRC database features a curated collection of genomes paired with AST data from several different sources, such as NCBI or the United States National Institute of Allergy and Infectious Diseases (NIAID) [52]. This database was used to construct and obtain the data used in this project (see table 4.1) [39].

Antibiotic	Pharmalogical class	Number of genomes	Genomes resistant
cefoxitin	beta lactam	1149	895 (78%)
ciprofloxacin	quinolone	2439	1111 (45%)
clindamycin	lincosamide	1080	408 (38%)
erythromycin	macrolide	2491	886 (36%)
fusidic acid	fusidane	2167	331 (15%)
gentamicin	aminoglycoside	2517	299 (12%)
methicillin	beta lactam	1925	1158 (60%)
oxacillin	beta lactam	562	93 (16%)
penicillin	beta lactam	1474	1288 (87%)
tetracycline	tetracycline	2203	364 (16%)
co-trimoxazole	antifolate	687	192 (30%)
vancomycin	glycopeptide	1357	45 (3%)

Table 4.1: Overview of the final input data set used for training and testing of the machine learning models in this work.

The size of the samples in the data set depends on different aspects, such as the purpose of prediction and the genomic diversity of the organism [20]. Highly clonal microbiota like *Mycobacterium tuberculosis* will require fewer genomes to identify key predictors than *Staphylococcus*

aureus, which generally shows greater genomic diversity and also relies on HGT to confer resistance against antibiotics [53]. The amount of data and the choice of antibiotics to examine was based solely on the access of the data, which is a common problem researchers face in predicting AMR [20].

Only complete whole-genome sequences of good quality were used for further analysis. This refers to the output of the genome quality analysis performed by the BV-BRC database. Quality is measured by completeness, contamination, coarse consistency and fine consistency [40]. High contamination or poor genome assembly quality can affect the process of finding the key features that contribute to a certain phenotype [20]. The data set was constructed using all genomes after filters were applied, including different hosts (primarily human) from different geographic backgrounds and time frames. Due to limitations in the data availability, this could potentially result in a sampling bias leading to confounding effects in the model. It would be interesting to evaluate how many different clinical sources are represented. However, there are no filterable metadata on whether it is a clinical sample. It often also lacks information about kind of disease it was isolated from. The ALPAR pipeline provides a way to control population structure in the data by calculating a phylogeny-related parallelism score (PRPS score), which is recommended when using genomic data from various backgrounds [20].

Another important aspect to consider is class imbalance, where the two different phenotype groups are represented by different amounts of samples. A model trained on an imbalanced data set will direct its predictions towards the majority class, meaning it predominantly predicts the more abundant class, resulting in a skewed prediction performance. This is reflected in the ratio of sensitivity and specificity [52]. Whole-genome sequencing data combined with an antimicrobial resistance phenotype is not only limited to data availability, but also often exhibits a strong class imbalance [20]. This is potentially due to a sampling bias introduced by deep, nonrandom sampling during disease outbreaks or sequencing after antimicrobial treatment failure. Data also reflect a geographical sampling bias due to differences in available funding and surveillance between nations [51, 52]. Strategies to overcome class imbalance include adjusting weights in training samples, under-sampling the majority class or over-sampling the minority class [20, 28]. For almost all antibiotic data sets in 4.1, class imbalance can be observed to a different extent. The percentage of resistant genomes varies between a very small amount of resistant samples for vancomycin (3%) and a very strong resistant class represented for penicillin (87%), as can be seen in table 4.1. The pipeline handles an imbalanced data set by applying weights to the classifiers, increasing the contribution of the minor class relative to the majority class.

4.2 Phylogenetic Relationships within Training Data

In order to calculate a PRPS score for each feature, a phylogenetic tree needs to be generated in advance [29]. As described in the paper by Yurtseven et al., the ALPAR pipeline constructed a maximum-likelihood phylogenetic tree for each data set which can be seen as an overview in figure 4.1 (for a details see figures 6.1 to 6.12 in the supplementary section). All phylogenetic trees display mostly short branch lengths, consistent for closely related bacterial strains of the same species. Except from vancomycin, all data sets show resistant genomes clustering in multiple different branches throughout the tree. The vancomycin dataset only comprises 45 resistant genomes (see table 4.1), which all came from the same publication on a MRSA outbreak in a

special care baby unit in the United Kingdom [54]. The close phylogenetic relationship among these isolates confirmed a local transmission event within this single care unit.

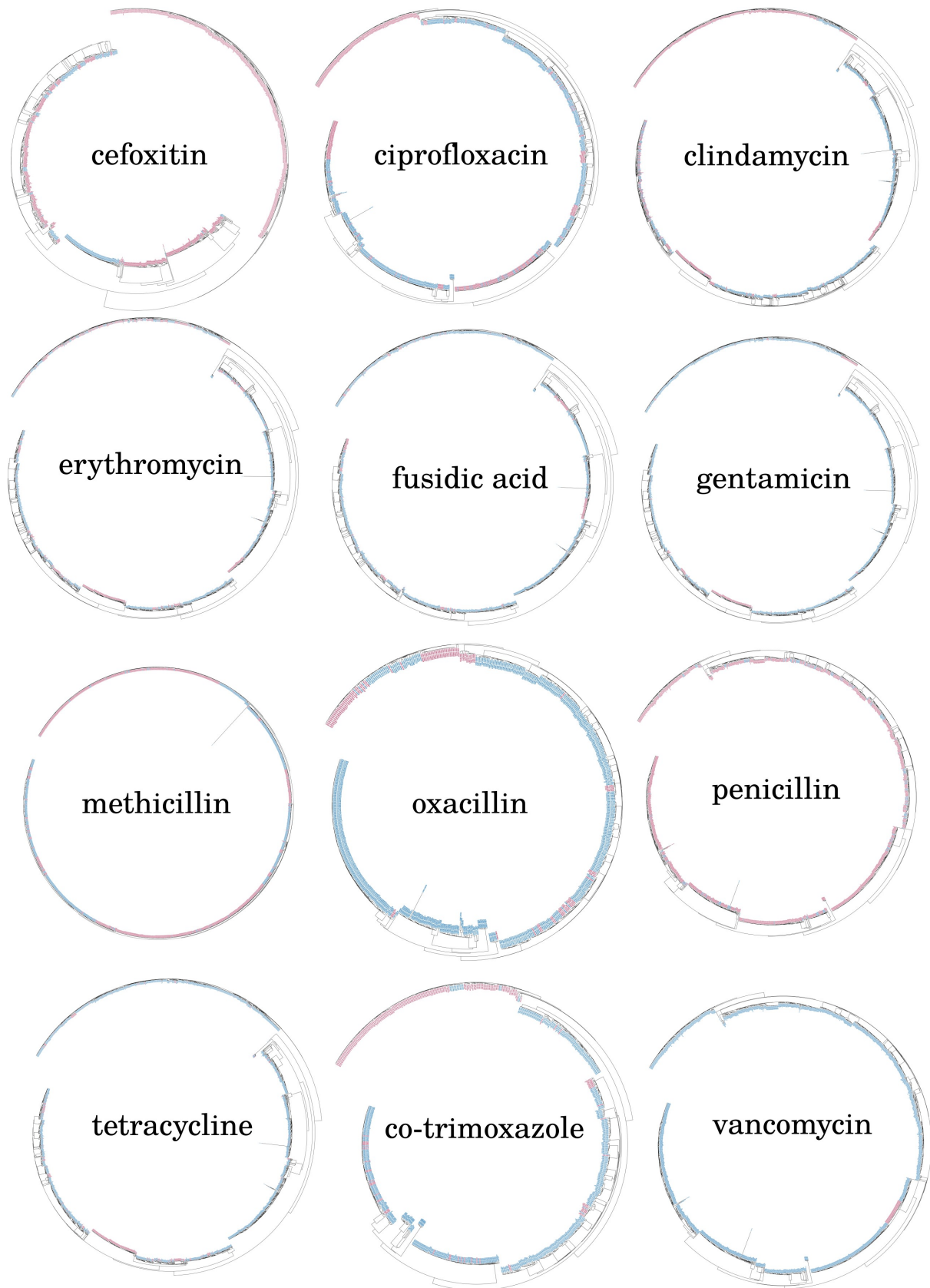


Figure 4.1: Overview of the phylogenetic trees of the training sets of the different antibiotics. Resistant genomes are colored red, susceptible ones are colored blue.

4.3 Selection of informative Features via Phylogenetic Scoring

The ALPAR pipeline creates binary feature tables, including binary information (0 = absent, 1 = present) on annotated gene presence and absence, as well as single nucleotide polymorphisms. In variant calling, mutations and genes relative to a reference genome are retained. In a first feature filtering step, variants that are present in less than 0.2% of all genomes are dropped. In a second step, a PRPS score is calculated for each feature to account for linkage disequilibrium and population structure. This is done by computing the pairwise distances between all nodes of collapsed clades with a specific SNP. The assigned score reflects how monophyletic or polyphyletic a SNP is. A high PRPS score indicates an independent acquisition of the SNP by different lineages of the tree, which is associated with resistance [29]. 30% of the features with the highest scores are selected. The exact number of features reduced by each filtering step can be obtained from table 4.2.

antibiotic	number of features		
	after variant calling	after 0.2%	after 30% PRPS
cefoxitin	290598	173946	52184
ciprofloxacin	401434	189636	56891
clindamycin	299492	192281	57685
erythromycin	423134	197082	59125
fusidicacid	407101	195614	58685
gentamicin	423387	189703	56911
methicillin	374852	161027	48309
oxacillin	266509	193097	57930
penicillin	373056	208127	62439
co-trimoxazole	238735	183656	55097
tetracycline	394885	189478	56844
vancomycin	357114	199255	59777

Table 4.2: Feature reduction by filtering and PRPS scoring.

4.4 Model Training Procedure and Performance Validation

The ALPAR pipeline comes with a set of three different algorithms: random forest, gradient boosting, and support vector machine. In general, it is recommended to evaluate multiple different machine learning algorithms for their antimicrobial resistance prediction performance, before making an informed choice [20, 51].

While random forest and gradient boosting are decision tree classifiers, SVMs approach classification differently. They find a separating hyperplane by using a kernel function that projects the data from a lower dimension into a high-dimensional space to partition it [20, 25]. The interpretability of machine learning models through a transparent prediction process is an important factor in clinical healthcare applications where the outcome can impact human lives [20, 55]. Support vector machines are typical black-box-based models, meaning they are mathematically hard to explain and their prediction flow is not traceable [56]. Unless a linear kernel is used, it is also not possible to evaluate important features in the prediction process [20]. But black-box models are not necessarily a bad choice. There are practical applications such as the use of deep learning in mammography, where the model is difficult to understand, but its output is explainable by experts [57]. The reason why the results are often not explainable is mostly due to a transformation of the model’s output data into something very different [56]. In this project, the output of the SVM is clearly understandable in the form of a classification between resistance and susceptibility. However, the potential use of such a model in the clinical decision-making process would require a fully interpretable model. When ALPAR is used for research purposes, SVMs are still a valuable choice as an algorithm, and were therefore trained and evaluated.

The most simple and widely used white-box algorithm are decision trees that feature not only accurate predictions but are also understandable in terms of their traceability and explainability. Decision trees are constructed in a hierarchical tree structure, with a top root node branching into internal child nodes and finally resulting in leaf nodes. Every node poses a question about a certain feature, sorting the data into two groups with different answers (yes / no), which again make up nodes. The final associated class is decided by the leaf node which the path reaches [26, 56]. Ensemble decision trees, which are collections of multiple decision trees, are among the best performing machine learning algorithms [58]. Random forest and gradient boosting are such approaches, where a set of different trees is built [26]. Due to their explanatory and understandable fashion, they are great options for the prediction of clinical antimicrobial resistance.

For each antibiotic, the three algorithms were used for training. The input data set was split into training and validation sets using different strategies in order to assess the performance of the model after training. For the support vector machine models, a four-fold cross-validation technique was used. For models trained with random forest and gradient boosting algorithms, a holdout validation set was randomly constructed using a test-train split of 0.2 %. For the assessment of model performance, different metrics were evaluated. Accuracy is a widely known evaluation metric for classifier performance. It measures how many correct predictions are made by the model. But it can be misleading, when the data set is imbalanced, as it could overestimate the performance of a model that predicts the majority class over and over [59, 60].

The Area under the ROC curve (AUROC) is a metric that can be used to compare the performance of models between each other. The ROC curve shows the true positive rate (TPR)

as a function of the false positive rate (FPR). A classifier making random decisions for predictions will have an AUROC of 0.5, while a perfect model will have an AUROC of 1.0 [60]. Generally, the AUROC scores of all models are very high (see table 4.3), with values between 0.842 for penicillin with a support vector machine and 1.0 for vancomycin regardless of the algorithm used. The AUROC can be a misleading metric, when looking at classifiers trained on imbalanced datasets. The model may be doing a good job in predicting susceptible bacteria resulting in a high false positive rate, but struggles to predict whether a bacterium is resistant resulting in a low true positive rate [60]. This could be the case for vancomycin, as it has the highest class imbalance with only 3 % resistant genomes and the best model performance when comparing the AUROC.

ROC AUC (area under the curve)				
Antibiotic	Support Vector Machine	Random Forest	Gradient Boosting	
cefoxitin	0.955	0.967	0.944	
ciprofloxacin	0.964	0.989	0.989	
clindamycin	0.948	0.973	0.977	
erythromycin	0.939	0.976	0.973	
fusidic acid	0.905	0.970	0.969	
gentamicin	0.934	0.985	0.978	
methicillin	0.987	0.990	0.990	
oxacillin	0.873	0.965	0.921	
penicillin	0.842	0.857	0.861	
tetracycline	0.954	0.997	0.980	
co-trimoxazol	0.910	0.939	0.866	
vancomycin	1.0	1.0	1.0	

Table 4.3: ROC AUC (area under the curve) Score for different machine learning algorithms for all trained models.

For binary classifications with imbalanced data sets, it is recommended to look at the confusion matrix, as it displays the performance of the different classes individually by counting the prediction outcomes [60]. It visualizes the relationship between the actual class labels and the predicted outcomes [59]. From the confusion matrix, the ALPAR pipeline directly calculates a range of evaluation metrics for each model.

The precision scores (see table 4.4) are defined as the number of true resistant predictions divided by the total number of resistant predictions. The recall or sensitivity (see table 4.5) of the model quantifies the number of true resistant predictions out of all actual resistant genomes. The specificity values reflect the proportion of correctly predicted susceptible genomes out of all actually susceptible genomes (see table 4.6) [59, 60]. These are especially meaningful to evaluate when the data set is imbalanced [20].

Precision			
Antibiotic	Support Vector Machine	Random Forest	Gradient Boosting
cefoxitin	0.979	0.984	0.973
ciprofloxacin	0.982	1.0	1.0
clindamycin	0.933	0.988	0.966
erythromycin	0.913	0.972	0.972
fusidic acid	0.932	1.0	0.984
gentamicin	1.0	1.0	1.0
methicillin	0.987	0.991	0.991
oxacillin	0.909	0.812	0.774
penicillin	0.967	0.970	0.970
tetracycline	0.974	0.976	0.987
cotrimoxazol	0.905	0.891	0.854
vancomycin	1.0	1.0	1.0

Table 4.4: Precision for different machine learning algorithms for all trained models.

Sensitivity (recall, true positive rate)			
Antibiotic	Support Vector Machine	Random Forest	Gradient Boosting
cefoxitin	1.0	1.0	1.0
ciprofloxacin	0.942	0.978	0.978
clindamycin	0.943	0.954	0.977
erythromycin	0.929	0.967	0.961
fusidic acid	0.821	0.940	0.940
gentamicin	0.868	0.971	0.956
methicillin	0.991	0.991	0.991
oxacillin	0.769	1.0	0.923
penicillin	0.985	0.981	0.989
tetracycline	0.915	1.0	0.963
co-trimoxazol	0.864	0.932	0.795
vancomycin	1.0	1.0	1.0

Table 4.5: Sensitivity (recall score, true positive rate) for different machine learning algorithms for all trained models.

Specificity (true negative rate)				
Antibiotic	Support Vector Machine	Random Forest	Gradient Boosting	
cefoxitin	0.911	0.933	0.889	
ciprofloxacin	0.985	1.0	1.0	
clindamycin	0.953	0.992	0.977	
erythromycin	0.949	0.984	0.984	
fusidic acid	0.989	1.0	0.997	
gentamicin	1.0	1.0	1.0	
methicillin	0.983	0.988	0.988	
oxacillin	0.977	0.931	0.919	
penicillin	0.700	0.733	0.733	
tetracycline	0.994	0.994	0.997	
co-trimoxazol	0.957	0.947	0.936	
vancomycin	1.0	1.0	1.0	

Table 4.6: Specificity (true negative rate) for different machine learning algorithms for all trained models.

The Matthews Correlation Coefficient (MCC) is considered a robust evaluation metric in cases of classification with an imbalanced data set. It takes into account all four cases of the confusion matrix and is calculated as follows:

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TN + FP) \cdot (TN + FN)}}$$

Calculation of the Matthews Correlation Coefficient. (TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives)

If a classifier makes random decisions, the MCC is 0, for a perfect model, it approaches +1, and for a model that predicts completely incorrectly, it approaches -1. Due to the imbalanced data sets in this project, the Matthews Correlation Coefficient was the most suitable metric and was therefore selected for model evaluation. Performance was generally very high, with values between 0.743 and 1.0, indicating well-predicting models (see table 4.7). The resulting vancomycin model with an MCC of 1.0 should be treated carefully, as this data set has the highest imbalance of only 3% resistant genomes. It also could be the case that the data on the resistant genomes is not very genetically diverse and only reflects a specific resistance mechanism that can be clearly distinguished from susceptible genomes.

Matthews Correlation Coefficient			
Antibiotic	Support Vector Machine	Random Forest	Gradient Boosting
cefoxitin	0.944	0.958	0.930
ciprofloxacin	0.930	0.980	0.980
clindamycin	0.895	0.952	0.952
erythromycin	0.875	0.952	0.948
fusidic acid	0.854	0.964	0.955
gentamicin	0.922	0.983	0.974
methicillin	0.975	0.980	0.980
oxacillin	0.793	0.870	0.795
penicillin	0.743	0.749	0.784
tetracycline	0.932	0.985	0.970
co-trimoxazol	0.832	0.869	0.746
vancomycin	1.0	1.0	1.0

Table 4.7: Matthews Correlation Coefficient for different machine learning algorithms for all trained models.

Following the evaluation results of the final model, learning curves show the generalization performance of classifiers with respect to the number of training examples. This is done through plotting the training and validation error against the size of the training set. They give insight into sample size determination, meaning they help researchers decide whether sample sizes should be increased for a specific performance or whether they can be decreased to save computational resources. Learning curves usually exhibit power-law or exponential behavior with smooth, monotone curves. Shapes of ill-behaved learning curves that display different behaviors like peaking or dipping can indicate model instability or misspecification of the classification problem. Phase transitions are characterized by abrupt, nearly discontinuous changes in the learning curve [61]. As the number of training examples increases, the training and testing errors are expected to decrease, gradually approaching each other and reducing the difference between the two error rates [62]. Additionally, one can estimate overfitting or underfitting of the models to the data by qualitatively looking at the learning curves. A model with high variance and low bias is too complex, fitting the training data too well. This is called overfitting, where the model learns to adapt to the noise in the data set and does not represent the true relationship of the underlying data set. It is not able to generalize to new, unseen data because it has learned patterns that are not representative of the whole population. The opposite behavior is underfitting, where the model is not complex enough to fit the training data well enough. It oversimplifies the data pattern with high bias and low variance. Furthermore, it generalizes poorly to new population data [63, 64]. Except for vancomycin, all curves for gradient boosting and SVM (figures 4.2 and 4.3) show a decreasing validation error with increasing training set size, indicating that the models generalize better with a larger data set. When looking at the learning curves from the gradient boosting algorithm, the antibiotics ciprofloxacin, clindamycin, oxacillin, tetracycline and co-trimoxazole show a small increase of the validation error with higher sample size. This indicates overfitting, because the model starts to adapt to noise in the data. Here, it could

help to stop training earlier with a smaller training set size or reduce complexity [24]. For some antibiotics, the learning curves failed for smaller training subsets. Since a 4-fold cross-validation was used and data is very imbalanced for some antibiotics, it could be the case that the folds are just too small for the smaller training subsets. Only some of the learning curves show a power-law or exponential behavior. The training error is very low for all models in figure 4.2 and 4.3, while the validation error is still very high, indicating that the models are overfit and do not generalize well.

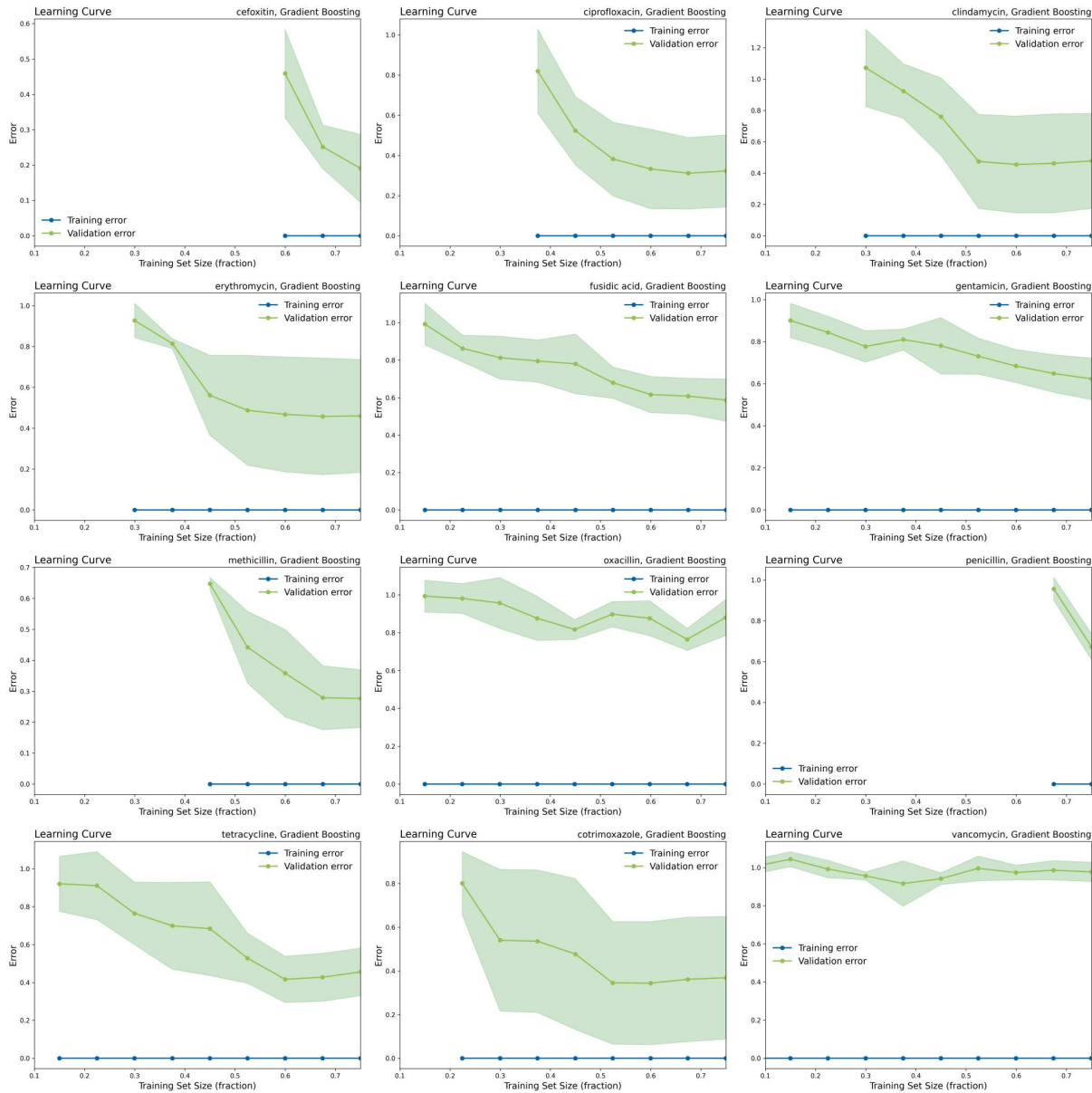


Figure 4.2: Learning curves for all antibiotic models with gradient boosting.

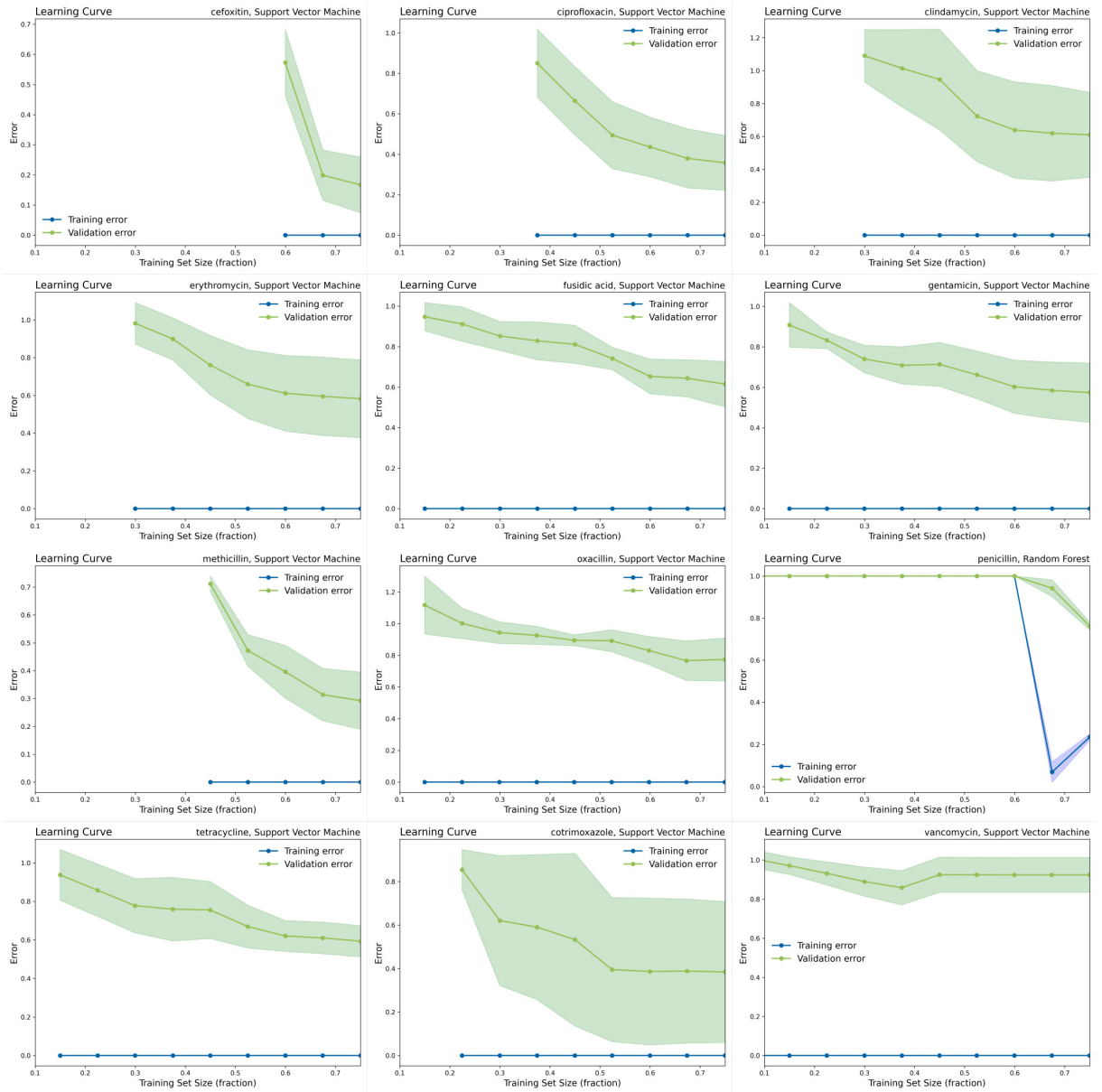


Figure 4.3: Learning curves for all antibiotic models with support vector machine.

The learning curves for the random forest models in figure 4.4 indicate an overall better performance than for SVM or gradient boosting. The training error is very high at the beginning, showing an underfit model that does not describe the underlying pattern well. With larger data subsets, it drops drastically, almost like in a phase-transition. For some models, the training error then gets higher again, but with a steadily decreasing validation error. The models learn patterns better and decrease overfitting at the cost of the training error, which in turn benefits generalization to new data. For most of the models, the difference between training and validation error is not very high in the end, indicating a well-performing model. Therefore, further analysis was constricted to random forest models.

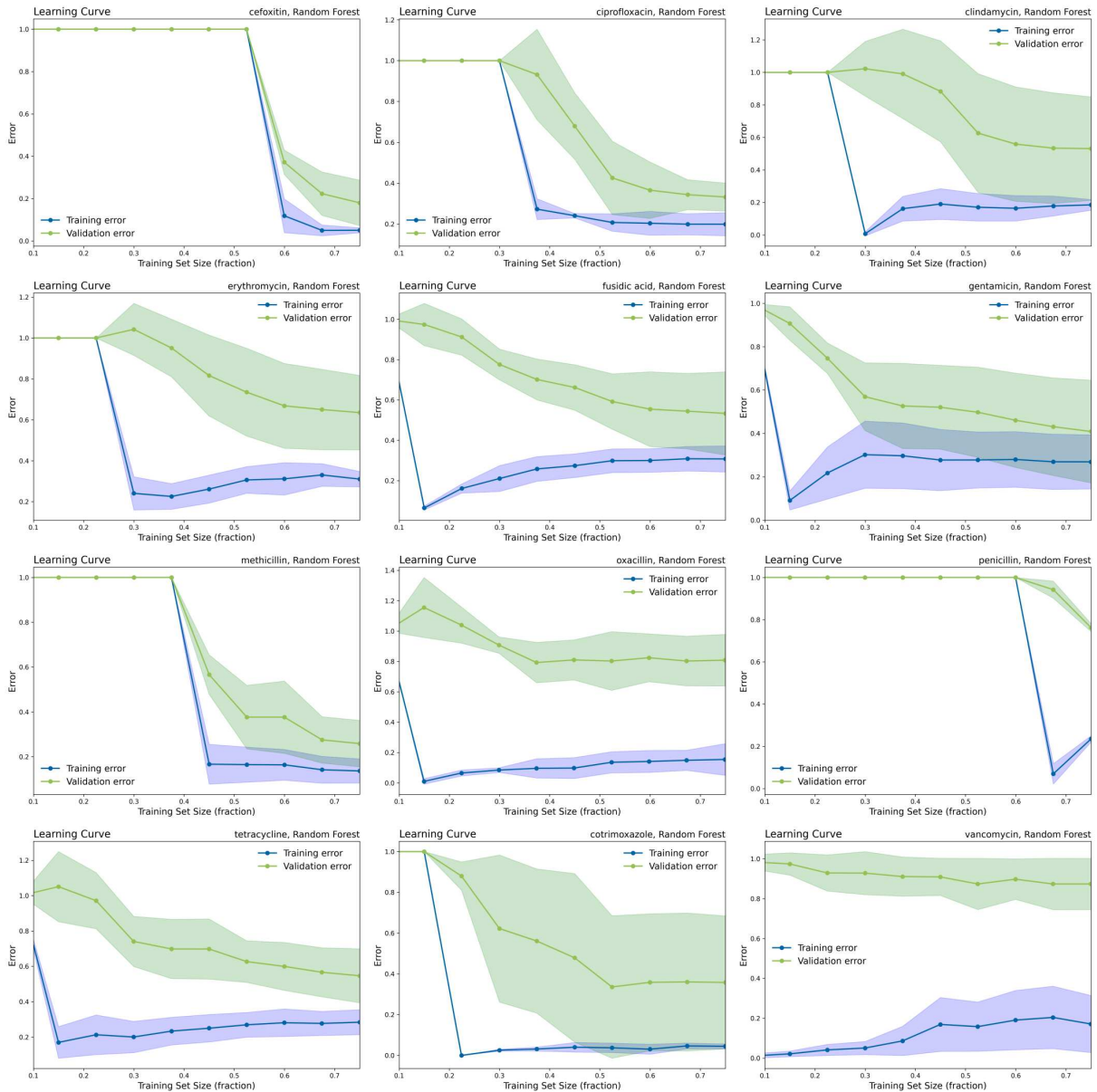


Figure 4.4: Learning curves for all antibiotic models with random forest.

4.5 Assessment of Feature Importance in Random Forest Models

A feature importance analysis was performed to study which features (SNPs, genes) have the greatest impact on the prediction process of the random forest models. It includes how often a certain feature occurred among the top 10 features in the importance analysis. In the feature importance ranking, the Gini coefficient is a number that scores how much a feature contributes to the classification tree [65]. A mean Gini importance coefficient was calculated from all Gini coefficients of the trained models where the feature was among the top 10 features. Examining the most relevant features for the model is an important part in verifying how generalizable and reliable a model really is. Highly ranked features can also give insight into new potential antibiotic resistance determinants or evolving resistance mechanisms [66]. Therefore, a feature importance analysis of machine learning models that predict AMR has great potential for antibiotic resistance research.

4.5.1 Penicillin

In the subsequent section, the results of the feature importance analysis for penicillin are discussed (see table 4.8). Penicillin resistance is typically conferred by a β -lactamase, which was found as the most important feature in the model. It is an inducible extracellular enzyme that cleaves the β -lactam ring of penicillin, thus inactivating the antibiotic [67]. The highest-ranked variant for the penicillin model is the class A β -lactamase PC1-1 (100% identity). The Beta-Lactamase Database (BLDB) was used to determine the exact type of β -lactamase, as there are many different classes and structures [68]. The second highest-ranked hit in the feature importance analysis is the gene encoding the penicillinase repressor. The penicillinase of *S. aureus* is an inducible enzyme that needs the repressor protein BlaI to express β -lactamase [69]. Penicillin resistance is often transmitted via β -lactamase plasmids that can also carry resistance genes against heavy metals such as cadmium [67]. Two proteins associated with cadmium resistance were found among the top hits. These variants were probably co-localized on the same plasmid, which justifies their presence among the results. The Hin DNA invertase is a site-specific recombinase that mediates DNA inversion through recognition of inverted repeat sequences. This type of regulated phase variation between an ON and OFF state can be found in host-associated bacterial species and near antibiotic resistance genes [70]. According to Rowland and Dyke and Lyon and Skurray, plasmids containing the *bla* operon contain flanking inverted repeat sequences that are associated with the Hin DNA invertase, which explains its importance in the model. Another promising feature is the ABC transporter ATP-binding protein, which resembles the VRA-F protein by 38.342 % [46]. According to Cao et al., this is an ABC transporter ATP-binding protein that confers resistance to vancomycin in *S. aureus*. This transporter is likely to function as an export system for the antibiotic itself or to contribute to virulence [74]. A BLAST search identified a hypothetical protein that was 4 times among the top 10 most important features in the penicillin model as the bacteriocin lactococcin 972. This points to horizontally acquired virulence genes against lactococci, as there is no known connection between the possession of lactococcin 972 genes and penicillin resistance [75]. Feature analysis of the most important features holds great potential for finding new antibiotic resistance determinants or mechanisms. For instance, the YxeA family protein is part of the *yxdLMyxeA* operon, which might be involved in the resistance against an unknown group of antibiotics [76]. The Yold-like family protein is assumed to be a type V DNA polymerase that repairs damaged DNA as part of the SOS stress response. Infection of the host, as well as exposure to antibiotics can trigger this pathway, indirectly linking it to antibiotic resistance [77]. For the last feature, the ISNCY family transposase, there was no literature found in connection to antibiotic resistance in *S. aureus*.

variant	variant type	occurrence (out of 10)	mean gini coefficient
Beta lactamase	annotated gene	8	0.02592
Penicillinase repressor	annotated gene	8	0.10011
DNA invertase hin	annotated gene	8	0.00919
ABC transporter ATP binding protein putative	annotated gene	7	0.00589
Cadmium resistance transcriptional regulatory protein CadC	annotated gene	6	0.00733
Cadmium resistance protein [Staphylococcus aureus]*	annotated gene	5	0.0054
YxeA family protein [Staphylococcus aureus]*	annotated gene	4	0.00728
lactococcin 972 family bacteriocin [Staphylococcus aureus]*	annotated gene	4	0.00924
YolD-like family protein [Staphylococcus aureus]*	annotated gene	4	0.01409
ISNCY family transposase ISBli29	annotated gene	4	0.0085

Table 4.8: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic penicillin. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

4.5.2 Other β -lactams (methicillin, cefoxitin, oxacillin)

The following section will discuss the results of the feature importance analysis for the β -lactamase-resistant β -lactams methicillin, oxacillin and the cephalosporin cefoxitin [78]. Resistance against these β -lactams is conferred through the expression of PBP2a, a low-affinity penicillin binding protein (PBP) that β -lactams cannot bind. It is encoded by the *mecA* gene, which is part of the *mec* operon together with the regulatory genes *mecI* and *mecR1*. This operon is typically part of the staphylococcal cassette chromosome *mec* (SCC*mec*) complex. Another element of the SCC*mec* are the chromosome cassette recombinases (*ccr*), that integrate the complex into the genome at a specific site. Depending on the exact composition of the SCC*mec*, multiple different types have been reported [67, 69].

For all three antibiotic models, the characteristic *mecA* gene was found among the most important variants (tables 4.9, 4.10 and 4.11). Other prominent parts of the SCC*mec* like the *mecR1* gene and recombinases such as the CcrA recombinase were also identified. Sabat et al. describe a new SCC*mec* subtype that includes a glycerophosphoryl diester phosphodiesterase and a MaoC family dehydratase, both among the highest ranked features in the methicillin

and oxacillin model. In the cefoxitin model, only the MaoC family dehydratase was found in the results. A paper published in 1991, already reported a gene product in the SCC*mec* similar to the *sn*-glycerophosphoryl diester phosphodiesterase in *Escherichia coli* [80]. Another variant deemed important across all models is the HMG CoA synthase. This enzyme plays an essential role in the mevalonate pathway of the peptidoglycan synthesis in *S. aureus* and is vital for growth, but little is known about a connection between this synthase and the SCC*mec* [81]. Zong explored the genomic region around the *mecA* gene of a clinical *Staphylococcus haemolyticus* isolate that did not contain *ccr* genes. In their work, HMG CoA synthase is found near the *mecA* gene. Due to the fact that this enzyme is one of the most important features in all three antibiotic models and is found near the SCC*mec* locus, this suggests that the SCC*mec* probably also contains metabolic genes too. It may not have a direct impact on antibiotic resistance, but it would be interesting to further explore this connection. Insertion sequences like IS431, which was identified in the feature analysis of the methicillin model, can lead to constitutive expression of *mecA* [67]. This variant was therefore correctly associated with resistance against β -lactams. Liu et al. also identified IS6 transposase as an important feature for their methicillin machine learning model.

Many SCC*mec* related features were found across all three β -lactam models, but the following paragraphs will focus on variants unique to each specific antibiotic model. Restriction-modification systems regulate the acquisition of foreign DNA and are ubiquitous in all *S. aureus* strains. Type 1 restriction-modification system endonucleases, a variant found in the results for methicillin, are said to block horizontal gene transfer in and between different lineages of *S. aureus*. Especially for MRSA and vancomycin resistant *S. aureus* (VRSA) this is typical [83]. While this system is important for the dissemination of mobile genetic elements, it is not directly linked to β -lactam resistance. A BLAST search for the hypothetical protein in the methicillin model as well as the oxacillin model resulted in a sequence alignment with the MrcB family-domain containing protein, that encodes for PBP1B in *Escherichia coli* [45, 84]. This protein is associated with resistance to the β -lactam mecillinam, but is misleading in the context of resistance prediction for methicillin and oxacillin [84].

Virulence genes such as the serine protease SplA or the missense mutation in the signal transduction protein TRAP are also among the most important features of the cefoxitin model [85, 86]. Acquisition and regulation of antibiotic resistance are often intertwined with increased virulence, which explains why virulence genes are often considered important [87]. Intergenic SNPs or other noncoding regulatory sequences can also play a potential role in resistance acquisition [88]. There is no known connection between the RES domain-containing protein and resistance to cefoxitin, making this a spurious feature.

For the oxacillin model, the DUF1643 domain-containing protein and the hypothetical protein that aligns with a protein from *Staphylococcus epidermidis* are uncharacterized proteins. Here, more experimental research is needed to investigate a potential relationship with antibiotic resistance. The synonymous mutation in the gene encoding fructose-bisphosphate aldolase, which is involved in virulence, is again a spurious feature [89].

variant	variant type	occurence (out of 10)	mean gini coefficient
MaoC family dehydratase [Staphylococcus aureus]*	annotated gene	10	0.02455
HMG CoA synthase putative	annotated gene	9	0.02058
PBP2a family beta lactam resistant peptidoglycan transpeptidase MecA	annotated gene	9	0.02493
glycerophosphoryl diester phosphodiesterase [Staphylococcus aureus]*	annotated gene	9	0.0224
recombinase family protein [Staphylococcus aureus]*	annotated gene	8	0.01104
beta lactam sensor signal transducer MecR1	annotated gene	7	0.01346
IS6 family transposase IS431mec	annotated gene	7	0.01468
MrcB family domain-containing protein [Staphylococcus aureus]*	annotated gene	6	0.00634
type I restriction enzyme endonuclease domain-containing protein, partial [Staphylococcus aureus]*	annotated gene	6	0.01108
recombinase family protein [Staphylococcus aureus]*	annotated gene	6	0.01

Table 4.9: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic methicillin. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

variant	variant type	occurence (out of 10)	mean gini coefficient
PBP2a family beta lactam resistant peptidoglycan transpeptidase MecA	annotated gene	6	0.00686
MaoC family dehydratase [Staphylococcus aureus]*	annotated gene	5	0.00513
HMG CoA synthase putative	annotated gene	4	0.00564
serine protease SplA	snp: missense	4	0.0034
intergenic	snp: intergenic	4	0.00409
intergenic	snp: intergenic	4	0.00207
intergenic	snp: intergenic	4	0.00583
signal transduction protein TRAP [Staphylococcus aureus]*	snp: missense	4	0.00215
intergenic	snp: intergenic	3	0.00338
RES family NAD+ phosphorylase [Staphylococcus aureus]*	snp: missense	3	0.00257

Table 4.10: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic cefoxitin. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

variant	variant type	occurence (out of 10)	mean gini coefficient
PBP2a family beta lactam resistant peptidoglycan transpeptidase MecA	annotated gene	9	0.02723
MaoC family dehydratase [Staphylococcus aureus]*	annotated gene	6	0.03524
MrcB family domain-containing protein [Staphylococcus aureus]*	annotated gene	6	0.02735
glycerophosphoryl diester phosphodiesterase [Staphylococcus aureus]*	annotated gene	6	0.03213
HMG CoA synthase putative	annotated gene	5	0.02756
intergenic	snp: intergenic	4	0.00275
cassette chromosome recombinase CcrA [Staphylococcus aureus]*	annotated gene	4	0.00299
DUF1643 domain-containing protein [Staphylococcus aureus]*	annotated gene	4	0.00986
class IIb fructose-bisphosphate aldolase FdaB [Staphylococcus aureus]*	snp: synonymous	3	0.00306
conserved hypothetical protein [Staphylococcus epidermidis ATCC 12228]*	annotated gene	2	0.02089

Table 4.11: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic oxacillin. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

4.5.3 Tetracycline

Table 4.12 shows the most important features of the tetracycline model. This antibiotic arrests protein synthesis by binding to the 30S subunit of the bacterial ribosome. Known AMR determinants for tetracycline include *tetA(K)* and *tetA(M)* [67]. As expected, genes encoding the efflux

pump *tetA(K)* and the ribosomal protection protein *tetA(M)* were found among the important features. The hypothetical protein CGP98 00055 appears to be an interesting candidate involved in antibiotic resistance, as it is considered important in all ten tetracycline models. Although the mean gini coefficient is higher for *tetA(K)*, this hypothetical protein occurred more frequently in the top ten features of the models. Interestingly, Liu et al. also found a hypothetical protein (SAOUHSC_02693) with 33 amino acids among the feature importance results for their tetracycline model. While both have the exact same length, their sequence differs. Among the results are also features that are not resistance determinants, but related to the dissemination of resistance in general. Relaxases of the MobV family are metal-dependent nucleases that break and join DNA in horizontal gene transfer, which is directly linked to the spread of antibiotic resistance [90]. Replication initiation proteins are vital for plasmid progression and are therefore also linked to evolving resistances through mobile genetic elements. A BLAST search identified both replication initiation proteins as the *rep* gene on pCW7 [45]. This plasmid is part of the pT181 family, antibiotic resistance plasmids typically associated with resistance to tetracycline [67, 91, 92]. One way bacteria are able to react to environmental stresses such as antibiotic exposure is by extracytoplasmic sigma factors (ECF), that activate specific target genes [93]. The RNA polymerase sigma factor found in the results was characterized as an ECF using BLAST [45]. In response to stress, bacteria can change their transcriptome. Chu et al. reported multiple proteins containing helix-turn-helix domains (HTH) that are involved in the transcriptome modulation of *S. aureus* and states that their impact on gene expression is underestimated. More knowledge of HTH-domain-containing proteins could provide insight into alternative ways in which bacteria can survive under stress, such as exposure to antibiotics. This could then explain their importance in the model. Another feature found among the results is a signal peptidases of type 1 (*spsA* and *spsB*), which is an attractive antibiotic target for novel antibiotic drugs, because SpsB is vital for the processing of peptides in *S. aureus* [95]. SpsA is a catalytically inactive signal peptidase upstream of SpsB. With the continuous development of drugs against SpsB, it could be possible that gain-of-function mutations in SpsA could lead to its activation and potentially be linked to resistance [96].

variant		variant type	occurrence (out of 10)	mean gini coefficient
hypothetical	protein	annotated gene	10	0.03161
CGP98 00055 [Staphylococcus aureus]*				
tetracycline efflux	MFS	annotated gene	8	0.0666
transporter Tet K				
MobV family	relaxase	annotated gene	7	0.03554
[Staphylococcus aureus]*				
Replication	initiation	annotated gene	7	0.03395
protein				
RNA polymerase	sigma	annotated gene	6	0.01557
factor [Staphylococcus aureus]*				
tetracycline	resistance	annotated gene	5	0.04528
ribosomal protection				
protein Tet(M) [Staphylococcus aureus]*				
Signal peptidase IA,		snp: synonymous	4	0.00318
putative				
Replication	initiation	annotated gene	4	0.01555
protein				
helix-turn-helix		annotated gene	4	0.01835
domain-containing				
protein [Staphylococcus aureus]*				
hypothetical	protein	snp: other	3	0.00272
[Staphylococcus aureus]*				

Table 4.12: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic tetracycline. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

4.5.4 Gentamicin

Resistance to gentamicin is mediated by different strategies. The most common mechanism is modification of the antibiotic itself through enzymes such as aminoglycoside acetyltransferase (AAC), aminoglycoside adenylyltransferase (AAD) and aminoglycoside phosphotransferase (APH). Bacteria can also alter their membrane that interferes with antibiotic uptake or modify the ribosomal structure through mutations [67]. Among the most important features listed is a bifunctional enzyme consisting of two enzymatic activities, as well as the aminoglycoside acetyltransferase of the GNAT family [97]. Using the CARD database, the subunits of the bifunctional enzyme were aligned to the AAC(6')-Ie and to APH(2'')-Ia [46]. The hypothetical proteins could also be a potential aminoglycoside-modifying enzyme and should be further studied. In staphylococci,

the insertion sequence IS256 is also associated with resistance to gentamicin and therefore a plausible variant found by the pipeline [6]. It is interesting that the HTH-domain-containing protein is considered important not only in the tetracycline model but also in the gentamicin model. Since both antibiotics target the same subunit of the ribosome, this could indeed be a promising candidate for a new resistance mechanism. The co-occurrence of heavy metal resistance with antibiotic resistance determinants is well-established. Resistance to mercury is encoded by the *mer* operon which is often found on multi-resistance plasmids such as pI258. However, these are typically associated with the dissemination of β -lactamase genes [6]. There is no known connection between the found variant dihydrolipoamide dehydrogenase and antibiotic resistance. However, this enzyme was associated with membrane-bound ribosomes involved in protein synthesis, which has a modest connection to the resistance mechanism of gentamicin.

variant		variant type	occurrence (out of 10)	mean gini coefficient
GNAT family N-acetyltransferase [Staphylococcus aureus]*		annotated gene	8	0.07139
IS256 family transposase IS256		annotated gene	7	0.0409
Bifunctional AAC-APH		annotated gene	6	0.08629
Uncharacterised protein [Staphylococcus aureus]*	pro-	annotated gene	5	0.02758
helix-turn-helix domain-containing protein [Staphylococcus aureus]*		annotated gene	5	0.03188
mercuric transport protein MerT [Staphylococcus aureus]*	pro-	annotated gene	4	0.01873
Alkylmercury lyase		annotated gene	4	0.02523
dihydrolipoamide dehydrogenase		annotated gene	4	0.01997
hypothetical protein [Staphylococcus aureus]*	protein	annotated gene	3	0.03044
Mercuric operon regulatory protein	resistance	annotated gene	3	0.02298

Table 4.13: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic gentamicin. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

4.5.5 Ciprofloxacin

Ciprofloxacin is a fluoroquinolone antibiotic that interferes with DNA replication through inhibition of DNA gyrase and topoisomerase IV. Common mechanisms of resistance against ciprofloxacin include mutations in the genes *gyrA* and *gyrB* encoding the DNA gyrase and *grlA* and *grlB* encoding the topoisomerase IV in *S. aureus* [98]. Three mutations in those genes are among the highest-ranked features (see table 4.14). The missense mutation (C:T) in DNA gyrase gene *gyrA* is resulting in a substitution of serine-84 to leucine. This is an already known mutation in ciprofloxacin-resistant *S. aureus* [99]. The two missense mutations in the topoisomerase IV gene *grlA* codon 80 are serine-to-phenylalanine (second highest-ranked mutation) and serine-to-tyrosine (fifth highest-ranked mutation) exchanges that have previously been reported [100]. The intergenic mutation found in the results, lies in the *mepRAB* operon, which is an efflux pump associated with resistance to ciprofloxacin [101]. This highlights again that it is important to further explore mechanisms of resistance conferred through non coding regions. Another feature that was termed important in 7 out of 10 models is the ImmA/IrrE family metallo-endopeptidase. According to Karimaei et al. it is only indirectly related to resistance, as it is part of the stress response under antibiotic exposure. Many of the genomes used to train the ciprofloxacin model were presumably also resistant to methicillin or other β -lactams, as some of the most important features are PBP2a, MaoC family dehydratase and glycerophosphoryl diester phosphodiesterase. These variants are nonspecific to the prediction of resistance to ciprofloxacin and introduce confounding effects. For the uncharacterized DUF6119 family protein and the YwqG family protein no known connection to resistance was found.

variant	variant type	occurrence (out of 10)	mean gini coefficient
DNA gyrase, A subunit	snp: missense	8	0.06185
DNA topoisomerase IV, A subunit	snp: missense	8	0.03814
ImmA/IrrE family metallo-endopeptidase [Staphylococcus aureus]*	annotated gene	7	0.00475
PBP2a family beta lactam resistant peptidoglycan transpeptidase MecA	annotated gene	5	0.0163
DNA topoisomerase IV, A subunit	snp: missense	5	0.01165
MaoC family dehydratase [Staphylococcus aureus]*	annotated gene	4	0.01181
DUF6119 family protein [Staphylococcus aureus]*	snp: other	3	0.00496
glycerophosphoryl diester phosphodiesterase [Staphylococcus aureus]*	annotated gene	3	0.02584
intergenic	snp: intergenic	3	0.00663
YwqG family protein [Staphylococcus aureus]*	snp: synonymous	3	0.00772

Table 4.14: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic ciprofloxacin. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

4.5.6 Erythromycin and clindamycin

Erythromycin and clindamycin both target the 50S ribosomal subunit, thereby arresting protein synthesis. Resistance is conferred by reducing the binding affinity of the antibiotic through methylation of the 23S rRNA, export, or modification of the antibiotic itself. Genes that are associated with macrolide or lincosamide resistance include *erm(C)*, *erm(A)* and *erm(B)* [67]. As expected, known resistance determinants were considered important for both models. Using the CARD database, the top dimethyladenosine transferase was identified as *ermC* (99.590 %) [46]. The rRNA (adenine-N-6)-methyltransferase in table 4.16 (erythromycin) is encoded by the *ermA* gene. Another class I SAM-dependent methyltransferase was found. However, it was not possible to match any existing AMR determinant of the CARD database with a high percentage of identity against it [46]. For both antibiotics, many genes were found that are related

to other antibiotic resistances. Both show variants associated with resistance to methicillin (PBP2a, glycerophosphoryl diester phosphodiesterase, MecR1, MaoC family dehydratase), with resistance to ciprofloxacin (mutations in DNA gyrase and topoisomerase IV) and with resistance to gentamicin (aminoglycoside nucleotidyltransferase ANT(9)) [103]. Liu et al. also report that they find the glycerophosphoryl diester phosphodiester in the results for their feature importance analysis. The presence of mobile genetic elements like recombinases is highly correlated with the acquisition of antibiotic resistance in *S. aureus* as they enable resistance genes to spread between bacteria. They leave traces of their existence in the genome, like the conserved RepL domain, which is part of the Rep protein in RC-replicating plasmids [6].

variant	variant type	occurrence (out of 10)	mean gini coefficient
glycerophosphoryl di- ester phosphodiesterase [Staphylococcus au- reus]*	annotated gene	10	0.02888
DNA gyrase, A subunit	snp: missense	9	0.01776
PBP2a family beta lac- tam resistant peptido- glycan transpeptidase MecA	annotated gene	8	0.02083
dimethyladenosine transferase	annotated gene	8	0.02832
RepL [Staphylococcus aureus]*	annotated gene	7	0.02052
class I SAM-dependent methyltransferase [Staphylococcus au- reus]*	annotated gene	7	0.01463
DNA topoisomerase IV, A subunit	snp: missense	7	0.02525
site specific recombinase putative	annotated gene	5	0.00947
beta lactam sensor sig- nal transducer MecR1	annotated gene	5	0.01352
aminoglycoside nu- cleotidyltransferase ANT(9) [Staphylococ- cus aureus]*	annotated gene	5	0.01768

Table 4.15: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic clindamycin. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

variant	variant type	occurrence (out of 10)	mean gini coefficient
RepL [Staphylococcus aureus]*	annotated gene	10	0.03239
dimethyladenosine transferase	annotated gene	10	0.03829
class I SAM-dependent methyltransferase	annotated gene	10	0.00885
[Staphylococcus aureus]*			
rRNA (adenine-N-6)-methyltransferase	annotated gene	10	0.01086
PBP2a family beta lactam resistant peptidoglycan transpeptidase	annotated gene	10	0.00968
MecA			
glycerophosphoryl diester phosphodiesterase	annotated gene	9	0.01038
[Staphylococcus aureus]*			
MaoC family dehydratase [Staphylococcus aureus]*	annotated gene	9	0.00925
site specific recombinase	annotated gene	8	0.00962
putative			
aminoglycoside nucleotidyltransferase	annotated gene	8	0.00987
ANT(9) [Staphylococcus aureus]*			
DNA gyrase, A subunit	snp: missense	5	0.0062

Table 4.16: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic erythromycin. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

4.5.7 Co-trimoxazole

Co-trimoxazole is a combination drug of trimethoprim and sulfamethoxazole. They both target different enzymes in the synthesis of tetrahydrofolic acid. Sulfamethoxazole is a sulfonamide that inhibits the dihydropteroate synthase, while trimethoprim binds to dihydrofolate reductase (DHFR). Resistance can be conferred through mutations in the *dfrB* gene encoding dihydrofolate reductase, which was identified as *dfrG* using the CARD database (see table 4.17) [46]. This gene was originally acquired from *Enterococcus faecium* [104]. It is very interesting that a hypothetical protein and an insertion element are considered the most important features for the co-trimoxazole model. They were found together in at least one strain directly next to each other and next to the dihydrofolate reductase. Further exploration of these genes could yield valuable

insights into new AMR determinants. The variant in the lipase gene and the ABC transporter VraE are both spurious features that do not contribute to the prediction of resistance to co-trimoxazole [105, 106]. Apparently, there is a connection between tRNA modification enzymes and an increased sensitivity to antibiotics targeting the bacterial cell wall, but other than that the mutation in the tRNA methyltransferase TrmB gene does not have a direct effect in the prediction of resistance against this antibiotic [107]. The nucleotidyltransferase that was considered important could be involved in resistance to aminoglycosides, but no match was reported when using the CARD database [46, 103]. The DNA translocase FtsK coordinates segregation in cell division and has no connection to antibiotic resistance [108]. Controlling the balance of metal ions in the cell is essential for the virulence of *Staphylococcus aureus*. The MntH Nramp family divalent metal transporter plays a role in the import of manganese ions, but is not relevant for resistance to co-trimoxazole [109]. Missense mutations in the *fabH* gene (beta-ketoacyl-ACP synthase III), are associated with resistance against antibiotics targeting the elongation condensing enzyme (*fabF*) but are nonspecific for this prediction purpose [110].

variant		variant type	occurence (out of 10)	mean gini coefficient
hypothetical	protein	annotated gene	9	0.05616
[Staphylococcus aureus]*				
insertion element	protein	annotated gene	8	0.07329
[Staphylococcus aureus]*				
dihydrofolate reductase		annotated gene	7	0.08795
YSIRK	domain-	snp: other	7	0.01725
containing triacylglycerol lipase Lip2/Geh				
[Staphylococcus aureus]*				
peptide resistance ABC transporter	permease subunit VraE	snp: missense	5	0.02227
MULTISPECIES:		snp: missense	5	0.01652
tRNA (guanosine(46)-N7)-methyltransferase				
TrmB [Staphylococcus]*				
nucleotidyltransferase		snp: missense	4	0.01409
[Staphylococcus aureus]*				
DNA translocase FtsK		snp: missense	3	0.02479
[Staphylococcus aureus]*				
Nramp family	divalent metal transporter	snp: synonymous	3	0.01966
[Staphylococcus aureus]*				
beta-ketoacyl-ACP synthase III		snp: missense	3	0.01796
[Staphylococcus aureus]*				

Table 4.17: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic co-trimoxazole. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

4.5.8 Fusidic acid

Resistance against fusidic acid is mediated by acquisition of *fusB*, a zinc finger domain containing protein that rescues the antibiotic-target-complex. Another known mechanism to inhibit fusidic acid binding is through mutations in the *fusA* gene, which encodes the elongation factor G [7, 67]. The gene *fusB* was not found in the analysis, but a missense mutation (TT:AA) in the *fusA* gene

resulting in the substitution of leucine-461 to lysin was considered important. Possession of the *fusC* gene is also a known resistance mechanism for fusidic acid and therefore a plausible important feature [15]. The first five ranked variants as well as the 7th importance feature were found next to each other in at least one strain and contain multiple unknown hypothetical proteins and a functionally uncharacterized FRG domain containing protein. Askarian et al. report that the gene *tirS* is located on the mobile genetic element SCC₄₇₆ next to a fusidic acid resistance gene and multiple hypothetical proteins. This organization was found here in the data as well. The single-stranded DNA binding protein is also part of SCC elements and responsible for their stable maintenance and dissemination [112]. As discussed for the AMR models of tetracycline and gentamicin, HTH-containing proteins are involved in modulating the transcriptome in response to stress conditions [94]. It may seem strange that the VOC family protein, which has no connection to antibiotic resistance occurs more often than the missense mutation in the *fusA* gene encoding the translation factor G. Various mutations in this gene contribute to resistance and they do not cluster into a single, easily detectable mutation [113]. The genomic region around the intergenic mutation was analyzed and revealed that it is located upstream of the *rrna* operon.

variant			variant type	occurrence (out of 10)	mean gini coefficient
FusC	family	fusidic acid resistance	annotated gene	10	0.01549
		EF-G-binding protein [Staphylococcus aureus]*			
	hypothetical	protein	annotated gene	10	0.01749
		[Staphylococcus aureus]*			
	FRG domain-containing protein	[Staphylococcus aureus]*	annotated gene	10	0.0109
	hypothetical	protein	annotated gene	9	0.00992
		[Staphylococcus aureus]*			
	NAD(+)	hydrolase	annotated gene	9	0.00853
	TirS	[Staphylococcus aureus]*			
	MULTISPECIES: family	VOC protein	annotated gene	8	0.00501
		[Staphylococcus]			
	MULTISPECIES: cassette	chromosome ssDNA-binding protein	annotated gene	7	0.00519
		[Staphylococcus]*			
	helix-turn-helix transcriptional	trans-regulator	annotated gene	7	0.00412
		[Staphylococcus aureus]*			
	translation factor G	elongation	snp: missense	6	0.00748
	intergenic		snp: intergenic	4	0.00217

Table 4.18: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic fusidic acid. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

4.5.9 Vancomycin

The glycopeptide vancomycin inhibits the synthesis of the bacterial cell wall through a different mechanism than β -lactams do and is therefore an alternative choice for MRSA. One must distinguish between intermediate resistance to vancomycin, which is linked to the *vraSR* operon or the *graS* gene and vancomycin-resistant *Staphylococcus aureus* (VRSA) that acquired the *vanRS* and *vanHAXYZ* genes [67]. None of the genes mentioned above appeared in the feature importance analysis. The top result is an ABC transporter that showed a 28.452% similarity

to the VRA-F ABC transporter [46]. It is an ATP-binding cassette antibiotic efflux pump that confers resistance to glycopeptides [73]. The MarR family transcription factor regulates multi-resistant efflux pumps that recognize various different antibiotics [114]. The other features are either uncharacterized proteins or are not associated with resistance to vancomycin or any other antibiotic. The intergenic mutation is also near an functionally unknown protein.

variant		variant type	occurence (out of 10)	mean gini coefficient
ABC	transporter	annotated gene	4	0.00535
ATP-binding	protein			
[Staphylococcus	au-			
reus]*				
DUF3113	family pro-	annotated gene	3	0.00506
tein [Staphylococcus	au-			
reus]*				
Fic	family protein	snp: synonymous	3	0.00428
[Staphylococcus	au-			
reus]*				
HAD	family hydrolase	snp: synonymous	2	0.00417
[Staphylococcus	au-			
reus]*				
DUF1108	family pro-	snp: missense	2	0.00495
tein [Staphylococcus	au-			
reus]*				
tandem-type	lipopro-	snp: synonymous	2	0.00387
tein Lpl10 [Staphylo-				
coccus aureus]*				
MarR	family tran-	snp: intergenic	2	0.00558
scriptional	regulator			
[Staphylococcus	au-			
reus]*				
intergenic		annotated gene	2	0.00306
hypothetical	protein	annotated gene	2	0.00635
[Staphylococcus	au-			
reus]*				
phage tail spike	pro-	snp: synonymous	2	0.00604
tein [Staphylococcus	au-			
reus]*				

Table 4.19: Top 10 results of the feature importance analysis of the 10 random forest models of the antibiotic vancomycin. Mutation in genes and hypothetical proteins were identified using BLAST and are indicated by a star [45].

4.6 Genome Wide Association Study (GWAS) based Identification of Resistance-Associated Variants

Genome-wide association studies (GWAS) provide an alternative method to identify potential new AMR determinants from genomic input data. This is done by statistical associations between genomic features and the resistance phenotype. In a single-locus approach, each variant is individually tested for its effect on the phenotype and assigned a P value. Machine learning methods are increasingly being incorporated into GWAS because phenotypes are usually the product of multiple epistatically linked variants. Microbial GWAS also face difficulties with linkage disequilibrium, population stratification or heterogeneity due to horizontally transferred AMR genes and population structure [20, 115]. The ALPAR pipeline features a GWAS as a classical way to detect mutations and genes relevant to the resistance phenotype. In this work, the most significant variants were compared to the features that were considered important in the machine learning models. Similar to the feature importance analysis, features associated with the *SCCmec* were found in the GWAS for the four β -lactams cefoxitin (A), methicillin (B), oxacillin (C) and penicillin (D) (see figure 4.5). Among those are the *mecA* and *mecR1* genes, the glycerophosphoryl diester phosphodiesterase, the MaoC dehydratase, the HMG CoA synthase and the cassette chromosome recombinase CcrA. In contrast to machine learning, the GWAS results for oxacillin also identified the cassette chromosome recombinase CcrB among the variants with the highest negative logarithm of the P value. Other features like the type 1 restriction enzyme, the recombinase family proteins and the MrcB family domain-containing protein were again identified by both methods. The associative methods differed in finding features that included the SAUGI family uracil-DNA glycosylase inhibitor, identified in the results for methicillin and oxacillin, the IS1182 family transposase and multiple uncharacterized proteins. In staphylococci the IS1182 family transposase is associated with resistance against streptomycin, kanamycin or streptothricin and therefore not relevant for resistance against β -lactams. According to Muselmani et al., SAUGI is also often localized on *SCCmec* cassettes, which explains why it was statistically associated with the resistant phenotype. The GWAS results for penicillin are similar to those of the feature importance analysis.

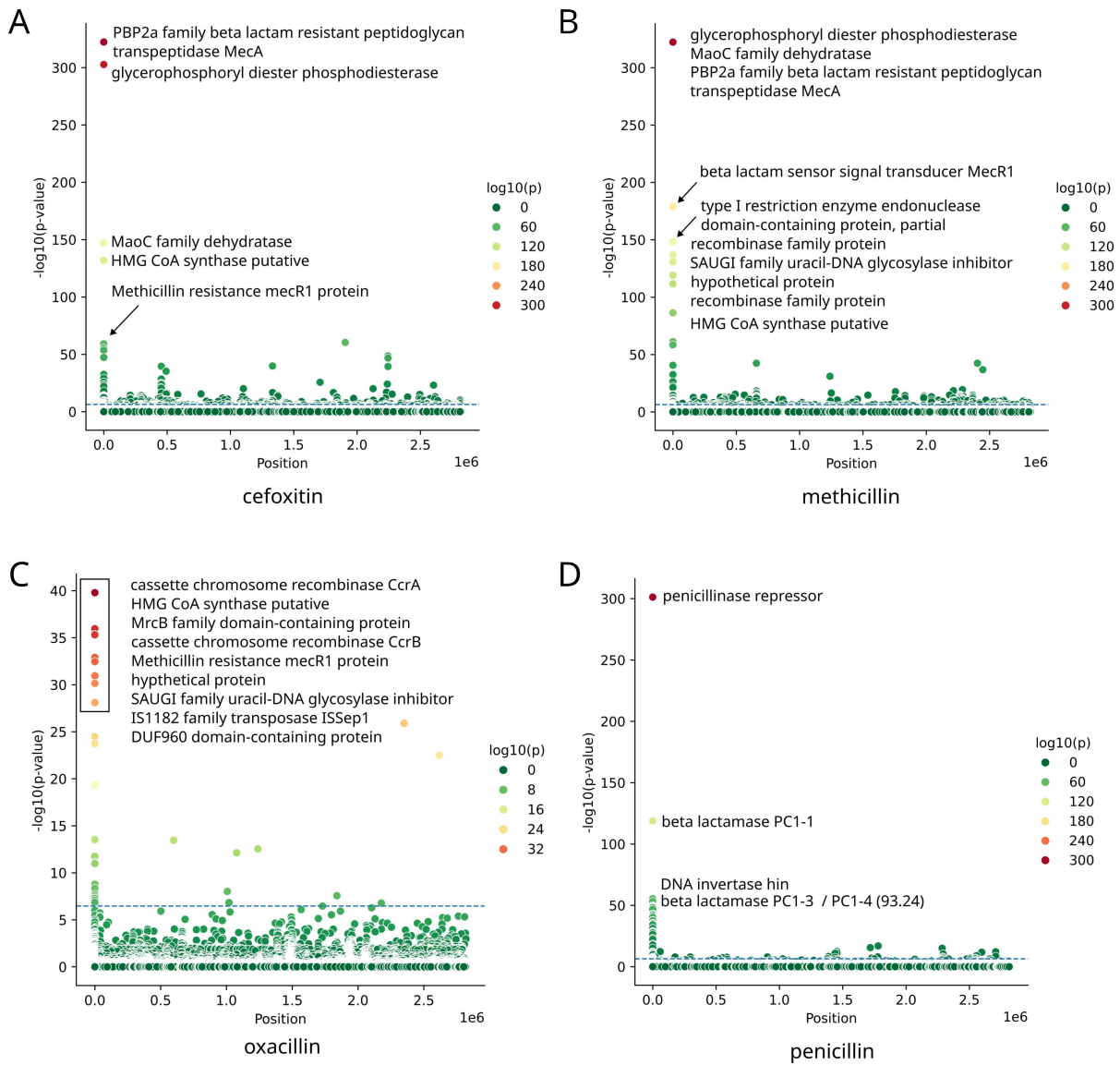


Figure 4.5: Annotated GWAS plots of the different beta lactams A) cefoxitin B) methicillin C) oxacillin D) penicillin.

The results of the GWAS for the antibiotics erythromycin (A) and clindamycin (B) show features that are also among the most important variants for the machine learning model (see figure 4.6). However, in comparison, the GWAS results are far more specific for both antibiotics, as the most significant features do not include resistance determinants associated with other antibiotics. The resistance determinant *ermA* appears among the GWAS results for clindamycin, whereas for the model-based analysis it was related to erythromycin.

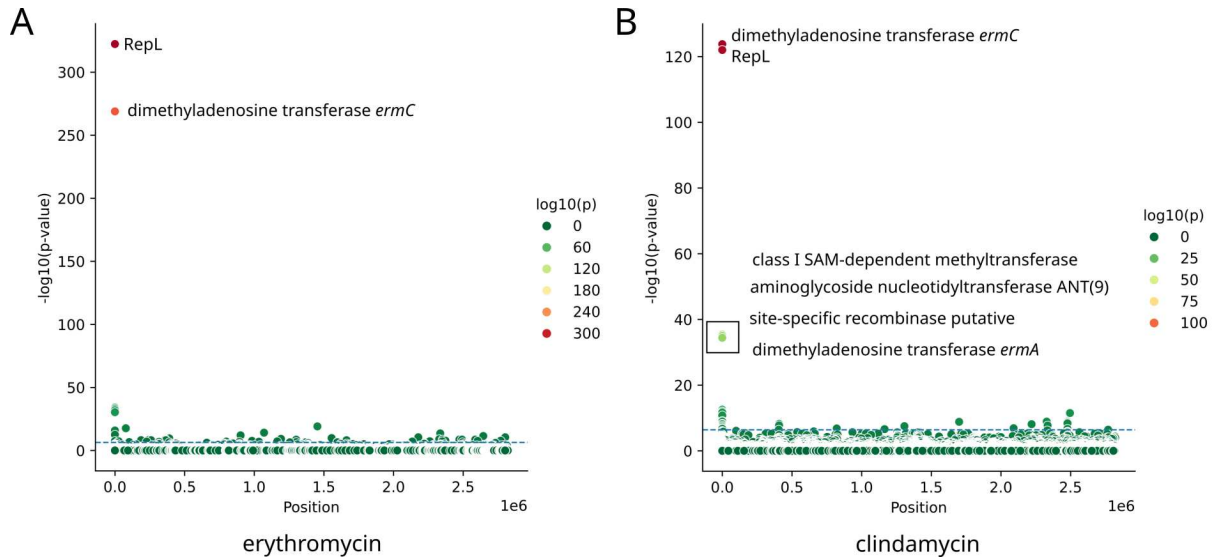


Figure 4.6: Annotated GWAS plots of A) erythromycin and B) clindamycin.

For gentamicin (A) and tetracycline (B), the GWAS results show a known AMR determinant as the most significant variant (see figure 4.7). Resistance against gentamicin can arise through different mechanisms, one of which involves the bifunctional AAC-APH in *S. aureus* [103]. For tetracycline, the *tetK* gene is known to confer resistance [67]. The MobV family relaxase and the hypothetical protein are also among the results of the feature importance analysis and their impact was discussed in chapter 4.5. Unlike in the machine learning models, a compelling pattern of SNPs was observed for both antibiotics around two regions. Except of one synonymous mutation, all of them are intergenic and localized upstream of the rRNA operon. A common resistance mechanism for aminoglycosides are mutations in the 16S rRNA, but not upstream of the operon [12]. No relevant publications on this topic were found, but mutations in regulatory regions have been reported to have an impact in antibiotic resistance [88, 117]. Because aminoglycosides target the 16S rRNA in the 30S subunit of the ribosome, mutations upstream of the rRNA operon can lead to upregulation of the 16S rRNA gene expression [12].

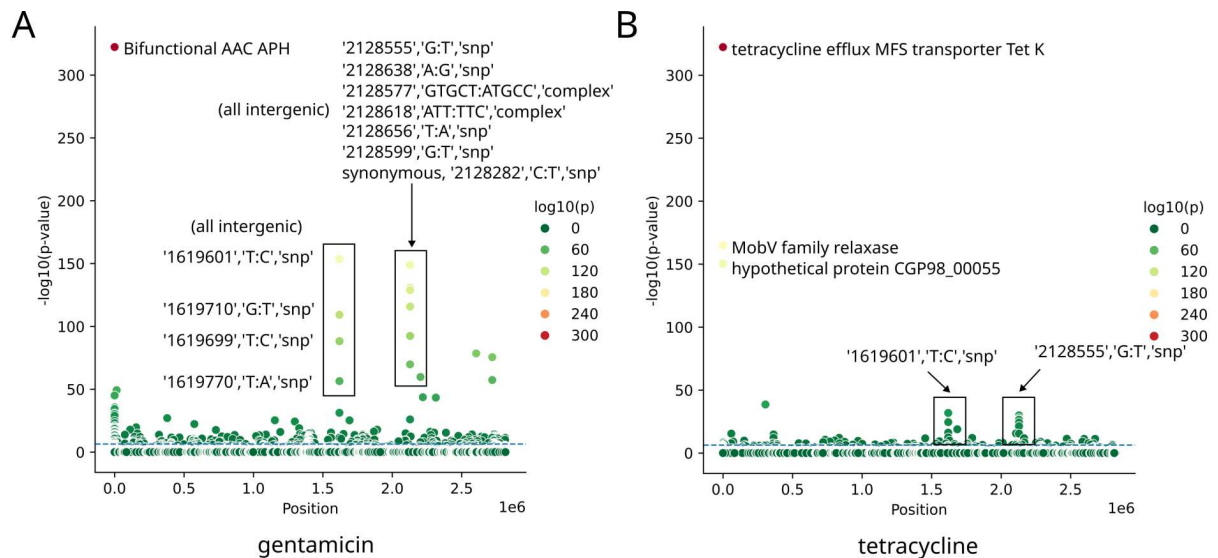


Figure 4.7: Annotated GWAS plots of A) gentamicin and B) tetracycline.

The GWAS results for ciprofloxacin contain mostly mutations (see figure 4.8 A). The highly significant missense mutation at position 7255 in the *gyrA* gene is a common resistance conferring mutation in *S. aureus*, resulting in a serine-84 to leucine substitution [99]. The frameshift variant that leads to a premature stop codon in the DUF6119 family protein was also listed among the most important features of the ciprofloxacin machine learning model. The GWAS for co-trimoxazol showed similar results to the feature importance analysis of the respective model, with an insertion element, an unknown hypothetical protein and a dihydrofolate reductase encoded by *dfrG* among the most significant variants (see figure 4.8 B).

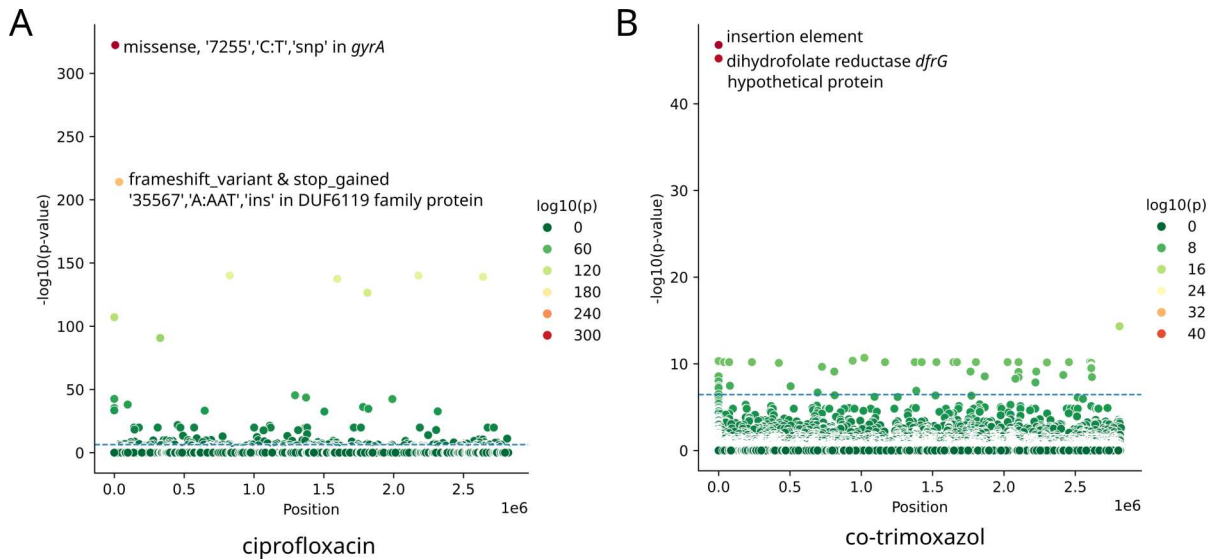


Figure 4.8: Annotated GWAS plots of A) ciprofloxacin and B) co-trimoxazol

The GWAS for vancomycin did not produce robust findings as no significant variants were detected (see figure 4.9 B). Although significant mutations were found in the GWAS for fusidic acid, no known resistant determinants of resistance against fusidic acid could be identified (see figure 4.9 A). Three significant mutations are among the top results. The relevance of the loss of the start codon in the unknown hypothetical protein for antibiotic resistance should be further studied. Duncan et al. report that *Escherichia coli* mutants deficient in uracil-DNA glycosylase accelerate acquisition of antibiotic resistance due to a higher mutation rate. The third mutation is in the glycytransferase FmhC, which is involved in peptidoglycan synthesis and is reported to confer resistance to lysostaphin [119]. These are all spurious variants that are not specific to resistance to fusidic acid.

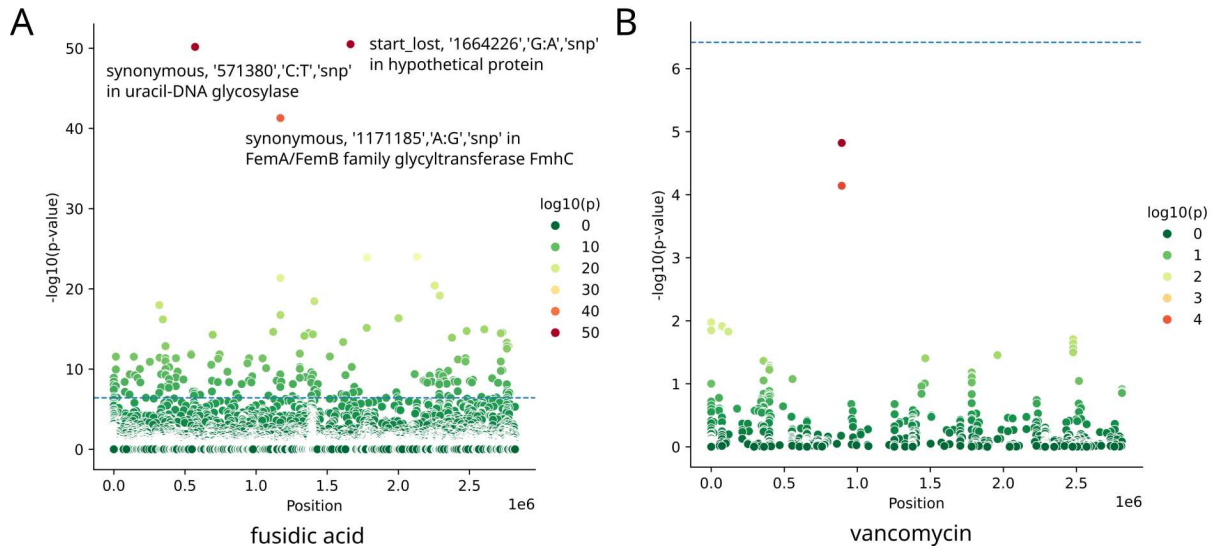


Figure 4.9: Annotated GWAS plots of A) fusidic acid and B) vancomycin.

4.7 Antimicrobial Resistance Predictions for a *Staphylococcus aureus* Isolate Collection

A *S. aureus* isolate of 87 isolates collected of patients suffering from atopic dermatitis was used to test the performance of the ALPAR pipeline on unseen data. The predictions were compared to ABRicate, a benchmarked AMR prediction tool [21]. ABRicate screens genomic data for acquired resistance genes using publicly available databases [21]. The predictions for all isolates from one patient are grouped together (see figure 4.10). For ciprofloxacin, methicillin, oxacillin, fusidic acid and vancomycin all isolates were predicted to be susceptible. This is surprising, given the widespread topical use of fusidic acid in atopic dermatitis, which would lead to the development of resistances in some cases [120]. Macrolides like erythromycin and lincosamides such as clindamycin are both part of the MLS antibiotics that share similar mechanisms of resistance. ABRicate does not distinguish between resistance to erythromycin and clindamycin, while ALPAR was trained with two different training data sets specifically for each antibiotic. The prediction results for tetracycline were questionable, because ABRicate predicted resistance for all isolates, while ALPAR predicted that all isolates were susceptible. ABRicate detected the resistance determinant *tet38* in all isolates, which is an efflux pump that confers resistance to tetracycline [46]. The *S. aureus* reference genome was screened by ABRicate too, revealing that it also exhibited the *tet38* gene. During variant calling, the ALPAR pipeline selects genes that are present in the training data genomes but are missing in the reference genome. This efflux pump is highly conserved in *S. aureus* strains and apparently also in the reference genome of strain NCTC 8325 [121].

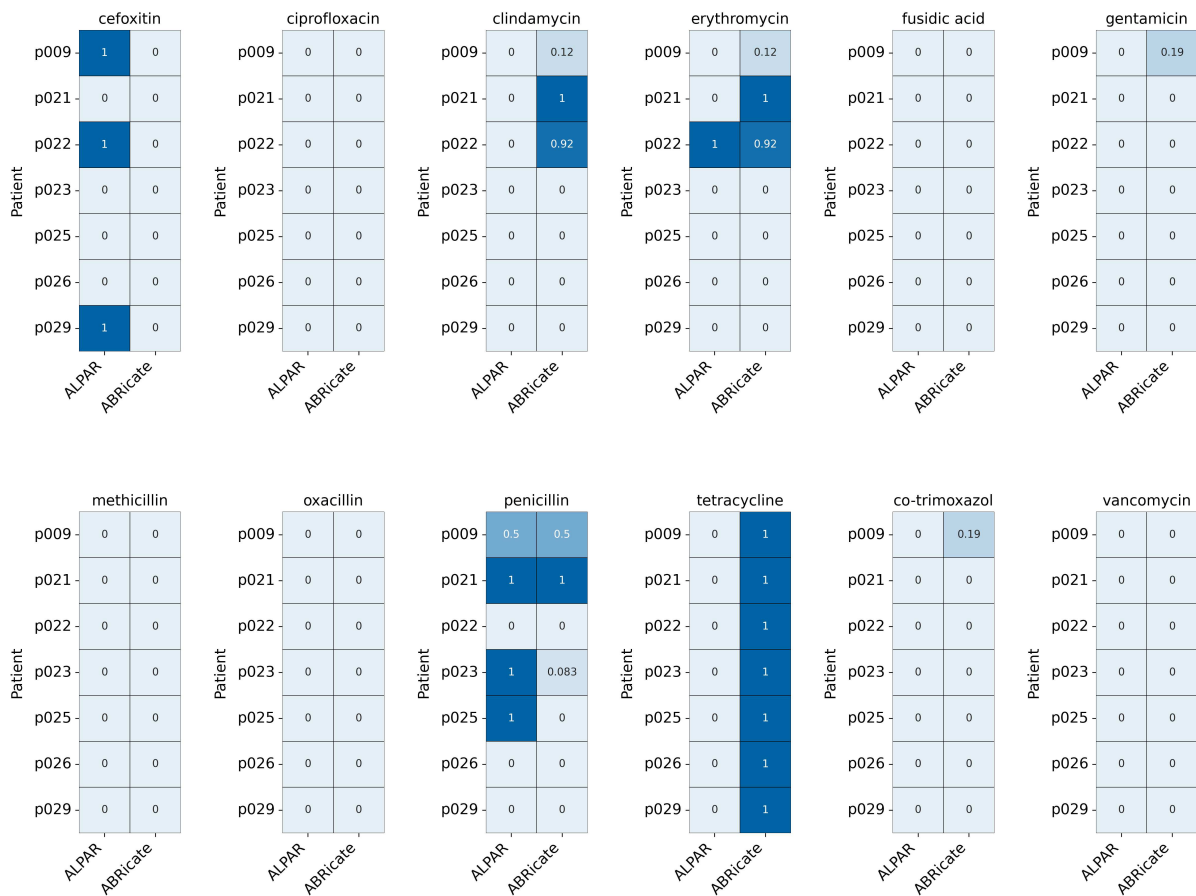


Figure 4.10: Antimicrobial Resistance Predictions by the ALPAR pipeline and ABRicate for the *Staphylococcus aureus* isolate collection. Predictions for all isolates from a patient are clustered together and percentage of isolates predicted resistant are shown in a color gradient.

Prediction outcome matrices were generated for every antibiotic (see figure 4.11). Here, all predictions for the 87 isolates by the two pipelines were sorted into four categories: both classified as resistant, both predicted as susceptible, and two groups where classification outcomes differ between tools. These plots were used to decide which antibiotics should be tested experimentally and further evaluated. For the majority of the antibiotics, both pipelines shared the same results for the classifications. Erythromycin and penicillin showed the highest difference in predictions between ALPAR and ABRicate and were therefore chosen for experimental validation. For further details on the predictions for all isolates that were tested experimentally, please see the supplementary section.

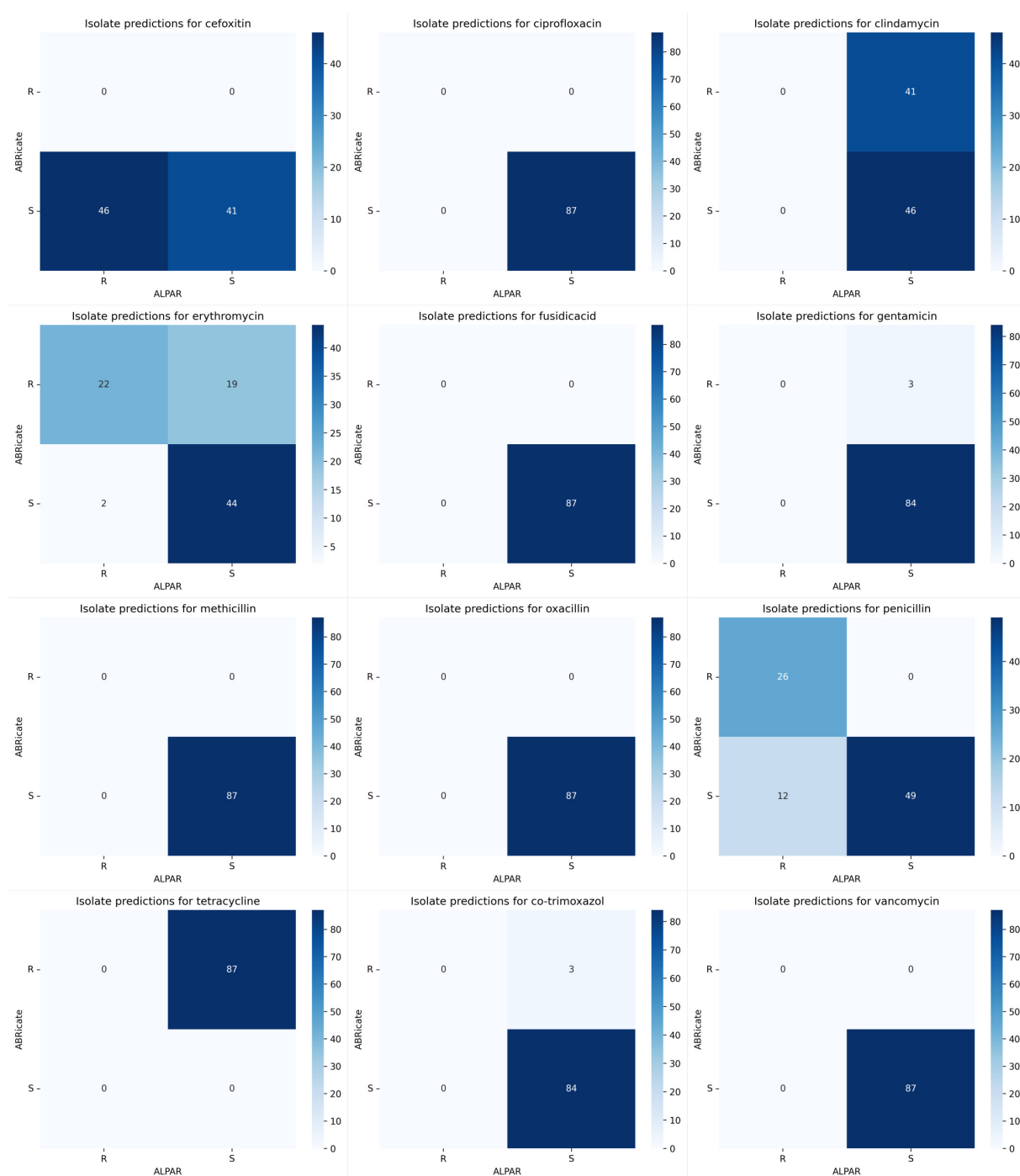


Figure 4.11: Prediction patterns of ALPAR for the isolate collection for all antibiotics.

4.8 Experimental validation of Antibiotic Resistance Predictions for Erythromycin and Penicillin

A minimum inhibitory concentration (MIC) assay was used to determine antibiotic susceptibility for erythromycin and penicillin. This assay is a gold standard in clinical settings that determines the lowest concentration of antibiotic exposure that inhibits visible growth of bacteria [122]. For each antibiotic, twelve isolates were chosen for antibiotic susceptibility testing.

For all isolates included in the experiment growth was visible (blue bars in figure 4.12). The error bars indicate the fluctuation between the six replicate values of the biological replicas (see supplementary section figure 6.13). While penicillin was dissolved in water, ethanol was used as

a solvent for erythromycin. The green bars are vehicle controls that were designed to test the effects of ethanol on bacteria. Although all isolates exhibited growth, isolate 6.07 showed the least growth. The use of ethanol did not demonstrate a negative impact on growth.

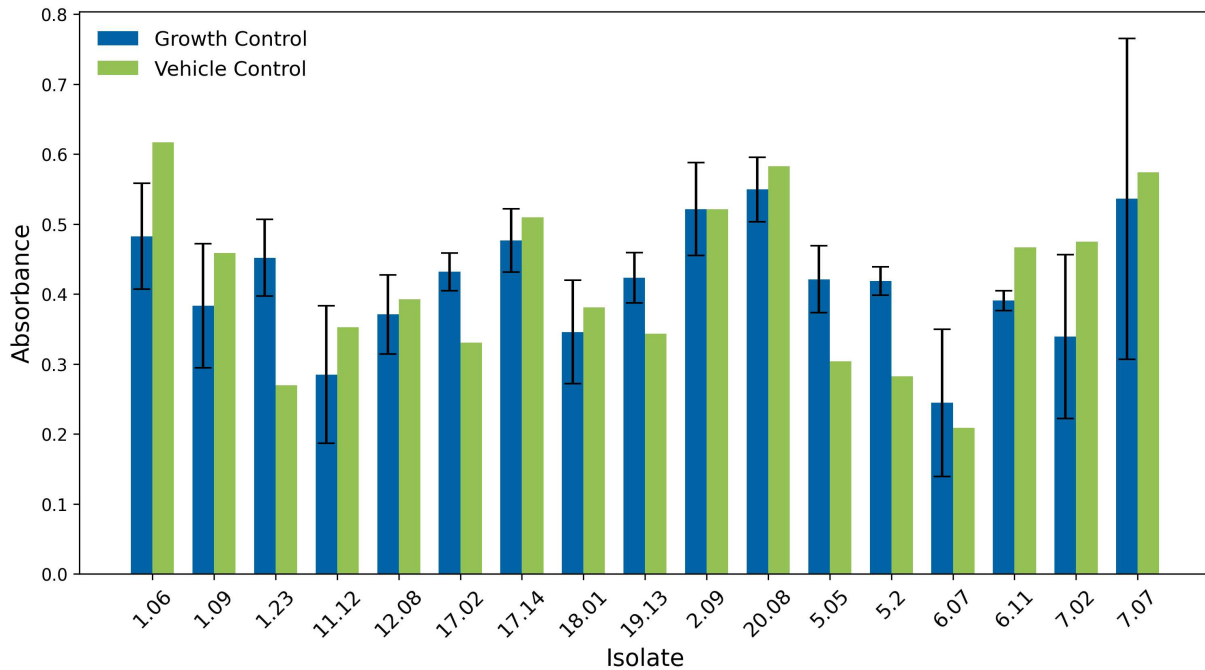


Figure 4.12: Growth controls (blue bars) and vehicle controls (green bars) of all tested isolates for both biological replicates. Vehicle controls were conducted for ethanol as a solvent of erythromycin.

In order to control for sterility, a negative control with medium alone and a sterile control for each antibiotic with medium and antibiotic stock solution were tested. With an absorbance of around 0.05 the controls exhibit no growth and confirm to be sterile.

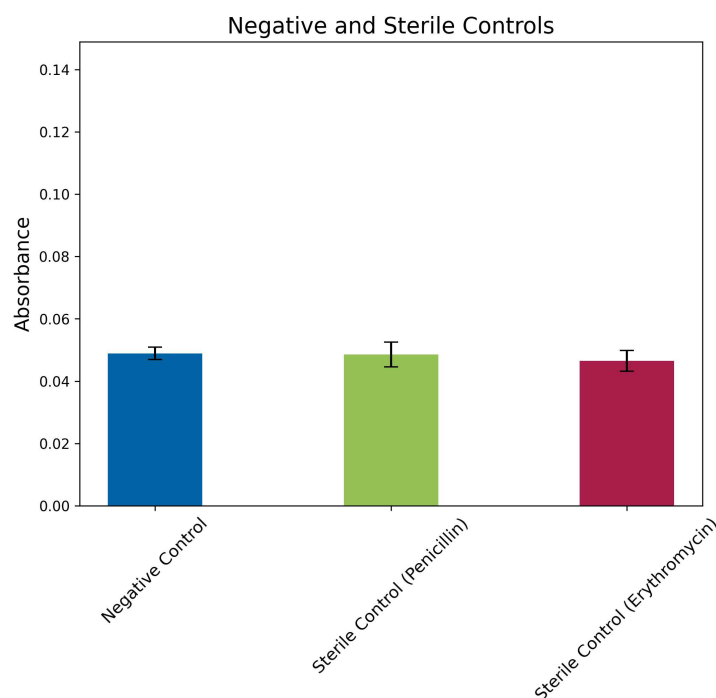


Figure 4.13: Negative and sterile controls of the two tested antibiotics penicillin and erythromycin for both biological replicates.

In each figure (see figures 6.14 to 6.20 in the supplementary section), all isolates of a category in the prediction outcome matrix are displayed together. The absorbance after incubation is shown as a function of the different antibiotic concentrations. Clinical breakpoints are indicated via a red line, which is used to determine whether an isolate is counted as resistant or susceptible. In the following plots, the results from the MIC assay are compared to the predictions by both pipelines ALPAR and ABRicate. The first section will discuss the findings for the antibiotic erythromycin (see figure 4.14) and the second one will focus on penicillin (see figure 4.16).

Isolates 1.09 and 2.09 are each predicted to be resistant by both pipelines. The experiment reveals that one of the two isolates is susceptible to erythromycin and one is resistant, indicating that both pipelines failed to correctly predict the antibiotic resistance phenotype for isolate 2.09. ABRicate detected in both isolates the resistance determinant *ermT*, a 23S rRNA methyltransferase that confers resistance to macrolides, lincosamides and streptogramins [123]. It is not clear why isolate 2.09 is susceptible to erythromycin. As *ermT* is a gene that can be inducible expressed, it could be the case that there is a mutation in its regulation sequence [123]. For the next category discussed, the classifications of the pipelines differ. ALPAR predicts resistance to erythromycin, whereas ABRicate suggests susceptibility. This category includes isolates 1.06 and 1.23. As before, both isolates show different resistance phenotypes. Isolate 1.06 shows resistance to erythromycin, even though the gold standard tool predicted otherwise. The three isolates 18.01, 17.14 and 7.07 were predicted to be susceptible to erythromycin by ALPAR, while ABRicate identifies the isolates as resistant. Two isolates, 18.01 and 17.14, showed resistance, while isolate 7.07 was susceptible. The resistance prediction by ABRicate for isolate 07.07 is supported by the detection of *ermC*. For the last category, both pipelines classified the isolates 6.07, 5.05, 19.13, 20.08 and 12.08 as susceptible to erythromycin. Only three of the five isolates exhibited susceptibility. Isolates 19.13 and 6.07 demonstrated a resistant phenotype.

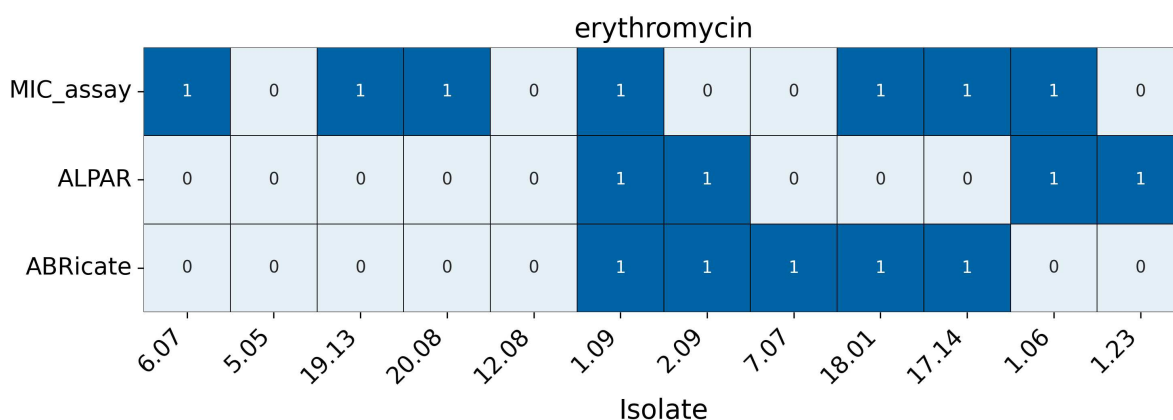


Figure 4.14: Antibiotic resistance predicted by ALPAR and ABRicate for erythromycin, compared to the observed phenotypes in the MIC assay (1 = resistant, 0 = susceptible).

Figure 4.15 shows the presence and absence of the top ten variants from the feature importance analysis for erythromycin in the tested isolates. On the one hand, many resistant isolates do not exhibit typical resistance determinants such as *ermC* (dimethyladenosine transferase), on the other hand susceptible isolates like 7.07 show presence of resistance determinants in their genome. Focusing only on the presence of genes ignores the possibility that genes might not be expressed [20]. Since only the top ten features are displayed, it is possible that some relevant resistant determinants were not highlighted, as their importance might have been obscured by the presence of spurious features.

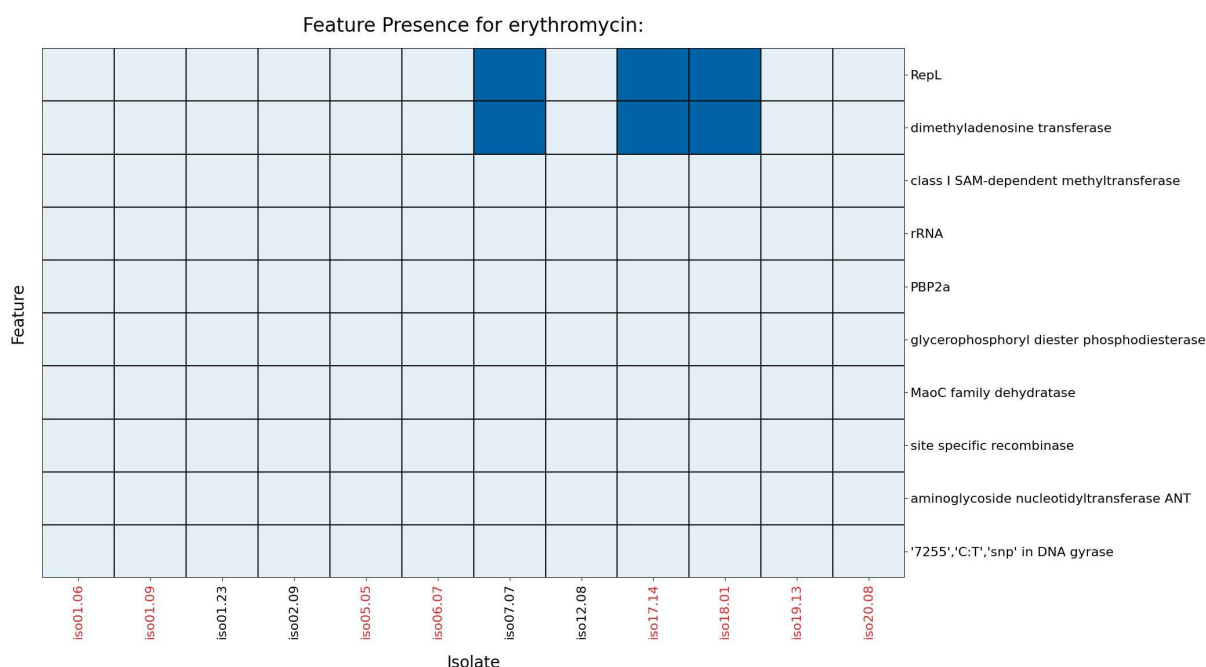


Figure 4.15: Presence (dark blue) and absence (light blue) of the top 10 features from the random forest feature importance analysis of erythromycin for the tested isolates. Isolates tested resistant are colored red.

For all isolates that were classified as resistant to penicillin by ALPAR and ABRicate, the predictions of both pipelines align with the observed phenotype in the experiments (see figure

4.16). The two tools disagree on the predicted phenotype of isolates 5.05 and 11.12, with ALPAR predicting resistance and ABRicate predicting susceptibility. Isolate 11.12 was indeed susceptible as predicted by ABRicate and isolate 5.05 was correctly classified as resistant to penicillin by ALPAR. All isolates tested that were predicted susceptible by both pipelines exhibited a susceptible phenotype except isolate 6.07.

Figure 4.16: Antibiotic resistance predicted by ALPAR and ABRicate for penicillin, compared to the observed phenotypes in the MIC assay (1 = resistant, 0 = susceptible).

Figure 4.17: Presence (dark blue) and absence (light blue) of the top 10 features from the random forest feature importance analysis of penicillin for the tested isolates. Isolates tested resistant are colored red.

To evaluate the performance of the antibiotic resistance predictions of both tools, ALPAR and ABRicate are compared to the experimentally observed phenotypes from the MIC assays. Based on the overlap of predictions and experimental test results, the predictions for penicillin align overall better than those for erythromycin. Evaluation of the performance of the tools on the experimentally tested isolates can be obtained from tables 4.20 and 4.21. The error rates of 0.58 (ALPAR) and 0.50 (ABRicate) are quite high for the erythromycin model. Surprisingly, the benchmarked tool also fails to achieve high prediction accuracy. The negative MCC of the erythromycin model indicates that something is systematically wrong with the ALPAR model as it is making a lot of misclassifications. The MCC for ABRicate (0.03) is not good either and is comparable to a model making random decisions. Although ABRicate performed better at predicting resistance for this test data set, both pipelines generally did not work well for this antibiotic. Notably, the pipelines achieved better results when applied to penicillin. Both tools show the same error rate of 0.17 and overall exhibit similar performance. They outperform each other in different metrics. While the MCC of 0.66 is lower for ALPAR, compared to ABRicate with 0.71, the F1 score of 0.86 is higher for ALPAR than for ABRicate (0.83). Although both pipelines failed to make accurate predictions for erythromycin, they performed well in the case of penicillin. These findings highlight the importance of a diverse and high-quality data set as well as the impact that varying complexities of resistance mechanisms have on prediction outcomes.

	ALPAR	ABRicate
error rate	0.58	0.50
accuracy	0.42	0.50
precision	0.50	0.60
recall	0.29	0.43
specificity	0.60	0.60
F1	0.36	0.50
MCC	-0.12	0.03

Table 4.20: Evaluation of the prediction performance of ALPAR and ABRicate on the experimentally tested isolates for resistance to erythromycin.

	ALPAR	ABRicate
error rate	0.17	0.17
accuracy	0.83	0.83
precision	0.86	1.00
recall	0.86	0.71
specificity	0.80	1.00
F1	0.86	0.83
MCC	0.66	0.71

Table 4.21: Evaluation of the prediction performance of ALPAR and ABRicate on the experimentally tested isolates for resistance to penicillin.

5 | Conclusion and outlook

This thesis aimed to establish the antimicrobial resistance (AMR) prediction tool ALPAR for the high priority pathogen *Staphylococcus aureus*, compare AMR prediction to the established tool ABRicate, and evaluate its performance through experimental validation using MIC assays. The ALPAR pipeline takes phylogenetic relationships into account, making it an interesting new tool for predicting resistance using machine learning methods. This work sought to contextualize ALPAR with respect to clinical implementation, public AMR surveillance, or whether its use is limited to research purposes. A comparison with the widely used gold standard tool ABRicate should determine its reliability and effectiveness. The results showed that both pipelines perform comparably in the prediction of a test dataset composed of a *S. aureus* isolate collection. The random forest models demonstrated high performance during validation with an MCC between 0.749 to 0.985, but generalization to unseen data varied between the antibiotic models evaluated. Depending on the antibiotic, the error rates ranged from 0.58 to 0.17. A genome-wide association study as well as a feature importance analysis of the machine learning models were conducted to study the basis of the decisions, with several known resistance determinants detected among the key features. Multiple uncharacterized hypothetical proteins and mutations were identified as promising candidates that ought be investigated further. Data availability is currently one of the top limiting factors in AMR prediction using machine learning [51]. The training set in this work comprised all publicly available data on the BV-BRC database including all antibiotics associated with *S. aureus*, but it remained insufficient for some antibiotics like daptomycin and exhibited considerable imbalance for others such as vancomycin with only 3% resistant genomes. However, with the ongoing emergence of antimicrobial resistance and the accelerating improvement of next-generation sequencing techniques, more genomic data linked to an antibiotic resistance phenotype will be available in the future. For tools like ALPAR to be applicable in clinical settings, they must demonstrate improved reliability and provide experimental evidence to support the confidence of their results. Nevertheless, ALPAR remains a valuable research tool for identifying novel antimicrobial resistance determinants and investigating emerging resistance mechanisms.

Statement of Assistive Technologies

In this thesis, the OpenAI tool ChatGPT was used to explore ideas, in development of code and to improve the author's writing style [124]. All computations and data were generated on the Life Science Compute Cluster (LISC) [125]. For code development and correction, GitHub Copilot was used [126]. This work was written in Overleaf, graphics were made using Python (version 3.13.3) and Inkscape [127, 128].

References

- [1] J. O'Neill, Tech. Rep. 20 (2014).
- [2] M. A. Salam, M. Y. Al-Amin, M. T. Salam, J. S. Pawar, N. Akhter, A. A. Rabaan, and M. A. A. Alqumber, Healthcare **11**, 1946 (2023), ISSN 2227-9032, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10340576/>.
- [3] C. J. L. Murray, K. S. Ikuta, F. Sharara, L. Swetschinski, G. R. Aguilar, A. Gray, C. Han, C. Bisignano, P. Rao, E. Wool, et al., The Lancet **399**, 629 (2022), ISSN 0140-6736, 1474-547X, publisher: Elsevier, URL [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(21\)02724-0/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02724-0/fulltext).
- [4] M. I. Hutchings, A. W. Truman, and B. Wilkinson, Current Opinion in Microbiology **51**, 72 (2019), ISSN 13695274, URL <https://linkinghub.elsevier.com/retrieve/pii/S1369527419300190>.
- [5] J. M. Munita and C. A. Arias, Microbiology spectrum **4**, 10.1128/microbiolspec.VMBF (2016), ISSN 2165-0497, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4888801/>.
- [6] S. R. Patridge, S. M. Kwong, N. Firth, and S. O. Jensen, Clinical Microbiology Reviews **31** (2018), URL <https://pubmed.ncbi.nlm.nih.gov/30068738/>.
- [7] E. M. Darby, E. Trampari, P. Siasat, M. S. Gaya, I. Alav, M. A. Webber, and J. M. A. Blair, Nature Reviews Microbiology **21**, 280 (2023), ISSN 1740-1534, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/s41579-022-00820-y>.
- [8] S. Zimmermann, M. Klinger-Strobel, J. A. Bohnert, S. Wendler, J. Rödel, M. W. Pletz, B. Löffler, and L. Tuchscher, Frontiers in Microbiology **10** (2019), ISSN 1664-302X, publisher: Frontiers, URL <https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2019.02762/full>.
- [9] E. M. Halawa, M. Fadel, M. W. Al-Rabia, A. Behairy, N. A. Nouh, M. Abdo, R. Olga, L. Fericean, A. M. Atwa, M. El-Nablaway, et al., Frontiers in Pharmacology **14**, 1305294 (2024), ISSN 1663-9812, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10820715/>.
- [10] L. M. Lima, B. N. M. d. Silva, G. Barbosa, and E. J. Barreiro, European Journal of Medicinal Chemistry **208**, 112829 (2020), ISSN 0223-5234, URL <https://www.sciencedirect.com/science/article/pii/S0223523420308011>.

- [11] Y. Uehara, *Antibiotics* **11**, 86 (2022), ISSN 2079-6382, number: 1 Publisher: Multidisciplinary Digital Publishing Institute, URL <https://www.mdpi.com/2079-6382/11/1/86>.
- [12] Y. Doi, J.-i. Wachino, and Y. Arakawa, *Infectious disease clinics of North America* **30**, 523 (2016), ISSN 0891-5520, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878400/>.
- [13] M. C. Shutter and H. Akhondi, in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025), URL <http://www.ncbi.nlm.nih.gov/books/NBK549905/>.
- [14] J. Spížek and T. Řezanka, *Biochemical Pharmacology* **133**, 20 (2017), ISSN 0006-2952, URL <https://www.sciencedirect.com/science/article/pii/S0006295216304622>.
- [15] A. J. O'Neill, F. McLaws, G. Kahlmeter, A. S. Henriksen, and I. Chopra, *Antimicrobial Agents and Chemotherapy* **51**, 1737 (2007), ISSN 0066-4804, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1855526/>.
- [16] N. G. Bush, I. Diez-Santos, L. R. Abbott, and A. Maxwell, *Molecules* **25**, 5662 (2020), ISSN 1420-3049, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7730664/>.
- [17] A. Wróbel, K. Arciszewska, D. Maliszewski, and D. Drozdowska, *The Journal of Antibiotics* **73**, 5 (2020), ISSN 1881-1469, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/s41429-019-0240-6>.
- [18] A. Arnold, S. McLellan, and J. M. Stokes, *npj Antimicrobials and Resistance* **3**, 1 (2025), ISSN 2731-8745, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/s44259-025-00085-4>.
- [19] H. Bilal, M. N. Khan, S. Khan, M. Shafiq, W. Fang, R. U. Khan, M. U. Rahman, X. Li, Q.-L. Lv, and B. Xu, *Computational and Structural Biotechnology Journal* **27**, 423 (2025), ISSN 2001-0370, URL <https://www.sciencedirect.com/science/article/pii/S2001037025000078>.
- [20] J. I. Kim, F. Maguire, K. K. Tsang, T. Gouliouris, S. J. Peacock, T. A. McAllister, A. G. McArthur, and R. G. Beiko, *Clinical Microbiology Reviews* **35**, e00179 (2022), URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC9491192/>.
- [21] Thorsten Seemann, *Abricate*, URL <https://github.com/tseemann/abricate>.
- [22] M. Su, S. W. Satola, and T. D. Read, *Journal of Clinical Microbiology* **57**, 10.1128/jcm.01405 (2019), publisher: American Society for Microbiology, URL <https://journals.asm.org/doi/10.1128/jcm.01405-18>.
- [23] J. Grus, *Data Science from Scratch* (O'Reilly Media, 2019), 2nd ed.
- [24] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, *Nature Reviews Molecular Cell Biology* **23**, 40 (2022), ISSN 1471-0080, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/s41580-021-00407-0>.
- [25] W. S. Noble, *Nature Biotechnology* **24**, 1565 (2006), ISSN 1546-1696, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/nbt1206-1565>.

- [26] C. Kingsford and S. L. Salzberg, *Nature Biotechnology* **26**, 1011 (2008), ISSN 1546-1696, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/nbt0908-1011>.
- [27] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, *Artificial Intelligence Review* **54**, 1937 (2021), ISSN 1573-7462, company: Springer Distributor: Springer Institution: Springer Label: Springer Publisher: Springer Netherlands, URL <https://link-springer-com.uaccess.univie.ac.at/article/10.1007/s10462-020-09896-5>.
- [28] Y. Li, X. Cui, X. Yang, and G. Liu, *Frontiers in Cellular and Infection Microbiology* **14** (2024), ISSN 2235-2988, publisher: Frontiers, URL <https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2024.1482186/full>.
- [29] A. Yurtseven, S. Buyanova, A. A. Agrawal, O. O. Bochkareva, and O. V. Kalinina, *BMC Microbiology* **23**, 404 (2023), ISSN 1471-2180, URL <https://doi.org/10.1186/s12866-023-03147-7>.
- [30] Y. Liu, X. Wang, L. Wu, Z. Shi, M. Zhu, L. Ye, and P. Xu, *Microbiology Spectrum* **0**, e00065 (2025), publisher: American Society for Microbiology, URL <https://journals.asm.org/doi/10.1128/spectrum.00065-25>.
- [31] kalininalab, *ALPAR*, URL <https://github.com/kalininalab/ALPAR/tree/main>.
- [32] T. Seemann, **30**, 2068 (2014).
- [33] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, *Bioinformatics* **28**, 3150 (2012), ISSN 1367-4803, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3516142/>.
- [34] G. Tonkin-Hill, N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, J. A. Lees, R. A. Gladstone, S. Lo, C. Beaudoin, R. A. Floto, et al., *Genome Biology* **21**, 180 (2020), ISSN 1474-760X.
- [35] T. Seemann, *Snippy: rapid haploid variant calling and core genome alignment*, URL <https://github.com/tseemann/snippy>.
- [36] L. Katz, T. Griswold, S. Morrison, J. Caravas, S. Zhang, H. Bakker, X. Deng, and H. Carleton, *Journal of Open Source Software* **4**, 1762 (2019), ISSN 2475-9066, URL <https://joss.theoj.org/papers/10.21105/joss.01762>.
- [37] A. Perrin and E. P. C. Rocha, *NAR Genomics and Bioinformatics* **3**, lqaa106 (2021), ISSN 2631-9268, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7803007/>.
- [38] J. A. Lees, M. Galardini, S. D. Bentley, J. N. Weiser, and J. Corander, *Bioinformatics* **34**, 4310 (2018), ISSN 1367-4803, 1367-4811, URL <https://academic.oup.com/bioinformatics/article/34/24/4310/5047751>.
- [39] R. D. Olson, R. Assaf, T. Brettin, N. Conrad, C. Cucinell, J. J. Davis, D. M. Dempsey, A. Dickerman, E. M. Dietrich, R. W. Kenyon, et al., *Nucleic Acids Research* **51**, D678 (2023), ISSN 1362-4962.

- [40] BV-BRC Database, *Analyzing Genome Quality*, URL https://www.bv-brc.org/docs/tutorial/genome_quality_report/genome_quality_report.html.
- [41] National Center for Biotechnology Information (NCBI) GenBank, *Staphylococcus aureus subsp. aureus NCTC 8325 ASM1342v1 (GCA_000013425.1)* (2014), URL https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000013425.1/.
- [42] The UniProt Consortium, *Staphylococcus aureus (strain NCTC 8325 / PS 47)*, URL <https://www.uniprot.org/proteomes/UP000008816>.
- [43] A. F. Gillaspy, V. Worrell, J. Orvis, B. A. Roe, D. W. Dyer, and J. J. Iandolo, in *Gram-Positive Pathogens* (John Wiley & Sons, Ltd, 2006), pp. 381–412, ISBN 978-1-68367-182-4, section: 32 _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1128/9781555816513.ch32>, URL <https://onlinelibrary.wiley.com/doi/abs/10.1128/9781555816513.ch32>.
- [44] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, and Vincent Michel, *Journal of Machine Learning Research* **12**, 2825 (2011).
- [45] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, *BMC Bioinformatics* **10**, 421 (2009), ISSN 1471-2105, URL <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-421>.
- [46] B. P. Alcock, W. Huynh, R. Chalil, K. W. Smith, A. R. Raphenya, M. A. Wlodarski, A. Edalatmand, A. Petkau, S. A. Syed, K. K. Tsang, et al., *Nucleic Acids Research* **51**, D690 (2022), ISSN 0305-1048, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825576/>.
- [47] I. Letunic and P. Bork, *Nucleic Acids Research* **52**, W78 (2024), ISSN 0305-1048, 1362-4962, URL <https://academic.oup.com/nar/article/52/W1/W78/7645242>.
- [48] European Committee on Antimicrobial Susceptibility Testing, *Clinical breakpoints and dosing of antibiotics*, URL https://www.eucast.org/clinical_breakpoints.
- [49] M. Feldgarden, V. Brover, N. Gonzalez-Escalona, J. G. Frye, J. Haendiges, D. H. Haft, M. Hoffmann, J. B. Pettengill, A. B. Prasad, G. E. Tillman, et al., *Scientific Reports* **11**, 12728 (2021), ISSN 2045-2322, URL <https://www.nature.com/articles/s41598-021-91456-0>.
- [50] S. Mohammed, L. Budach, M. Feuerpfel, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch, *Information Systems* **132**, 102549 (2025), ISSN 0306-4379, publisher: Elsevier BV, URL <https://linkinghub.elsevier.com/retrieve/pii/S0306437925000341>.
- [51] L. Lüttlinger, P. Májek, S. Beisken, T. Rattei, and A. E. Posch, *Frontiers in Cellular and Infection Microbiology* **11** (2021), ISSN 2235-2988, publisher: Frontiers, URL <https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2021.610348/full>.

- [52] M. VanOeffelen, M. Nguyen, D. Aytan-Aktug, T. Brettin, E. M. Dietrich, R. W. Kenyon, D. Machi, C. Mao, R. Olson, G. D. Pusch, et al., Briefings in Bioinformatics **22**, bbab313 (2021), ISSN 1467-5463, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8575023/>.
- [53] J. R. Fitzgerald, D. E. Sturdevant, S. M. Mackie, S. R. Gill, and J. M. Musser, Proceedings of the National Academy of Sciences **98**, 8821 (2001), publisher: Proceedings of the National Academy of Sciences, URL <https://www.pnas.org/doi/10.1073/pnas.161098098>.
- [54] S. R. Harris, E. J. P. Cartwright, M. E. Török, M. T. G. Holden, N. M. Brown, A. L. Ogilvy-Stuart, M. J. Ellington, M. A. Quail, S. D. Bentley, J. Parkhill, et al., The Lancet. Infectious Diseases **13**, 130 (2013), ISSN 1474-4457.
- [55] C. Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead* (2019), arXiv:1811.10154 [stat], URL <http://arxiv.org/abs/1811.10154>.
- [56] O. Loyola-González, IEEE Access **7**, 154096 (2019), ISSN 2169-3536, URL <https://ieeexplore.ieee.org/document/8882211>.
- [57] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, Scientific Reports **8**, 4165 (2018), ISSN 2045-2322, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/s41598-018-22437-z>.
- [58] R. Caruana and A. Niculescu-Mizil, in *Proceedings of the 23rd international conference on Machine learning - ICML '06* (ACM Press, Pittsburgh, Pennsylvania, 2006), pp. 161–168, URL <http://portal.acm.org/citation.cfm?doid=1143844.1143865>.
- [59] H. M and S. M.N, International Journal of Data Mining & Knowledge Management Process **5**, 01 (2015), ISSN 2231-007X, 2230-9608, publisher: Academy and Industry Research Collaboration Center (AIRCC), URL <http://www.airconline.com/ijdkp/V5N2/5215ijdkp01.pdf>.
- [60] M. Ghanem, A. K. Ghaith, V. G. El-Hajj, A. Bhandarkar, A. de Giorgio, A. Elmi-Terander, and M. Bydon, Brain Sciences **13**, 1723 (2023), ISSN 2076-3425, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10741524/>.
- [61] T. Viering and M. Loog, IEEE Transactions on Pattern Analysis and Machine Intelligence **45**, 7799 (2023), ISSN 1939-3539, URL <https://ieeexplore.ieee.org/document/9944190/>.
- [62] A. Jain, R. Duin, and J. Mao, IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 4 (2000), ISSN 1939-3539, URL <https://ieeexplore.ieee.org/document/824819/>.
- [63] C. Aliferis and G. Simon, in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, edited by G. J. Simon and C. Aliferis (Springer International Publishing, Cham, 2024), pp. 477–524, ISBN 978-3-031-39355-6, URL https://doi.org/10.1007/978-3-031-39355-6_10.

- [64] X. Guan and H. Burton, Structures **46**, 17 (2022), ISSN 2352-0124, URL <https://www.sciencedirect.com/science/article/pii/S2352012422009018>.
- [65] X. Yuan, S. Liu, W. Feng, and G. Dauphin, Remote Sensing **15**, 5203 (2023), ISSN 2072-4292, publisher: Multidisciplinary Digital Publishing Institute, URL <https://www.mdpi.com/2072-4292/15/21/5203>.
- [66] J. I. Kim, A. Manuele, F. Maguire, R. Zaheer, T. A. McAllister, and R. G. Beiko, Canadian Journal of Microbiology **70**, 446 (2024), ISSN 0008-4166, 1480-3275, publisher: Canadian Science Publishing, URL <https://cdnscepub.com/doi/10.1139/cjm-2024-0049>.
- [67] S. O. Jensen and B. R. Lyon, Future Microbiology **4**, 565 (2009), ISSN 17460913, num Pages: 565-82 Place: London, United Kingdom Publisher: Future Medicine Ltd, URL <https://www.proquest.com/docview/223078806/abstract/FB1CED8D80DD4E0CPQ/1>.
- [68] T. Naas, S. Oueslati, R. A. Bonnin, M. L. Dabos, A. Zavala, L. Dortet, P. Retailleau, and B. I. Iorga, Journal of Enzyme Inhibition and Medicinal Chemistry **32**, 917 (2017), ISSN 1475-6366, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6445328/>.
- [69] H. Lade and J.-S. Kim, Antibiotics **12**, 1362 (2023), ISSN 2079-6382, number: 9 Publisher: Multidisciplinary Digital Publishing Institute, URL <https://www.mdpi.com/2079-6382/12/9/1362>.
- [70] V. Gor, R. L. Ohniwa, and K. Morikawa, Microorganisms **9**, 244 (2021), ISSN 2076-2607, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7911514/>.
- [71] S.-J. Rowland and K. Dyke, FEMS Microbiology Letters **50**, 253 (1988), ISSN 0378-1097, 1574-6968, publisher: Oxford University Press (OUP), URL <https://academic.oup.com/femsle/article-lookup/doi/10.1111/j.1574-6968.1988.tb02947.x>.
- [72] B. R. Lyon and R. Skurray, Microbiol. Rev. **51** (1987).
- [73] R. Cao, H. Su, Z. Wei, Z. He, T. Pan, Y. Li, and B. Sun, International journal of medical microbiology: IJMM **315**, 151624 (2024), ISSN 1618-0607.
- [74] P. Blanco, S. Hernando-Amado, J. A. Reales-Calderon, F. Corona, F. Lira, M. Alcalde-Rico, A. Bernardini, M. B. Sanchez, and J. L. Martinez, Microorganisms **4**, 14 (2016), ISSN 2076-2607, number: 1 Publisher: Multidisciplinary Digital Publishing Institute, URL <https://www.mdpi.com/2076-2607/4/1/14>.
- [75] B. Martínez, A. Rodríguez, and J. E. Suárez, Microbiology **146**, 949 (2000), ISSN 1465-2080, publisher: Microbiology Society, URL <https://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-146-4-949>.
- [76] P. Joseph, A. Guiseppi, A. Sorokin, and F. Denizot, Microbiology **150**, 2609 (2004), ISSN 1465-2080, publisher: Microbiology Society, URL <https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.27155-0>.
- [77] K. P. Ha and A. M. Edwards, Microbiology and Molecular Biology Reviews **85**, e00091 (2021), ISSN 1092-2172, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8483670/>.

- [78] K. Bush and P. A. Bradford, Cold Spring Harbor Perspectives in Medicine **6**, a025247 (2016), ISSN 2157-1422, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4968164/>.
- [79] A. J. Sabat, E. Bathoorn, K. Becker, V. Akkerboom, M. Miskoski, T. Durfee, and A. W. Friedrich, Journal of Antimicrobial Chemotherapy **76**, 1986 (2021), ISSN 0305-7453, 1460-2091, URL <https://academic.oup.com/jac/article/76/8/1986/6276546>.
- [80] C. Ryffel, R. Bucher, F. H. Kayser, and B. Berger-Bächi, Journal of Bacteriology **173**, 7416 (1991), ISSN 0021-9193, 1098-5530, URL <https://journals.asm.org/doi/10.1128/jb.173.23.7416-7422.1991>.
- [81] Y. Matsumoto, J. Yasukawa, M. Ishii, Y. Hayashi, S. Miyazaki, and K. Sekimizu, Scientific Reports **6**, 22894 (2016), URL <https://pmc-ncbi-nlm-nih-gov.uaccess.univie.ac.at/articles/PMC4790635/>.
- [82] Z. Zong, BMC Microbiology **13**, 64 (2013), ISSN 1471-2180, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3637587/>.
- [83] D. E. Waldron and J. A. Lindsay, Journal of Bacteriology **188**, 5578 (2006), ISSN 0021-9193, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1540015/>.
- [84] E. Thulin and D. I. Andersson, Antimicrobial agents and chemotherapy **63** (2019), ISSN 1098-6596, publisher: Antimicrob Agents Chemother, URL <https://pubmed.ncbi.nlm.nih.gov/31332059/>.
- [85] M. Korem, Y. Gov, M. D. Kiran, and N. Balaban, Infection and Immunity **73**, 6220 (2005), ISSN 0019-9567, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1230921/>.
- [86] J. Stec-Niemczyk, K. Pustelny, M. Kisielewska, M. Bista, K. T. Boulware, H. R. Stennicke, I. B. Thogersen, P. S. Daugherty, J. J. Enghild, K. Baczynski, et al., The Biochemical Journal **419**, 555 (2009), ISSN 1470-8728.
- [87] M. Schroeder, B. D. Brooks, and A. E. Brooks, Genes **8**, 39 (2017), ISSN 2073-4425, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5295033/>.
- [88] H. Zhang, D. Li, L. Zhao, J. Fleming, N. Lin, T. Wang, Z. Liu, C. Li, N. Galwey, J. Deng, et al., Nature Genetics **45**, 1255 (2013), ISSN 1546-1718, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/ng.2735>.
- [89] P. K. Párraga Solórzano, T. S. Bastille, J. N. Radin, and T. E. Kehl-Fie, mBio **14**, e0322322 (2023), ISSN 2150-7511.
- [90] R. Pluta, D. R. Boer, F. Lorenzo-Díaz, S. Russi, H. Gómez, C. Fernández-López, R. Pérez-Luque, M. Orozco, M. Espinosa, and M. Coll, Proceedings of the National Academy of Sciences of the United States of America **114**, E6526 (2017), ISSN 0027-8424, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5559014/>.
- [91] M. Mojumdar and S. A. Khan, **170**, 5522 (1988), URL <https://journals-asm-org.uaccess.univie.ac.at/doi/epdf/10.1128/jb.170.12.5522-5528.1988>.

- [92] D. F. Balson and W. V. Shaw, *Plasmid* **24**, 74 (1990), ISSN 0147-619X.
- [93] E. C. Woods and S. M. McBride, *Microbes and infection* **19**, 238 (2017), ISSN 1286-4579, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5403605/>.
- [94] L.-C. Chu, P. Arede, W. Li, E. C. Urdaneta, I. Ivanova, S. W. McKellar, J. C. Wills, T. Fröhlich, A. von Kriegsheim, B. M. Beckmann, et al., *Nature Communications* **13**, 2883 (2022), ISSN 2041-1723, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/s41467-022-30553-8>.
- [95] J. H. Morisaki, P. A. Smith, S. V. Date, K. K. Kajihara, C. L. Truong, Z. Modrusan, D. Yan, J. Kang, M. Xu, I. M. Shah, et al., *mBio* **7**, e00412 (2016), ISSN 2150-7511, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013292/>.
- [96] K. M. Cregg, E. I. Wilding, and M. T. Black, **178**, 5712 (1996), URL <https://journals.asm.org/doi/epdf/10.1128/jb.178.19.5712-5718.1996>.
- [97] M. W. Vetting, L. P. S. de Carvalho, M. Yu, S. S. Hegde, S. Magnet, S. L. Roderick, and J. S. Blanchard, *Archives of Biochemistry and Biophysics* **433**, 212 (2005), ISSN 0003-9861, URL <https://www.sciencedirect.com/science/article/pii/S000398610400503X>.
- [98] B. Chakrakodi, S. Prabhakara, S. Nagaraj, J. Etienne, and G. Arakere, *Advances in Microbiology* **04**, 133 (2014), ISSN 2165-3402, 2165-3410, URL <http://www.scirp.org/journal/doi.aspx?DOI=10.4236/aim.2014.42018>.
- [99] S. Sreedharan, M. Oram, B. Jensen, L. R. Peterson, and L. M. Fisher, *Journal of Bacteriology* **172**, 7260 (1990), ISSN 0021-9193, 1098-5530, URL <https://journals.asm.org/doi/10.1128/jb.172.12.7260-7262.1990>.
- [100] J. E. Fitzgibbon, J. F. John, J. L. Delucia, and D. T. Dubin, *Antimicrobial Agents and Chemotherapy* **42**, 2122 (1998), ISSN 0066-4804, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC105881/>.
- [101] Q. C. Truong-Bolduc, Y. Wang, J. L. Reedy, J. M. Vyas, and D. C. Hooper, *Antimicrobial Agents and Chemotherapy* **66**, e01845 (2022), publisher: American Society for Microbiology, URL <https://journals-asm-org.uaccess.univie.ac.at/doi/10.1128/aac.01845-21>.
- [102] S. Karimaei, B. S. Kalani, N. Shahrokhi, R. Mashhadi, and M. R. Pourmand, *Iranian Journal of Microbiology* **13**, 204 (2021), ISSN 2008-3289, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8408035/>.
- [103] M. S. Ramirez and M. E. Tolmasky, *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy* **13**, 151 (2010), ISSN 1368-7646, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992599/>.
- [104] J.-i. Sekiguchi, P. Tharavichitkul, T. Miyoshi-Akiyama, V. Chupia, T. Fujino, M. Araake, A. Irie, K. Morita, T. Kuratsuji, and T. Kirikae, *Antimicrobial Agents and Chemotherapy* **49**, 3948 (2005), ISSN 0066-4804, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1195413/>.

- [105] M.-T. Nguyen, A. Luqman, K. Bitschar, T. Hertlein, J. Dick, K. Ohlsen, B. Bröker, B. Schitteck, and F. Götz, *International journal of medical microbiology: IJMM* **308**, 653 (2018), ISSN 1618-0607.
- [106] K. L. Nawrocki, E. K. Crispell, and S. M. McBride, *Antibiotics* **3**, 461 (2014), ISSN 2079-6382, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4239024/>.
- [107] A. Kouvela, J. R. Jaramillo Ponce, N. Giarimoglou, J. Chicher, S. Marzi, C. Stathopoulos, and V. Stamatopoulou, *Nucleic Acids Research* **53**, gkaf51399 (2025), ISSN 0305-1048, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12256121/>.
- [108] H. Veiga, A. Jousselin, S. Schäper, B. M. Saraiva, L. B. Marques, P. Reed, J. Wilton, P. M. Pereira, S. R. Filipe, and M. G. Pinho, *The EMBO journal* **42**, e112140 (2023), ISSN 1460-2075.
- [109] E. E. Price and J. M. Boyd, *Trends in microbiology* **28**, 821 (2020), ISSN 0966-842X, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7494629/>.
- [110] J. B. Parsons, J. Yao, M. W. Frank, and C. O. Rock, *Antimicrobial Agents and Chemotherapy* **59**, 849 (2015), ISSN 0066-4804, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4335864/>.
- [111] F. Askarian, N. M. van Sorge, M. Sangvik, F. C. Beasley, J. R. Henriksen, J. U. Sollid, J. A. van Strijp, V. Nizet, and M. Johannessen, *Journal of Innate Immunity* **6**, 485 (2014), ISSN 1662-811X, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4198549/>.
- [112] I. Mir-Sanchis, Y. Z. Pigli, and P. A. Rice, *Structure* **26**, 1144 (2018), ISSN 0969-2126, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6084467/>.
- [113] J. Lannergård, T. Norström, and D. Hughes, *Antimicrobial Agents and Chemotherapy* **53**, 2059 (2009), ISSN 0066-4804, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2681530/>.
- [114] G. A. Beggs, R. G. Brennan, and M. Arshad, *Protein Science : A Publication of the Protein Society* **29**, 647 (2020), ISSN 0961-8368, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7020996/>.
- [115] P. E. Chen and B. J. Shapiro, *Current Opinion in Microbiology* **25**, 17 (2015), ISSN 1369-5274, URL <https://www.sciencedirect.com/science/article/pii/S1369527415000375>.
- [116] W. Muselmani, N. Kashif-Khan, C. Bagnéris, R. Santangelo, M. A. Williams, and R. Savva, *Viruses* **15**, 1348 (2023), ISSN 1999-4915, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10303552/>.
- [117] N. McCallum, B. Berger-Bächi, and M. M. Senn, *International Journal of Medical Microbiology* **300**, 118 (2010), ISSN 1438-4221, URL <https://www.sciencedirect.com/science/article/pii/S1438422109001088>.

- [118] B. K. Duncan, P. A. Rockstroh, and H. R. Warner, *Journal of Bacteriology* **134**, 1039 (1978), ISSN 0021-9193, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC222353/>.
- [119] S. Willing, E. Dyer, O. Schneewind, and D. Missiakas, *The Journal of Biological Chemistry* **295**, 13664 (2020), ISSN 0021-9258, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7521636/>.
- [120] A. Belloni Fortina and L. Neri, *Giornale italiano di dermatologia e venereologia* **150**, 321 (2015), ISSN 1827-1820.
- [121] Q. C. Truong-Bolduc, G. R. Bolduc, H. Medeiros, J. M. Vyas, Y. Wang, and D. C. Hooper, *Infection and Immunity* **83**, 4362 (2015), ISSN 0019-9567, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4598397/>.
- [122] N. Kadeřábková, A. J. S. Mahmood, and D. A. I. Mavridou, *npj Antimicrobials and Resistance* **2**, 1 (2024), ISSN 2731-8745, publisher: Nature Publishing Group, URL <https://www.nature.com/articles/s44259-024-00051-6>.
- [123] E. Gómez-Sanz, K. Kadlec, A. T. Feßler, M. Zarazaga, C. Torres, and S. Schwarz, *Antimicrobial Agents and Chemotherapy* **57**, 3275 (2013), ISSN 0066-4804, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3697343/>.
- [124] OpenAI, *ChatGPT-4o* (2022), URL <https://openai.com/index/chatgpt/>.
- [125] University of Vienna, *Life Science Compute Cluster* (2025), URL <https://lisc.univie.ac.at/>.
- [126] GitHub, *Copilot* (2025), URL <https://github.com/features/copilot>.
- [127] Digital Science, *Overleaf* (2025), URL <https://www.overleaf.com>.
- [128] Free Software Foundation, *Inkscape* (1989), URL <https://inkscape.org/de/>.

Supplementary Material

1280.51558	1280.51559	1280.51560	1280.51739	1280.51740	1280.51741	1280.51742
1280.51743	1280.51744	1280.51745	1280.51746	1280.51747	1280.51748	1280.51749
1280.51750	1280.51751	1280.51752	1280.51753	1280.51754	1280.51761	1280.51763
1280.51764	1280.51766	1280.51767	1280.51768	1280.51769	1280.51770	1280.51771
1280.51772	1280.51773	1280.51774	1280.51775	1280.51776	1280.51777	1280.51778
1280.51779	1280.51780	1280.51784	1280.51785	1280.51786	1280.51787	1280.51788
1280.51789	1280.51790	1280.51791	1280.51792	1280.51793	1280.51794	1280.51795
1280.51796	1280.51797	1280.51798	1280.51799	1280.51800	1280.51801	1280.51803
1280.51804	1280.51805	1280.51806	1280.51807	1280.51808	1280.51809	1280.51810
1280.51812	1280.51813	1280.51814	1280.51815	1280.51816	1280.51817	1280.51818
1280.51819	1280.51820	1280.51821	1280.51823	1280.51824	1280.51825	1280.51826
1280.51828	1280.51829	1280.51830	1280.51831	1280.51832	1280.51833	1280.51834
1280.51835	1280.51836	1280.51837	1280.51838	1280.51839	1280.51840	1280.51841
1280.51842	1280.51843	1280.51844	1280.51845	1280.51846	1280.51847	1280.51848
1280.51849	1280.51850	1280.51852	1280.51853	1280.51855	1280.51856	1280.51857
1280.51858	1280.51859	1280.51860	1280.51861	1280.51862	1280.51863	1280.51864
1280.51865	1280.51866	1280.51867	1280.51868	1280.51869	1280.51870	1280.51871
1280.51872	1280.51874	1280.51875	1280.51876	1280.51879	1280.51880	1280.51881
1280.51882	1280.51883	1280.51884	1280.51885	1280.51886	1280.51887	1280.51888
1280.51889	1280.51890	1280.51891	1280.51893	1280.51895	1280.51896	1280.51897
1280.51898	1280.51899	1280.51900	1280.51901	1280.51902	1280.51903	1280.51904
1280.51905	1280.51906	1280.51907	1280.51908	1280.51909	1280.51910	1280.51911
1280.51912	1280.51913	1280.51914	1280.51915	1280.51916	1280.51917	1280.51918
1280.51919	1280.51920	1280.51921	1280.51922	1280.51923	1280.51924	1280.51925
1280.51926	1280.51927	1280.51928	1280.51929	1280.51930	1280.51931	1280.51932
1280.51933	1280.51934	1280.51935	1280.51936	1280.51937	1280.51938	1280.51939
1280.51940	1280.51941	1280.52134	1280.52135	1280.52136	1280.52137	1280.52138
1280.52139	1280.52140	1280.52141	1280.52142	1280.52143	1280.52144	1280.52145
1280.52146	1280.52147	1280.52148	1280.52149	1280.52150	1280.52151	1280.52152
1280.52153	1280.52154	1280.52155	1280.52156	1280.52157	1280.52158	1280.52159
1280.52160	1280.52161	1280.52162	1280.52163	1280.52164	1280.52165	1280.52166
1280.52167	1280.52168	1280.52169	1280.52170	1280.52171	1280.52172	1280.52173
1280.52174	1280.52175	1280.52176	1280.52177	1280.52178	1280.52179	1280.52180
1280.52181	1280.52182	1280.52183	1280.52184	1280.52185	1280.52186	1280.52187
1280.52188	1280.52189	1280.52190	1280.52191	1280.52192	1280.52193	1280.52194
1280.52195	1280.52196	1280.52197	1280.52198	1280.52199	1280.52200	1280.52201
1280.52202	1280.52203	1280.52204	1280.52205	1280.52206	1280.52207	1280.52208
1280.52209	1280.52210	1280.52211	1280.52212	1280.52213	1280.52214	1280.52215
1280.52216	1280.52217	1280.52218	1280.52219	1280.52220	1280.52221	1280.52222
1280.52223	1280.52224	1280.52225	1280.52226	1280.52227	1280.52228	1280.52229
1280.52230	1280.52231	1280.52232	1280.52233			

Table 6.1: Faulty genomes excluded from this work.

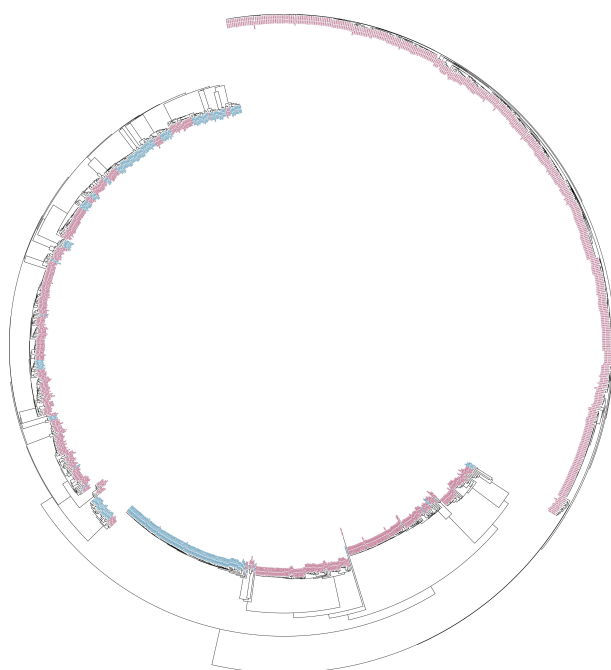


Figure 6.1: Phylogenetic tree of the training set for cefoxitin. Resistant genomes are colored red, susceptible ones are colored blue.

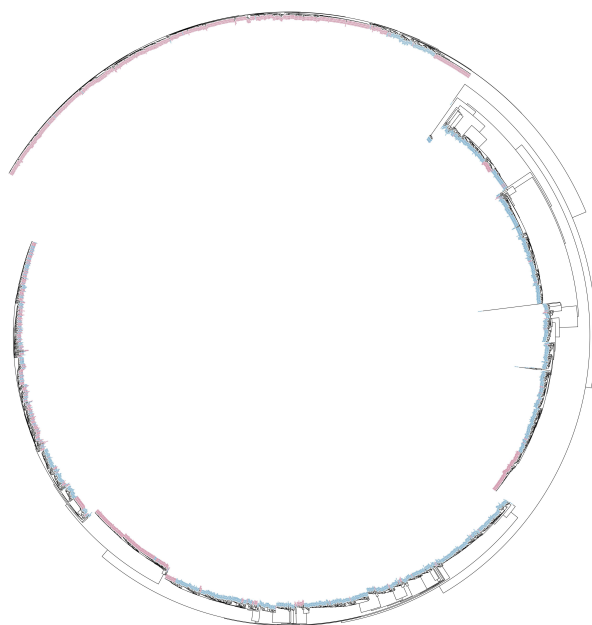


Figure 6.2: Phylogenetic tree of the training set for ciprofloxacin. Resistant genomes are colored red, susceptible ones are colored blue.

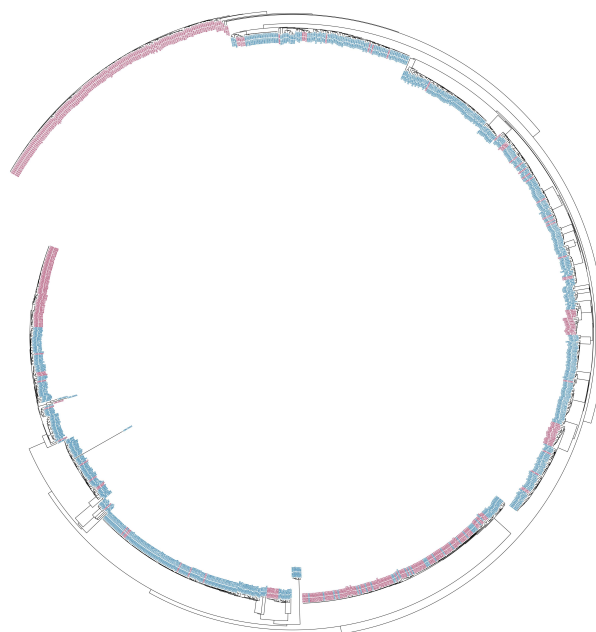


Figure 6.3: Phylogenetic tree of the training set for clindamycin. Resistant genomes are colored red, susceptible ones are colored blue.

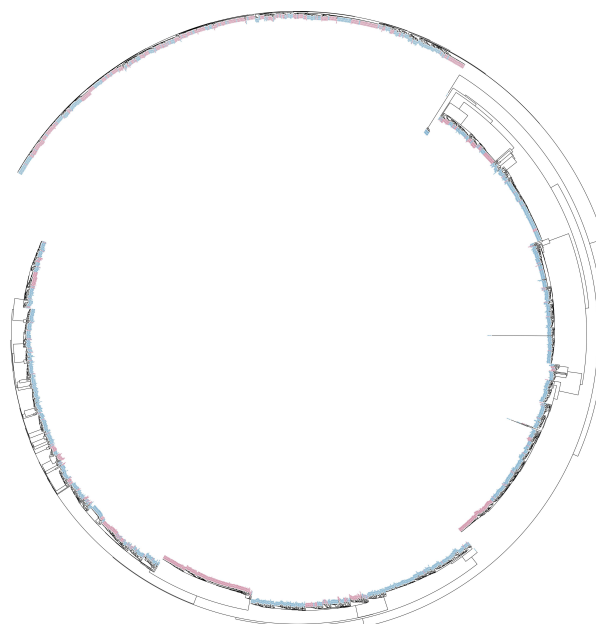


Figure 6.4: Phylogenetic tree of the training set for erythromycin. Resistant genomes are colored red, susceptible ones are colored blue.

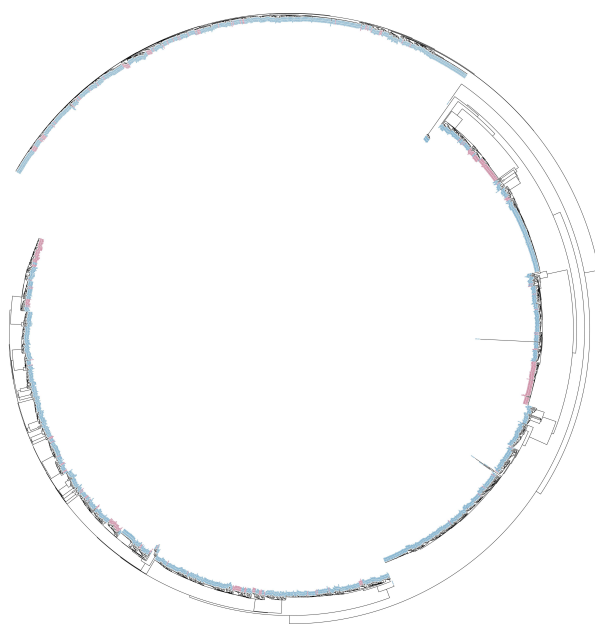


Figure 6.5: Phylogenetic tree of the training set for gentamicin. Resistant genomes are colored red, susceptible ones are colored blue.

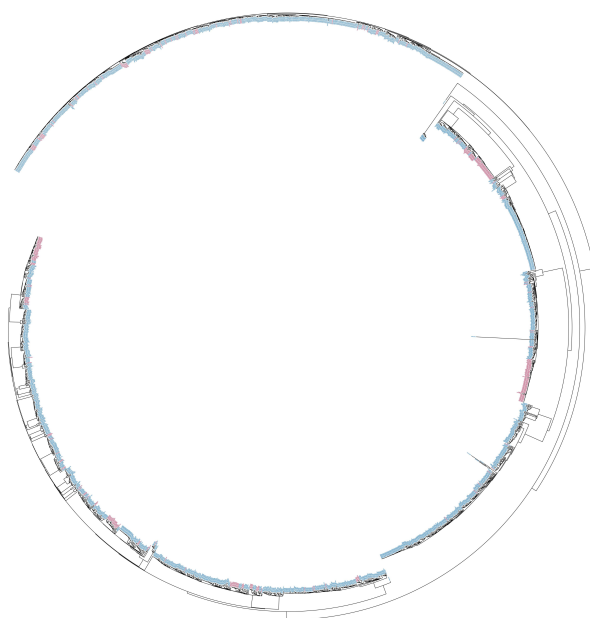


Figure 6.6: Phylogenetic tree of the training set for fusidic acid. Resistant genomes are colored red, susceptible ones are colored blue.

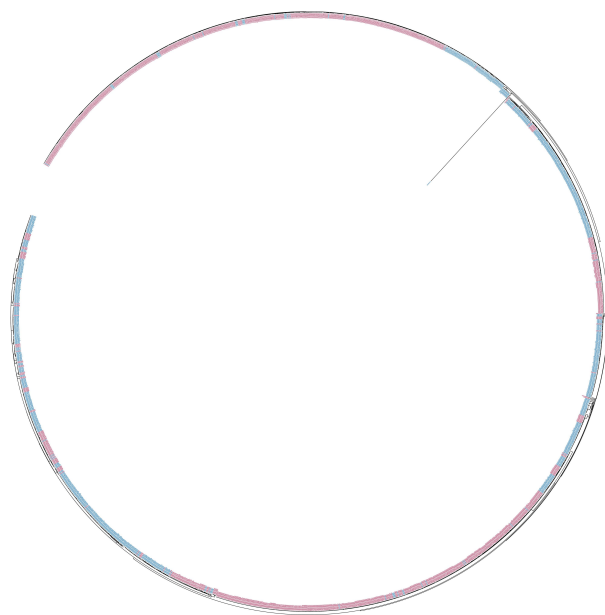


Figure 6.7: Phylogenetic tree of the training set for methicillin. Resistant genomes are colored red, susceptible ones are colored blue.

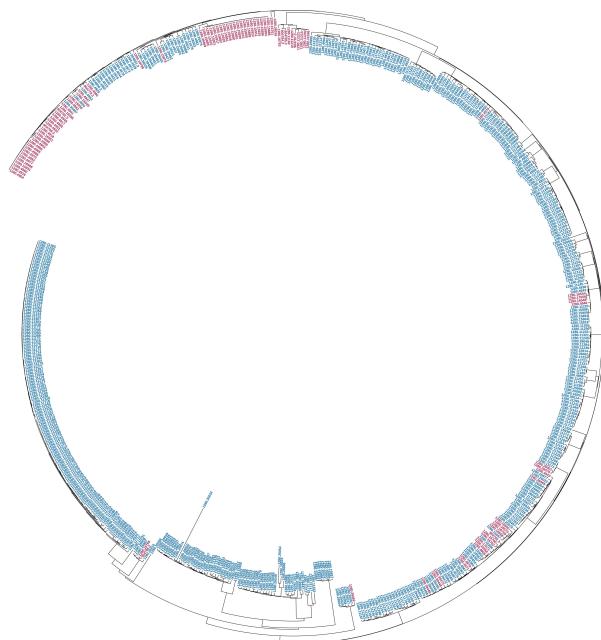


Figure 6.8: Phylogenetic tree of the training set for oxacillin. Resistant genomes are colored red, susceptible ones are colored blue.

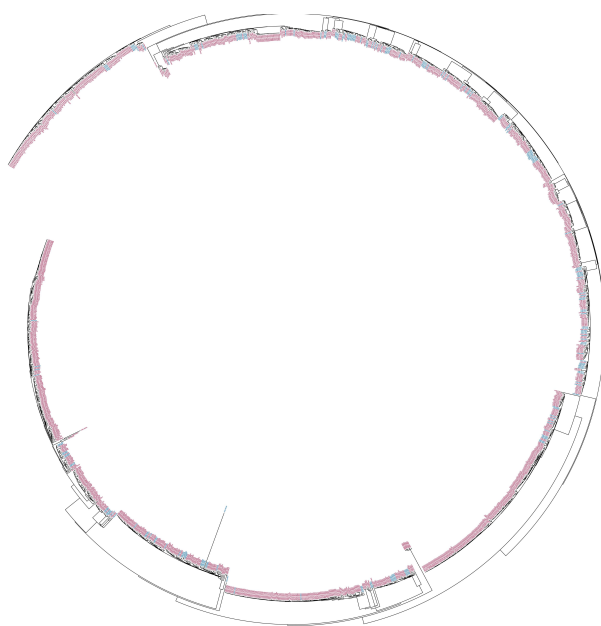


Figure 6.9: Phylogenetic tree of the training set for penicillin. Resistant genomes are colored red, susceptible ones are colored blue.

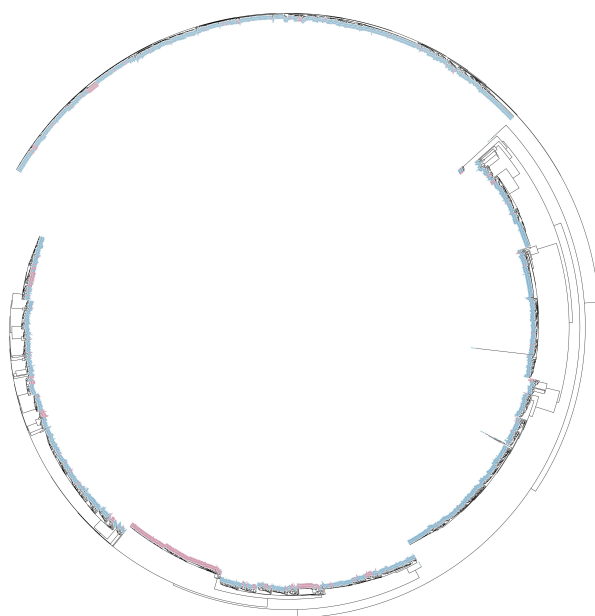


Figure 6.10: Phylogenetic tree of the training set for tetracycline. Resistant genomes are colored red, susceptible ones are colored blue.

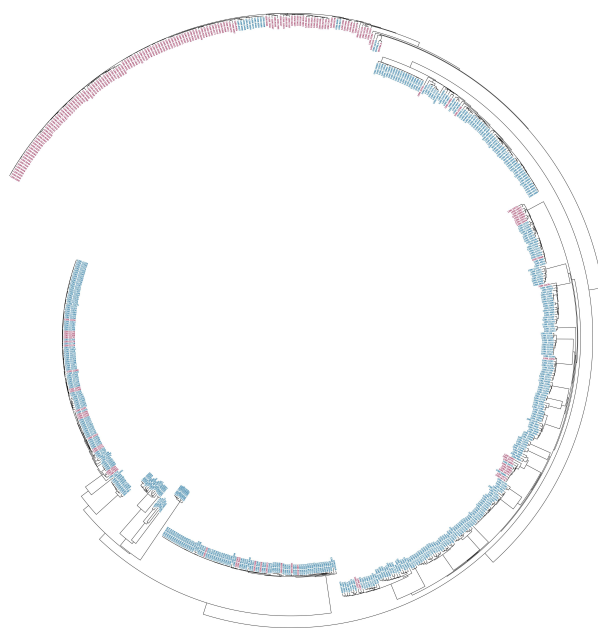


Figure 6.11: Phylogenetic tree of the training set for co-trimoxazole. Resistant genomes are colored red, susceptible ones are colored blue.

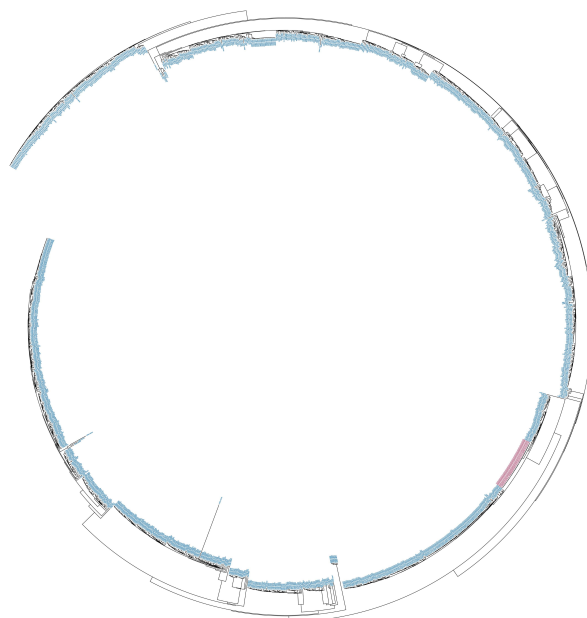


Figure 6.12: Phylogenetic tree of the training set for vancomycin. Resistant genomes are colored red, susceptible ones are colored blue.

isolate ID	patient ID	antibiotic tested	ALPAR prediction	ABRicate prediction
iso01.06	p022	ERY	R	S
iso01.09	p022	ERY	R	R
iso01.09	p022	PEN	S	S
iso01.23	p022	ERY	S	R
iso02.09	p022	ERY	R	R
iso05.05	p023	ERY	S	S
iso05.05	p023	PEN	R	S
iso05.20	p023	PEN	R	R
iso06.07	p009	ERY	S	S
iso06.07	p009	PEN	S	S
iso06.11	p009	PEN	R	R
iso07.02	p009	PEN	S	S
iso07.07	p009	ERY	S	R
iso07.07	p009	PEN	R	R
iso11.12	p025	PEN	R	S
iso12.08	p026	ERY	S	S
iso12.08	p026	PEN	S	S
iso17.02	p021	PEN	R	R
iso17.14	p021	ERY	S	R
iso18.01	p021	ERY	S	R
iso18.01	p021	PEN	R	R
iso19.13	p029	ERY	S	S
iso19.13	p029	PEN	S	S
iso20.08	p029	ERY	S	S

Table 6.2: Predictions from ALPAR and ABRicate for isolates that were experimentally tested. (ERY = erythromycin, PEN = penicillin, R = resistant, S = susceptible)

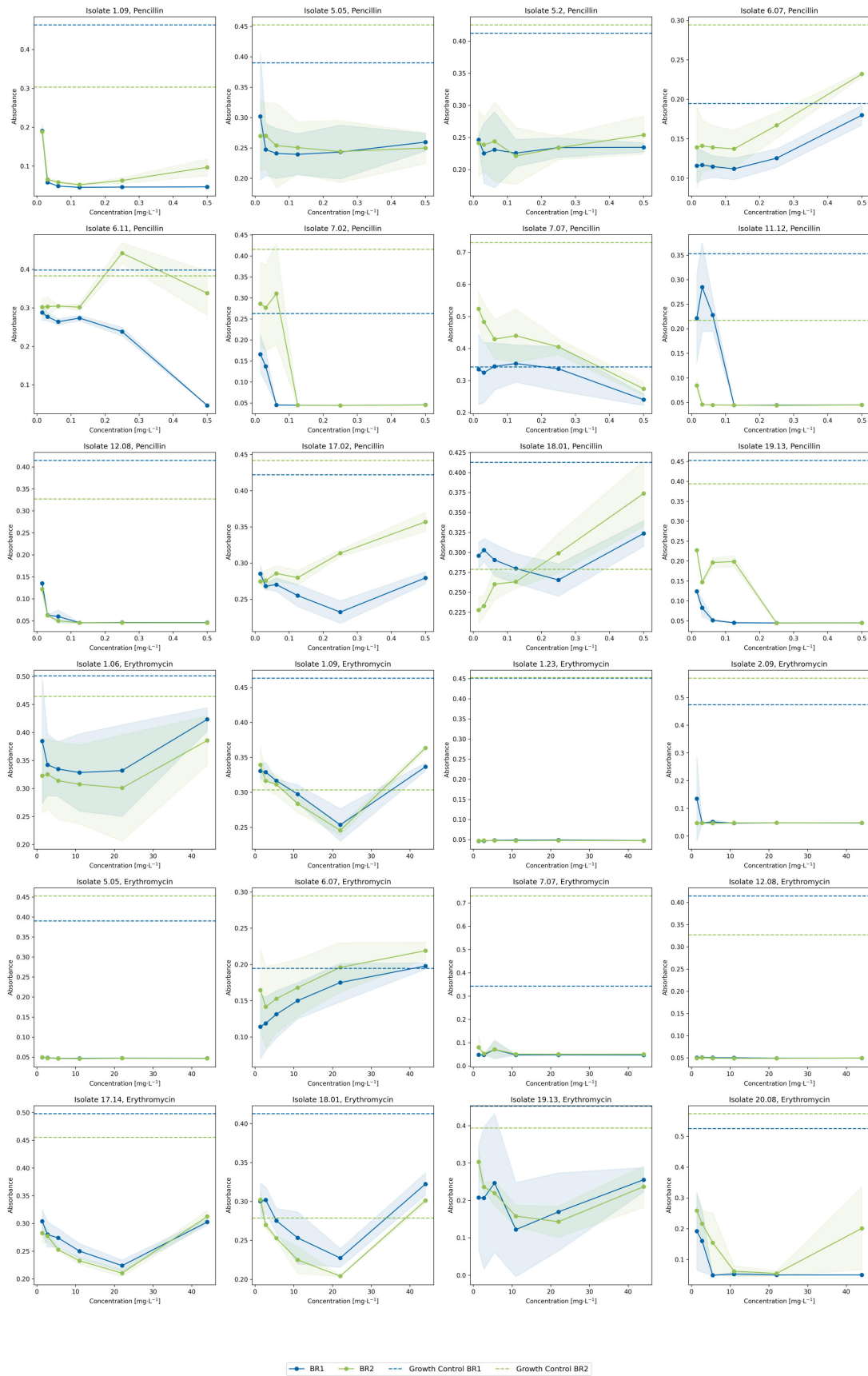


Figure 6.13: Comparison of the two biological replicas for every isolate and both antibiotics. The growth control is shown as a dotted line.

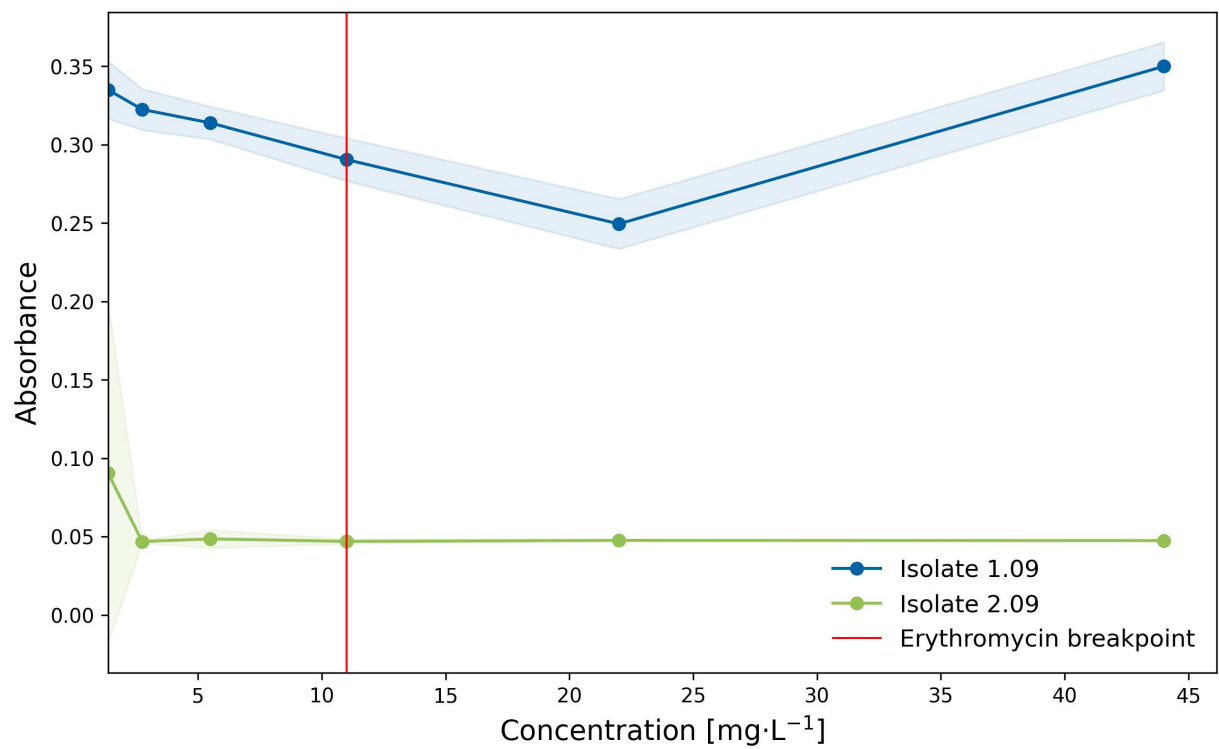


Figure 6.14: Growth inhibition curve of prediction patterns (ALPAR: resistant, ABRicate: resistant) for erythromycin for both biological replicates.

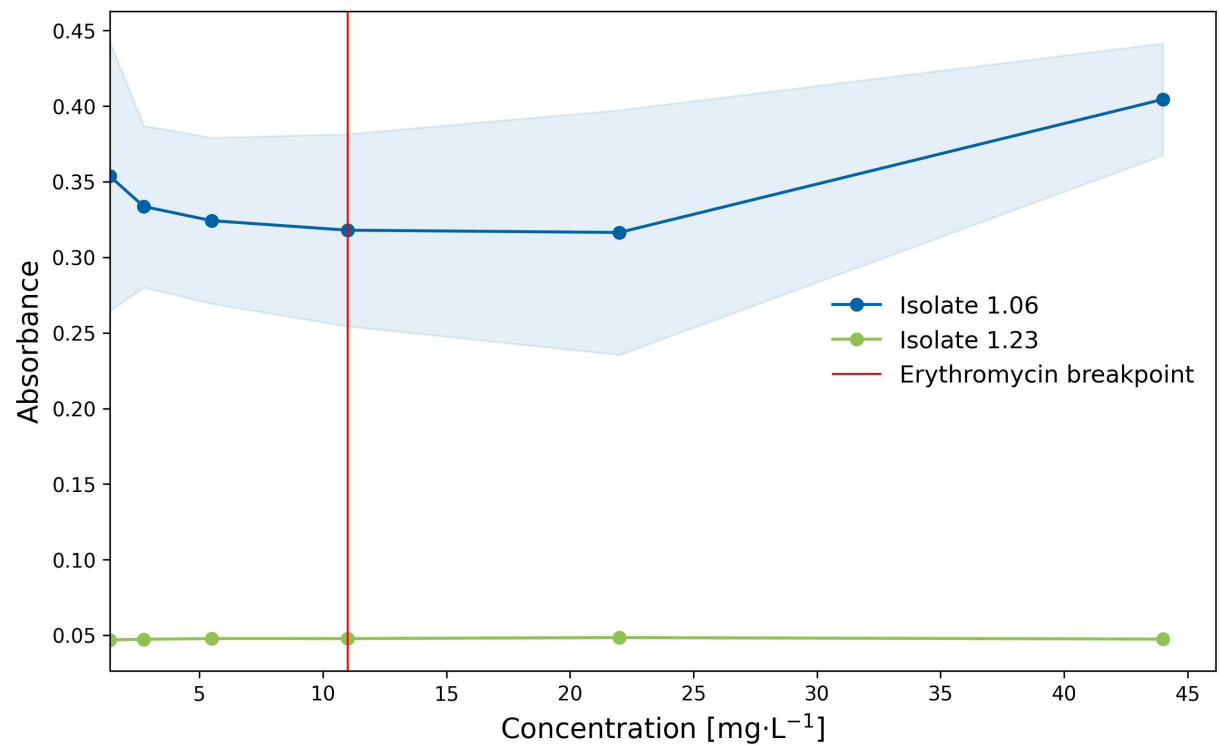


Figure 6.15: Growth inhibition curve of prediction patterns (ALPAR: resistant, ABRicate: susceptible) for erythromycin for both biological replicates.

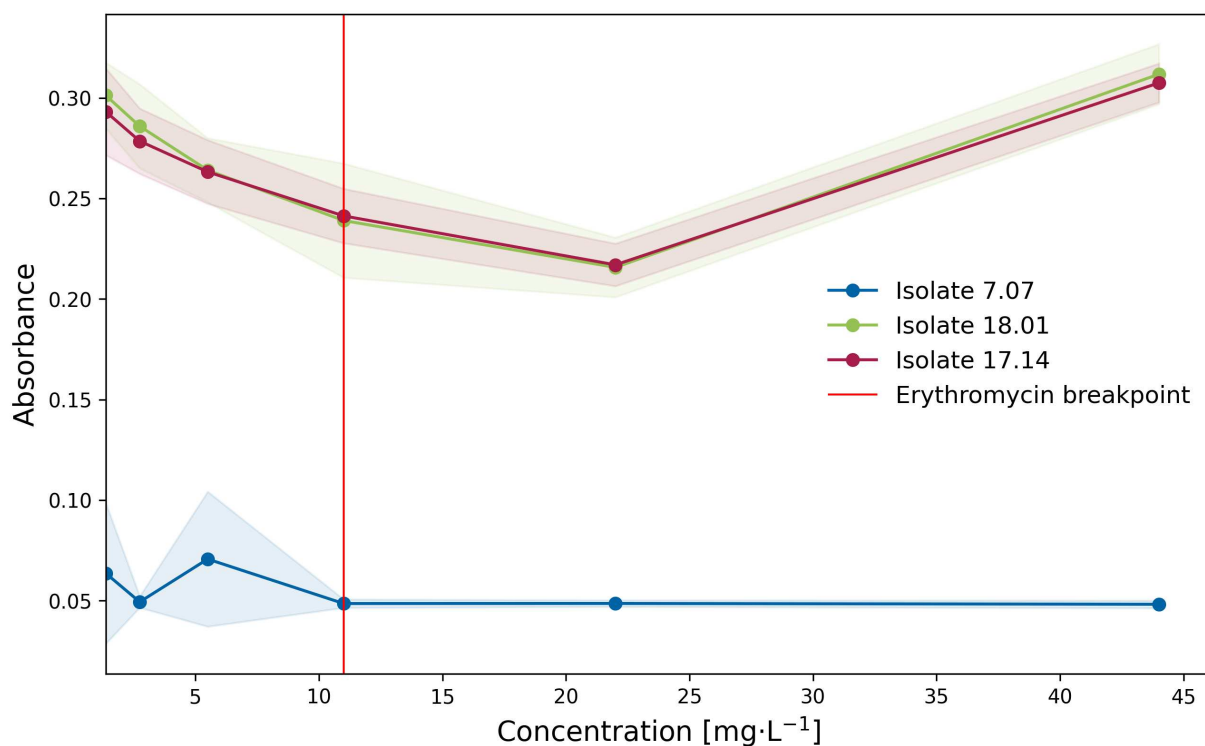


Figure 6.16: Growth inhibition curve of prediction patterns (ALPAR: susceptible, ABRicate: resistant) for erythromycin for both biological replicates.

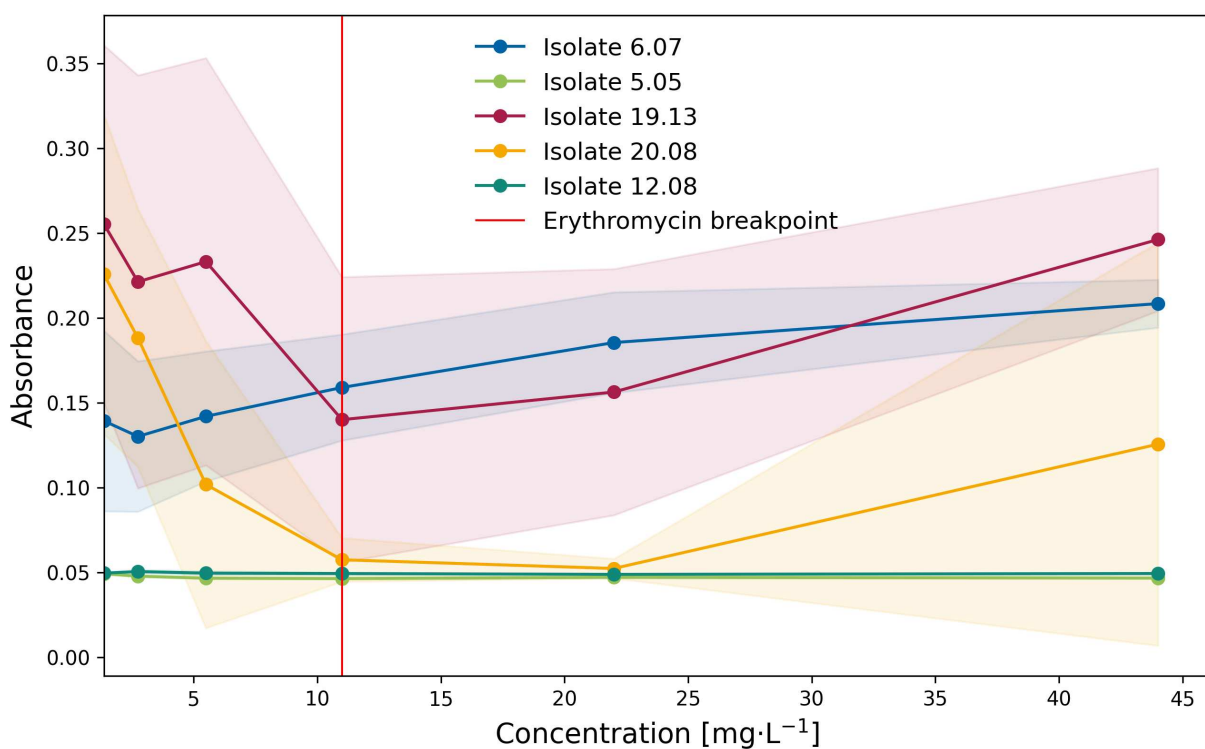


Figure 6.17: Growth inhibition curve of prediction patterns (ALPAR: susceptible, ABRicate: susceptible) for erythromycin for both biological replicates.

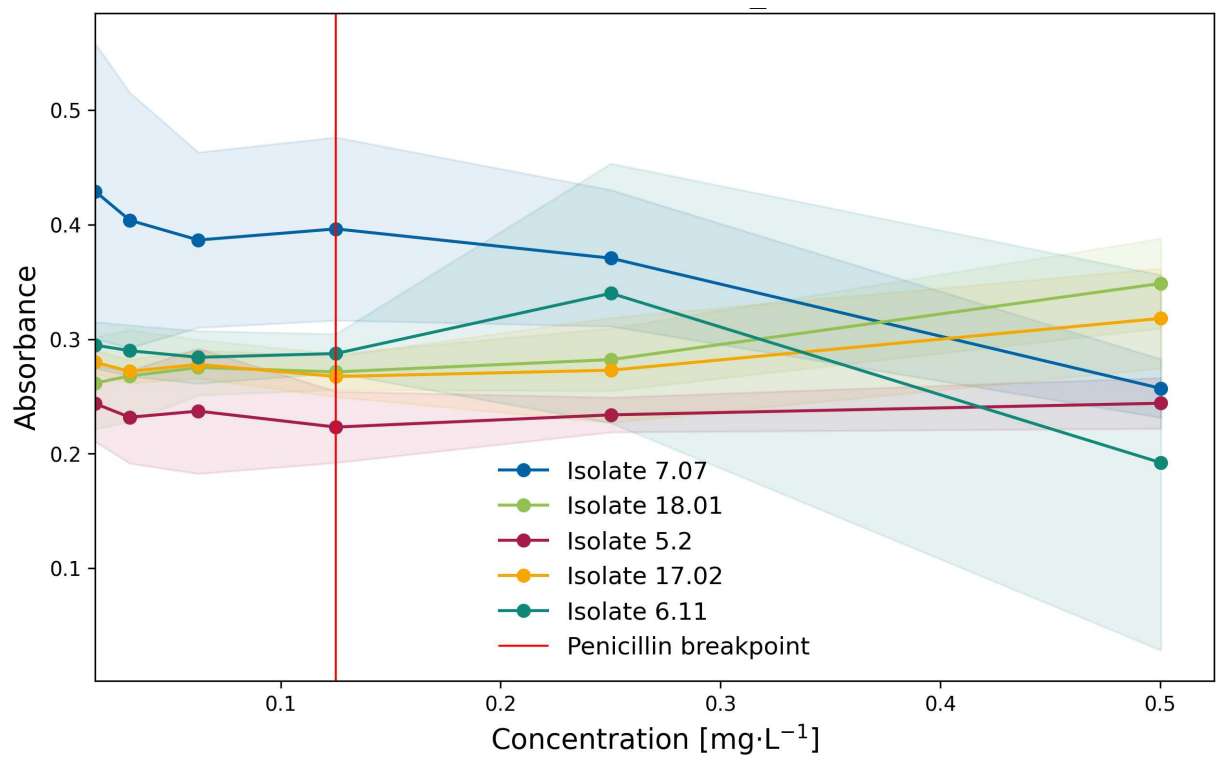


Figure 6.18: Growth inhibition curve of prediction patterns (ALPAR: resistant, ABRicate: resistant) for penicillin for both biological replicates.

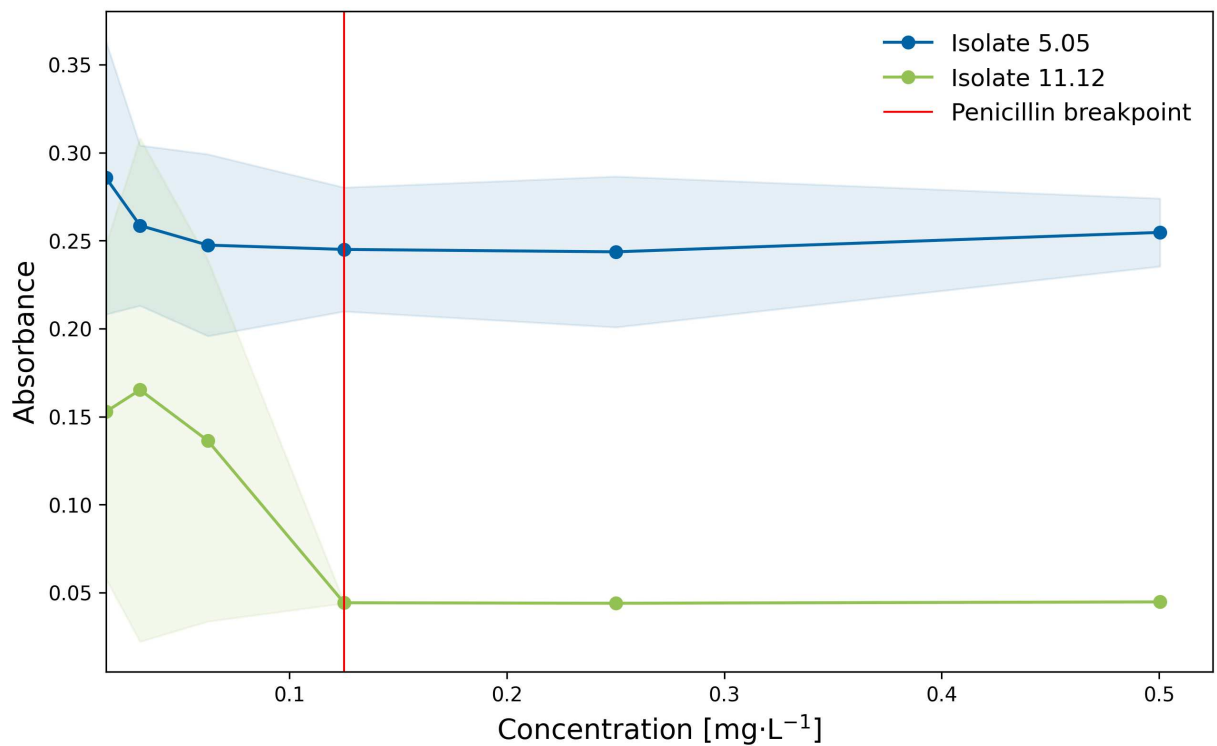


Figure 6.19: Growth inhibition curve of prediction patterns (ALPAR: resistant, ABRicate: susceptible) for penicillin for both biological replicates.

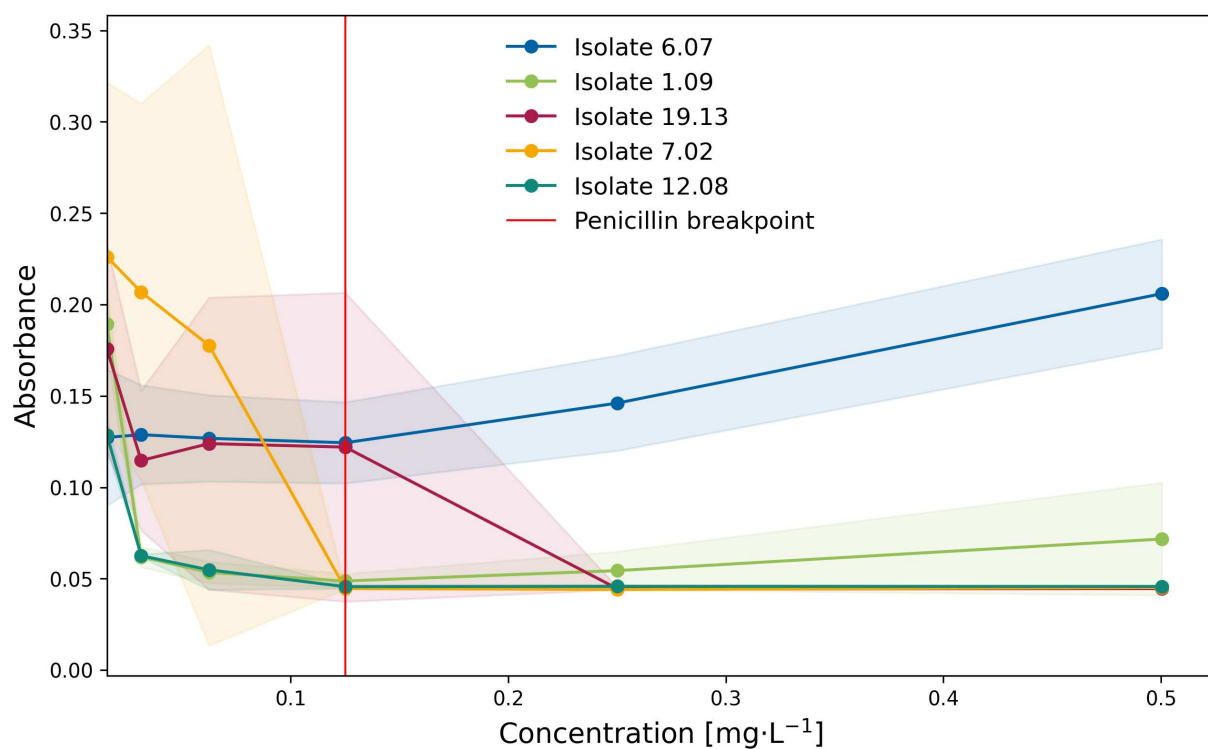


Figure 6.20: Growth inhibition curve of prediction patterns (ALPAR: susceptible, ABRicate: susceptible) for penicillin for both biological replicates.