PREDICTION OF ANTIMICROBIAL RESISTANCE PHENOTYPES FROM GENOTYPE

PREDICTION OF ANTIMICROBIAL RESISTANCE PHENOTYPES FROM GENOTYPE

By KARA K. TSANG, B.HSc.

A Thesis Submitted to the School of Graduate Studies in partial fulfilment of requirements for the degree Doctor of Philosophy

DESCRIPTIVE NOTE

McMaster University, DOCTOR OF PHILOSOPHY (2021), Hamilton, Ontario (Health Sciences)

TITLE: Prediction of antimicrobial resistance phenotypes from genotype

AUTHOR: Kara K. Tsang B.HSc. (McMaster University)

SUPERVISOR: Dr. Andrew G. McArthur

NUMBER OF PAGES: xx, 254

LAY ABSTRACT

Many surgeries, chemotherapy, and transplantation will be impossible if antibiotic resistance is not addressed. Antibiotic misuse, overuse, and time to definitive therapy exacerbate this global health problem. Phenotypic testing determines definitive therapy, but bacterial culturing is slow. A potentially faster and more accurate approach relies on sequencing the pathogen's genome.

I used machine learning to generate antibiotic resistance prediction models that achieved average accuracies of 94% and 89% for *Escherichia coli* and *Pseudomonas aeruginosa*, respectively. These models identified novel relationships between known resistance genes and resistance phenotypes, which were experimentally validated.

Resistance and susceptibility are interpretations of a minimum inhibitory concentration (MIC) using a clinical breakpoint guideline. Since there are different guidelines, I generated MIC prediction models with average accuracies of 86%, 41%, and 98% for *E. coli*, *P. aeruginosa*, and *Neisseria gonorrhoea*, respectively.

My findings work towards a world where clinical sequencing and genomics-based diagnostics are the gold standard.

ABSTRACT

Antimicrobial resistance (AMR) is a threat to global health, food security, and economic productivity. Infections caused by drug resistant Gram-negative pathogens, such as *Escherichia coli, Pseudomonas aeruginosa,* and *Neisseria gonorrhoeae*, are continuously becoming harder to treat due to limited treatment options and long turnaround times for culture-based phenotypic diagnosis. Alternatively, genotypic approaches that exploit whole genome sequencing have the potential to be faster and more accurate. Genotypic approaches rely on using bacterial genomes to predict AMR phenotypes.

I generated a rules-based algorithm and machine learning models using known resistance determinants from bacterial genomes to predict resistance or susceptibility.
I showed that machine learning was superior to a rules-based algorithm and achieved an average accuracy of 94% and 89% for *E. coli* and *P. aeruginosa,* respectively. These machine learning models identified novel AMR genotype-phenotype relationships between known resistance determinants and resistance phenotypes, which were experimentally validated.

To identify the parameters that can improve machine learning models, I tested a variety of genetic features, algorithms, and evaluation metrics. I observed an intricate dependency between parameters for AMR prediction performance, illustrating that careful selection of parameters is required to generate accurate AMR prediction models.

A limitation of this work was its prediction of resistance and susceptibility categories, as these are interpretations of minimum inhibitory concentrations defined by clinical breakpoint guidelines. Since multiple guidelines exist, these prediction models are not generalizable, so prediction of MIC values was explored. The average accuracy of my MIC prediction models was 86%, 41%, and 98% for *E. coli*, *P. aeruginosa,* and *N. gonorrhoea,* respectively.

Despite the multifactorial and intricate nature of the resistome, I was able to accurately predict AMR phenotypes for many antibiotics for these pathogens. This is a step towards advanced diagnostic microbiology methods driven by genomics.

ACKNOWLEDGEMENTS

It truly takes a village.

To my supervisor Dr. Andrew McArthur – thank you for nurturing the best environment for me to grow as a scientist and as a person. You have taught me the type of supervisor I want to be, and I can't thank you enough for the years of opportunities, support, and advice.

To Dr. Jennifer Stearns and Dr. Gerry Wright – thank you for being on my supervisory committee. Your guidance and insightful contribution to this work has been instrumental.

To the McArthur Lab members, both past and present – I have learned something from every one of you. Brian, you are my lab rock, what would I do without you? Amos, your patience is unparalleled, thank you for always knowing what to do. Jalees, thank you for your puns and putting up with my questions. Arman, thank you for your humour and kindness. Tammy, Bhavya, and Sally – I am incredibly grateful for our friendship that stemmed in this lab and continues to bloom.

To those who I've had the privilege of collaborating with – Dr. Sarah Khan, Dr. John Whitney, Dr. Finlay Maguire, Haley, Sommer, Sheri – I am grateful to you all for the opportunity and your willingness to share your science with me.

To my loved ones – Phil, Esther, Elizabeth, Anika, Laura, Leanne, and Jackie, thank you for loving me; you are the support system I didn't know I needed until I did.

To my brother Derek, thank you for always having my back. To my parents, Silas and Dorothy, thank you for the unconditional love and sacrifices you made for our family. From immigrating to Canada, to starting your hair salon, to now being awarded national and international hairdresser of the year awards. I am so proud of your achievements which have always been my beacon of motivation.

This accomplishment is as much all of yours as it is mine.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

LIST OF ABBREVIATIONS

AdaB	AdaBoost
AMR	Antimicrobial Resistance
ARP	Antibiotic Resistance Platform
AST	Antimicrobial susceptibility testing
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Alignment
CARD	Comprehensive Antibiotic Resistance Database
CART	Classification and Regression Trees
CBMM	Class-Conditional Bernoulli Mixture model
CLSI	Clinical Laboratory Standards Institute
DT	Decision Trees
EUCAST	European Committee on Antimicrobial Susceptibility Testing
GBT	Gradient Boosted Trees
GS	Grantham Score
LinR	Linear Regression
LogR; LR	Logistic Regression
MIC	Minimum inhibitory concentration
NB	Naïve Bayes
NCBI	National Center for Biotechnology Information
NN	Neural Network
PS	Perfect and Strict Representation
RD	Resistance Determinants
RF	Random Forest
RGI	Resistance Gene Identifier
SCM	Set Covering Machine
SVM	Support Vector Machine
XGBoost	Extreme gradient boosting

DECLARATION OF ACADEMIC ACHIEVEMENT

I have performed all of the research in this body of work except where indicated in the preface of each chapter.

CHAPTER ONE: Introduction

ANTIMICROBIAL RESISTANCE

By 1942, a mere 14 years after the discovery of penicillin by Sir Alexander Fleming in 1928, penicillin resistant infections were documented in hospitals (Lobanovska & Pilla, 2017). Over the next few years, penicillin-resistant *Staphylococcus aureus* infections became widespread in the community and hospitals (Rammelkamp, Maxon, & Medicine, 1942). Antimicrobial discovery followed by resistance is not unique to penicillin and has been observed time and time again (Ventola, 2015). Antimicrobial resistance (AMR) is a natural phenomenon whereby microorganisms resist the effectiveness of antimicrobials. Over time, AMR has become a global health problem that many national and international organizations are addressing with varying degrees of success. The World Health Organization approved a Global Action Plan in 2015 which included the launch of the Global Antimicrobial Resistance and Use Surveillance System, the first worldwide effort to standardize AMR surveillance (World Health Organization, 2015, 2017a). The US Centers for Disease Control and Prevention estimated that in 2019 more than 2.8 million AMR infections occur every year and 35,000 people die as a result (CDC, 2019). This was an increase from their 2013 report, where 2 million people had an AMR infection and 23,000 people died, illustrating AMR as a growing public health threat (CDC, 2013). In 2018, approximately 5,400 people died as a result of AMR in Canada, with an economic impact of $2 billion due to deaths and illnesses related to AMR infections (Council of Canadian Academies, 2019). The Canadian Antimicrobial Resistance Surveillance System also indicates that AMR infections are becoming more

prevalent, leading to increased illness, death, and healthcare costs (Public Health Agency of Canada, 2020).

Alongside misuse and overuse of antibiotics, we are currently in a shortage of new antibiotics because of scientific and structural (e.g., financial and regulatory) challenges (Silver, 2011; Zorzet, 2014). Most newly approved antibiotics are modifications of existing antibiotic classes, rather than novel classes (World Health Organization, 2021). One of the few exceptions is teixobactin, which is the first of a new antibiotic class that was discovered in 2015 and is currently in late-stage preclinical development (Ling *et al.*, 2015). One main barrier is that emergence of resistance to antibiotics is nearly inevitable (Ventola, 2015) and bacteria do so through a few main mechanisms, which include enzyme catalyzed antibiotic modifications, bypass of antibiotic targets, and efflux of drugs from the cell (G. D. Wright, 2011). Bacteria destroy or modify antibiotics to resist their action through hydrolysis or transfer of a chemical group, but bacteria can also protect or modify the antibacterial target site to reduce affinity for the antibiotic. Bacteria can further reduce permeability of the cell to antibiotics (i.e., reduce uptake of drug) or utilize machinery called efflux pumps to actively expel antibiotics from the cell. Bacteria can either acquire or intrinsically possess one or more of these resistance mechanisms. Acquisition of resistance can be attributed to either horizontal gene transfer from other cells or the generation of spontaneous AMR mutations that can be vertically transmitted. Yet even without acquired resistance, not all antibiotics are effective for all infections due to intrinsic resistance, a universal trait that is independent of antibiotic selective pressure and horizontal gene transfer (G. Cox & Wright, 2013). For example, Gram-negative

bacteria have an additional outer membrane which makes them impermeable to many antibiotics that are effective for Gram-positive bacteria. In addition, it is important to acknowledge the effects of AMR in microbial communities since bacteria seldom live as planktonic bacteria. Bacteria without AMR mechanisms are able to survive antibiotic treatment through tolerance, where they slow essential processes or interact with the host's immune system or other bacterial species (Bottery, Pitchford, & Friman, 2021). Persistence, on the other hand, is when a small population of bacteria become dormant to allow survival of sub-population despite high antibiotic concentrations, allowing post-treatment re-emergence of infection (Bottery *et al.*, 2021). Additionally, bacteria can form biofilms when they attach to a surface and produce a hydrated matrix to create microenvironments that promote growth while protecting the microorganisms from external insult, such as antibiotic treatment (Patel, 2005). Collectively, all of the antibiotic resistance determinants and their precursors that are encoded in bacterial genomes or mobile genetic elements are called the 'resistome' (G. D. Wright, 2007).

GENOMICS OF BACTERIA AND RESISTANCE

Bacterial genomics is the study of the hereditary information of bacteria that can be vertically or horizontally transmitted. The field of bacterial genomics has grown exponentially since the bacterial genomes of *Haemophilus influenzae* Rd and *Mycoplasma genitalium* were first completely sequenced in 1995 (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995). Between 2015 and 2021, NCBI's sequenced bacterial genome database grew from over 30,000 to 335,910 sequenced bacterial genomes (Land *et al.*,

2015; NCBI Resource Coordinators, 2018). Simplification and cost reduction of sequencing technology has made bacterial genome sequencing affordable to researchers, which has shifted the cost and workload towards downstream bioinformatics analysis and data management.

Sanger sequencing in combination with whole genome shotgun sequencing, high-throughput sequencing, and single molecule long-read sequencing are currently the three generations in sequencing technology (Loman & Pallen, 2015). Using any of these sequencing technologies generates short or long sequencing reads that are then trimmed for quality control, using tools such as Trimmomatic (Bolger, Lohse, & Usadel, 2014). Afterwards, they can be assembled into chromosomes and plasmids by aligning sequences to generate longer, contiguous consensus sequence fragments called contigs. Genome assembly methods include SPAdes, which will assemble chromosomes and plasmids (Bankevich *et al.*, 2012), whereas HyAsP identifies only plasmid sequences (Müller & Chauve, 2019). In an ideal scenario, contigs represent the complete chromosome and plasmid sequence of the bacteria. However, in practice, sequencing errors and biases, along with repetitive DNA regions, are challenges for genome assemblers and often result in draft fragmented genomes instead of complete circular genomes (Treangen & Salzberg, 2011; Utturkar, Klingeman, Hurt, & Brown, 2017). Assembling only plasmid genomes, particularly large plasmids (>50kbp), is challenging because they contain shared repeat sequences and different *k*-mer (short nucleotide sequences) abundance profiles that are difficult to resolve using de Bruijn graph-based assemblers (Arredondo-Alonso, Willems, van Schaik, & Schurch, 2017). The former

difficulties are only exacerbated by potential contamination from non-target organisms, whose sequences can be incorporated into the contigs or causes misassembly of contigs (Goig, Blanco, Garcia-Basteiro, & Comas, 2020). Even with all the complexities of bacterial genome sequencing and assembly, using draft genomes provides practical information that can be used to understand evolutionary origin, transmission route, pathogenicity, and resistance potential of bacteria. While the complete genome sequence itself may not be resolved, the gene content of draft genomes is in the majority accurately represented by assembly data.

To understand the drivers of resistance in bacteria, the resistome, the part of the genome that encodes for resistance, can be annotated using a number of bioinformatics databases and tools. Typically, comparative sequence analysis is used to annotate potential resistance determinants using either a read-based or assembly-based approach. Trimmed sequence reads can be mapped onto a reference database or trimmed sequence reads can be assembled into contigs and then compared to a reference database. In the read-based approach, pairwise alignment tools based on the Burrows–Wheeler transform such as Bowtie2 and BWA are commonly used (Langmead & Salzberg, 2012; H. Li & Durbin, 2009), although recent work shows that AMR in particular may require a new class of algorithms due to its complex network of similar alleles (P. Clausen, Aarestrup, & Lund, 2018). In comparison, after assembly, pairwise alignment tools such as BLAST or DIAMOND are used to compare the assembly to a reference sequence database (Buchfink, Xie, & Huson, 2015; Madden, 2013). While an assembly-based approach requires more computational power, it allows for understanding of genomic context (e.g.,

regulatory and mobile elements). However, genome mis-assembly, particularly if plasmids are involved (Robertson & Nash, 2018), can cause a loss of information compared to a read-based method.

Regardless of the method, antibiotic resistance annotation is highly dependent on the reference sequence database used (Boolchandani, D'Souza, & Dantas, 2019; Xavier *et al.*, 2016). The Comprehensive Antibiotic Resistance Database (CARD) is a curated collection of AMR determinants connected to the antibiotics they confer resistance towards using the Antibiotic Resistance Ontology (ARO) (Alcock *et al.*, 2020). CARD's Resistance Gene Identifier (RGI) software uses the information in CARD to predict AMR determinants from genomic sequence data. RGI works under three paradigms: 1) Perfect, the complete detection of a known AMR gene that has peer-reviewed evidence of experimentally elevated minimum inhibitory concentration, 2) Strict, the detection of previously unknown variants of known AMR genes, including mutations, that pass a curated bit score cut-off, and 3) Loose, the detection of new, emergent threats and distant homologs of AMR genes that fall below the bit score cut-off. The latter could be newly emerging AMR genes or spurious matches and require experimental validation. The Antibiotic Resistance Gene-ANNOTation database (Gupta *et al.*, 2014) and Pathosystems Resource Integration Center (Davis *et al.*, 2020; Wattam *et al.*, 2017) store a similar breadth of resistance determinants as CARD and also use BLAST-based tools for resistome annotations. Antibiotic Resistance Genes Online (Scaria, Chandramouli, & Verma, 2005) only catalogues β-lactam and vancomycin resistance determinants, in comparison to ResFinder (Bortolaia *et al.*, 2020) which primarily annotates acquired

resistance genes using BLASTN, while ResFams (Gibson, Forsberg, & Dantas, 2015) is a database of protein domain Hidden Markov Models associated with AMR functional domains. The dependency on reference sequence databases is an inherent limitation of antibiotic resistance determinant annotation because it requires active and, oftentimes manual, curation. Without addressing the dynamic nature of biocuration, exacerbated by the ever evolving resistome, important epidemiological data on newly discovered resistance determinants cannot be collected and incorrect annotations can perpetuate throughout literature. Antibiotic resistance genomic data increases as new mechanisms of resistance and genes are discovered, which then need biocuration into these databases, e.g., the discovery of NDM-1 and MCR-1 in 2009 and 2016, respectively (Y. Y. Liu *et al.*, 2016; Yong *et al.*, 2009), but also discovery of new antibiotics, e.g. teixobactin (Ling *et al.*, 2015). Biocurators are also responsible for resolving gene/protein nomenclature conflicts and redundancy within and across databases, e.g., *dhfr* is often curated as *dfrA* (Xavier *et al.*, 2016). Furthermore, curated resistance determinants may require editing when more accurate information is published. For example, CrpP was previously thought to be a ciprofloxacin-modifying enzyme, but it has recently been shown to not inactivate fluoroquinolones, including ciprofloxacin (Chavez-Jacobo *et al.*, 2018; Zubyk & Wright, 2021). Above all, biocurators already face lack of funding, curating increased volumes of data, and difficulties in convincing other scientists of the need and importance of databases (Burge *et al.*, 2012). Lastly, while resistome annotation can identify the presence of resistance determinants, it does not necessarily infer expression or repression of a particular resistance phenotype.

GENOTYPE TO ANTIBIOTIC RESISTANT PHENOTYPE

Antibiotic Resistant Phenotypes

To determine suitable treatment and to monitor the spread of resistant microbes, antimicrobial susceptibility testing (AST) is routinely performed in clinical microbiology labs. AST is typically performed after bacterial culturing and species identification. Currently, the gold standard is culture-based (or phenotypic) AST, where it measures bacterial cell growth in the presence of an antimicrobial to determine the minimum inhibitory concentration (MIC) to inhibit growth. There are a number of methods to perform phenotypic AST, including disk diffusion, E-test / gradient test, and broth microdilution. Disk diffusion and gradient tests detect decreased or no visible growth within a zone of inhibition, whereas broth microdilution assesses for the lack of visibility in broth. There are also commercial semi-automated machines, which can read zones of inhibition in disc diffusion assays, and a number of automated broth dilution assays. Common underlying issues with all of these tests are that they require bacterial culturing and there are years of lag before new antimicrobials can be used in these routine methods. Practically, a full test run or validation set must be complete before testing on samples, which requires staff availability (van Belkum *et al.*, 2019). Current turnaround times for most AST results are between 48h and 72h depending on the species and antibiotic combination. This time includes overnight incubation, where non-fastidious pathogens (e.g., *Escherichia coli*) require 16-18 hours of incubation but fastidious pathogens (e.g., *Neisseria gonorrhoeae*) could require 18-24 hours or even 4-8 weeks of incubation (e.g., *Mycobacterium tuberculosis*) (Ghodbane, Raoult, & Drancourt, 2014; Jorgensen &

Ferraro, 2000; Melendez, Hardick, Barnes, Page, & Gaydos, 2018). Another limitation of AST is that it is not globally standardized. Currently, there are two popular guidelines, the European Committee on Antimicrobial Susceptibility Testing (EUCAST) and the Clinical Laboratory Standards Institute (CLSI) (CLSI, 2018; EUCAST, 2015). There are a number of differences between EUCAST and CLSI guidelines that include AST methodology, cost, definitions of the 'intermediate' resistance category, and clinical breakpoints (Cusack, Ashley, Ling, Roberts, *et al.*, 2019). Clinical breakpoints organize MICs into 'resistant', 'intermediate', and 'susceptible' categories, which are then used to inform treatment or public health surveillance. Since clinical breakpoints are rarely aligned between EUCAST and CLSI, several laboratories have compared the consequences of using either guideline (Cusack, Ashley, Ling, Rattanavong, *et al.*, 2019; Hombach, Mouttet, & Bloemberg, 2013; Kassim, Omuse, Premji, & Revathi, 2016; Rodríguez-Baño, Picón, Navarro, López-Cerero, & Pascual, 2012; Wolfensberger *et al.*, 2013).

In contrast to culture-based methods, there are genotypic AST methods that rely on identifying specific resistance determinants using molecular or genomic approaches. Most currently used genotypic AST methods are rapid, culture independent, and polymerase chain reaction (PCR)-based. However, currently genotypic ASTs provide supplemental information that still requires validation with phenotypic AST and are only available for a small subset of clinically important resistance determinants. For example, the *mecA* resistance gene can be identified in *Staphylococcus aureus* through PCR within minutes, however phenotypic AST is still required to test for susceptibility towards other antibiotics to treat a *mecA*-identified methicillin-resistant infection (Banerjee &

Humphries, 2021; van Belkum *et al.*, 2019). Additionally, the identification of resistance genes using genotypic AST is rarely completely correlated with phenotypic AST due to the multifactorial nature of AMR (Banerjee & Humphries, 2021). While there are limited molecular tests, there is also hesitation towards implementing whole genome sequencing technology in clinical microbiology labs due to added costs, turnaround time, and lack of evidence towards the clinical utility of AMR predictions (Rossen, Friedrich, & Moran-Gilad, 2018). This latter point is best exemplified by the Gram-negative pathogens (e.g., *Pseudomonas aeruginosa*), where AMR genotype-phenotype relationships are difficult to predict.

Genotype-Phenotype Antibiotic Resistant Relationships

Since the resistome encompasses all genetic drivers of resistance, prediction of phenotypic resistance should be possible using bacterial genomes. In particular, acquired resistance mechanisms generally result in a predictable increase of resistance. For example, mutations in *rpsL, rpoB,* or *gyrA* genes always confer in an increased level of resistance towards streptomycin, rifampicin, or nalidixic acid, respectively (Hughes & Andersson, 2017). In addition, over 30 resistance genes in 76 *E. coli* correlate to phenotypes with 97.8% specificity and 99.6% sensitivity (Tyson *et al.*, 2015). However, after decades of research on the resistome, it is evident that genotype does not always result in the expected phenotype for AMR. Part of the reason for this dissociation is that we collectively have not characterized the entire resistome and there are unknown genes and mutations (e.g., genetic dark matter) that influence phenotype. In addition to

unknowns in the genome, we also poorly understand the genetic context (in addition to resistance genes and mutations) and environmental changes that can cause the dissociation of genotype and phenotype.

While we have characterized a number of resistance determinants, it is evident that some resistance phenotypes are shaped by the interplay of a combination of these resistance determinants and not individual genes or mutations. A notable exception is *M. tuberculosis,* which is genetically homogenous and where resistance can be caused by a single chromosomal mutation. Yet, in Gram-negative bacteria fluoroquinolone resistance involves mutations in antibiotic target genes, efflux regulators, and acquired resistance genes (Jacoby, 2005; Webber & Piddock, 2001). In fact, individual mutations were not shown to increase fluroquinolone MICs beyond clinical breakpoints (Marcusson, Frimodt-Moller, & Hughes, 2009). Resistance to β-lactams in *N. gonorrhoeae* and *Streptococcus pneumoniae* occurs by homologous recombination that produces mosaic genes, e.g., penicillin-binding proteins, whereas in *Klebsiella pneumoniae* β-lactamases are horizontally gene transferred via plasmids (Bush, 2013; Hakenbeck, Bruckner, Denapaite, & Maurer, 2012; Tapsall, 2009). There is also evidence of epistatic relationships, where resistance genes and/or mutations can interact and change a resistance phenotype. For example, in *E. coli* a mutation in GyrA (S83L) increases ciprofloxacin MIC to 0.25 μg/mL whereas a mutation in ParC S80I does not change the MIC (0.015 μg/mL), but together in a double mutant they increase the MIC by greater than additivity to 0.75 μg/mL (Huseby *et al.*, 2017). Lastly, antibiotic resistance genes are subject to genetic amplification that can alter the resistance phenotype through copy

number (Hjort, Nicoloff, & Andersson, 2016). Along with the complexities of genetic context, environmental modulation of resistance phenotypes have also been observed. Microbial communities can form biofilms which have increased resistance phenotypes due to increased physical protection and altered growth physiology (Olsen, 2015). Antibiotic induced and metabolite modulated antibiotic resistance can also occur, where the presence of other small compounds can change a resistance phenotype (Hanson & Sanders, 1999; Thulin, Sundqvist, & Andersson, 2015).

ANTIBIOTIC RESISTANCE PHENOTYPE PREDICTION

Methods for Predicting Antibiotic Resistance Phenotypes

To our knowledge, the first publication that compared AMR prediction based on whole genome sequencing and AMR phenotypes was published in 2013 (Zankari *et al.*, 2013). The authors demonstrated high concordance between AMR prediction and AMR phenotypes in *Salmonella, E. coli,* and *Enterococcus* (Zankari *et al.,* 2013). AMR phenotype prediction using whole genome sequencing has been a growing field with 74 publications between 2013-2021 (determined using the search terms "antimicrobial", "predict" "genome", “resistance”, “phenotype” in NCBI PubMed).

Predicting antibiotic resistance phenotypes begins with selecting a dataset that includes features and resistance phenotypes. These genetic features are derived from the genome in the form of short nucleotide sequences, known resistance genes, or all observed mutations. The resistance phenotypes are the ‘resistant’/‘susceptible’ categories (i.e., qualitative) or minimum inhibitory concentrations (MICs) (i.e., quantitative). The

dataset should be evaluated for the number of samples, sparsity of data, and balance of resistance phenotypes (e.g., the less representative phenotype should have at least >10% prevalence in the dataset). If there is a low number of samples and imbalance of resistance phenotypes, the dataset can still be used for AMR prediction, however these limitations must be acknowledged and generalization may be difficult. Following dataset selection, rules-based, machine learning, and neural network algorithms have all been applied for AMR phenotype prediction (Figure 1-1).

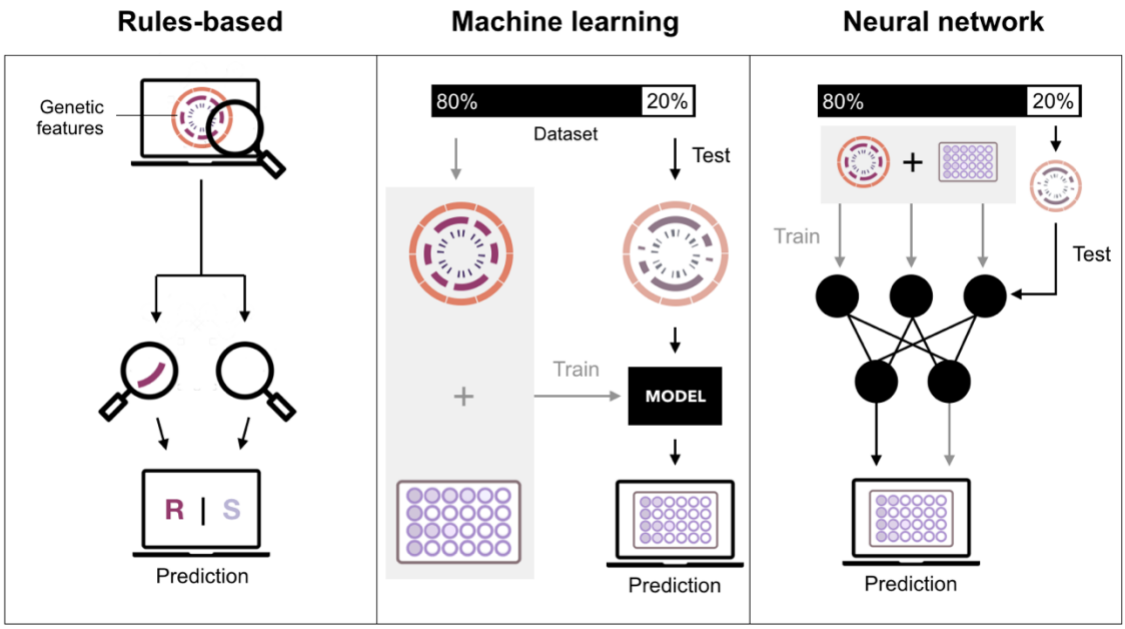


Figure 1-1. Methods for AMR phenotype prediction. Rules-based methods use known resistance determinants and can only predict resistance (R) or susceptibility (S). Machine learning algorithms generate a model using the training dataset, where then the performance of the model is evaluated by using the test dataset. Neural networks use the training dataset to develop a network, where it is then evaluated using the test set. Machine learning and neural networks are able to predict R/S or minimum inhibitory concentrations.

Rules-based algorithms predict resistance based on the presence of resistance determinants and rely on expert and current knowledge, generally stored in AMR databases like CARD. For machine learning, the dataset of genotypes and phenotypes is split into a training and test set (Larranaga *et al.*, 2006) and the algorithm learns patterns within the training dataset to generate a model. Typically, a number of different algorithms are tested and an evaluation metric (e.g., log loss) is used to select the final model that best predicts AMR phenotype for new data. Since the test set has not been used in training the final model, it provides an unbiased evaluation of the final model performance. Neural networks are similar to machine learning, except that the training set is used to make a network and the test set is used to evaluate the network's performance and that they generally require larger datasets (Kröse, Krose, van der Smagt, & Smagt, 1993). There are also neural network learning (optimization) algorithms that can be used, however testing many of them will require much more computational power than machine learning algorithms. Regardless of algorithm selected, prediction results are always compared to the laboratory-determined resistance phenotypes to evaluate prediction performance.

Antibiotic Resistance Phenotype Prediction Parameters

AMR phenotype prediction using rules-based, machine learning, and neural networks are summarized in the following reviews (Lv, Deng, Zhang, & Health, 2020; Macesic, Polubriaginof, & Tatonetti, 2017; McDermott & Davis, 2021; Su, Satola, & Read, 2018). There are a number of AMR prediction publications that use a variety of

different datasets that span many Gram-negative and Gram-positive pathogens, including *Acinetobacter baumannii*, *Campylobacter jejuni*, *Campylobacter coli*, *Escherichia coli*, *Enterococcus faecalis*, *Enterococcus faecium*, *Enterobacter aerogenes*, *Klebsiella pneumoniae*, *Mycobacterium tuberculosis*, *Neisseria gonorrhoeae*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Shigella sonnei*, *Salmonella enterica* serovar Typhi, non-serovar Typhi *Salmonella enterica,* and *Streptococcus pneumoniae* (Table 1-1, 1-2). These datasets also include testing on a number of different antibiotics and use a variety of rules-based and machine learning parameters (Figure 1-2).

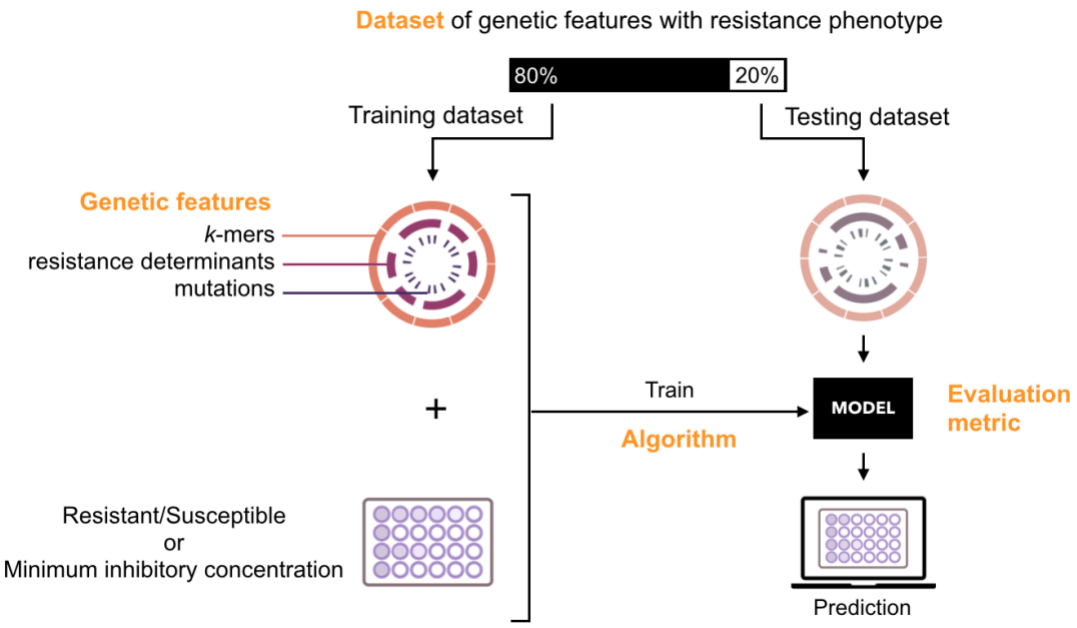


Figure 1-2. Parameters in machine learning for AMR phenotype prediction. In orange are the parameters of machine learning that can be selected by the researcher.

In addition to dataset selection, there are many aspects to predicting AMR phenotypes including genetic feature selection, algorithm, and evaluation metric choice. For feature selection, whole genome sequences from an isolate can be used in a number of methods for AMR phenotype prediction. Sequences can be assembled into contigs and then annotated with known resistance genes (as previously described), divided into short nucleotide sequences of *k* length, called *k*-mers, or used to identify mutations relative to a reference. Using known resistance determinants relies on the quality and quantity of information within a reference sequence database and is inherently blind to any unknown resistance determinants. The available antibiotic resistance reference databases are not harmonized, resulting in discordant AMR prediction (Doyle *et al.*, 2020; Mahfouz, Ferreira, Beisken, von Haeseler, & Posch, 2020). On the other hand, *k*-mers can span the entire bacterial genome, thus in theory they can be used to discover novel resistance determinants (Drouin *et al.*, 2016; Drouin *et al.*, 2019). Catalogs of mutations relative to a reference make a similar, but more discrete data set than *k*-mers but their performance is dependent upon the reference (Eyre *et al.*, 2017; Niehaus, Walker, Crook, Peto, & Clifton, 2014).

For algorithms, rules-based methods, which rely on expert and current knowledge, make a resistant phenotype prediction based on presence of an AMR determinant and inversely, a susceptible prediction if the AMR determinant is absent (Bradley *et al.*, 2015; Gordon *et al.*, 2014; Hasman *et al.*, 2014; Kos *et al.*, 2015; B.J. Metcalf *et al.*, 2016; Moran, Anantham, Holt, & Hall, 2016; Pesesky *et al.*, 2016; Stoesser *et al.*, 2013; Walker *et al.*, 2015; Zankari *et al.*, 2013). As such, rules-based methods are ‘hard-wired’ and lack

any potential unknown resistance determinants when making AMR phenotype predictions. In contrast, machine learning algorithms identify patterns within a dataset to infer AMR genotype-phenotype relationships when making AMR predictions. This approach can be advantageous because it is not limited to defined rules. Acknowledging that many techniques (e.g., logistic regression and random forest) are borrowed from the field of statistics for machine learning (ML) (Larranaga *et al.*, 2006), the algorithms discussed will collectively be called ML algorithms to prevent confusion.

There are two types of machine learning, supervised and unsupervised. Supervised learning uses prior knowledge of what the output (e.g., resistance phenotype) should be, whereas unsupervised learning seeks learn the structure of the data without explicitly provided prior knowledge. AMR phenotype prediction studies use supervised learning as we have the input (e.g., genetic features) and output (e.g., resistance phenotypes) and we want to learn the relationships (i.e., patterns) associating the input and the output to make new predictions. In this work, there are two types of supervised learning problems: classification and regression. The goal of a classification problem is to categorize the associations within the data into two or more classes (e.g., resistant or susceptible categories), whereas a regression problem predicts a continuous, quantitative output variable (e.g., minimum inhibitory concentration). The classification algorithms that have been used for resistant or susceptible prediction are logistic regression, decision tree, random forest, naïve Bayes, adaBoost, XGBoost, support vector machine, set covering machine, and neural networks. Over these, the simplest ML classification algorithm is logistic regression and is an extension of linear regression (D. R. Cox, 1958). The

decision tree algorithm creates a model by learning simple decision rules inferred from the features (Quinlan, 1986). Random forest is a combination of many decision trees, which makes it more difficult to interpret (Breiman, 2001). Naïve Bayes assumes independence among features when developing a model (Lewis, 1998), which may not best reflect interplay of AMR genes. AdaBoost includes sequentially growing decision trees that punish incorrectly predicted samples (i.e., assigns weights to incorrect values) to allow it to learn from previous mistakes (Freund & Schapire, 1996). Extreme gradient boosting (XGBoost) uses the best parts of random forests and AdaBoost to also increases speed and is able to ignore irrelevant features in the model (Friedman, 2001), which can be valuable for noisy data sets. Support vector machine is similar to logistic regression, but it attempts to find the largest margin that separates the prediction phenotypes (Suthaharan, 2016). Set covering machine learns conjunctions (e.g., gene 1 and gene 2) or disjunctions (e.g., gene 1 or gene 2) between the features to predict a phenotype (Marchand, Shawe-Taylor, Brodley, & Danyluk, 2002). Lastly, modelled after the brain, neural networks consists of many processing nodes that, in its simplest form, are interconnected where information only moves in one (forward) direction (Wang, 2003). Random forest and XGBoost have also been used for quantitative MIC prediction, as has linear regression, which predicts AMR phenotype using the best straight line fit to a set of genetic features (Lai, Robbins, & Wei, 1978).

One of the remaining choices in AMR phenotype prediction is evaluation metric choice. For rules-based algorithms, results are reported via a confusion matrix: true positive (prediction and laboratory phenotype are both resistant), true negative (prediction

and laboratory phenotype are both susceptible), false positive / major error (prediction is resistant, but laboratory phenotype is susceptible) or false negative / very major error (prediction is susceptible, but laboratory phenotype is resistant). With the confusion matrix values, accuracy, recall/sensitivity, precision, F1 score, and specificity can be calculated and are used to report the results of AMR phenotype prediction (Ting, 2017), elaborated on further in Chapter 3. Unlike rules-based approaches, in machine learning choosing an evaluation metric is important for both model selection and assessment of the final model performance (e.g., results of AMR phenotype prediction). Similar to evaluation metrics used in rules-based algorithms, accuracy, recall/sensitivity, precision, F1 score, specificity, plus log loss (described in detail in Chapter 2 and 3) can be used for model selection and final model performance. Choosing one evaluation metric will have consequences and trade-offs towards other evaluation metrics. For example, when two models have different accuracy, precision, and log loss values, prioritizing one evaluation metric may come at the cost of another. One challenge in currently published AMR prediction models is lack of standardization of evaluation metric reporting, which makes it difficult to compare results.

Table 1-1. AMR prediction using rules-based algorithms. NR indicates not reported. In *M. tuberculosis* analyses, if uncharacterized mutations were included, this is indicated with an asterisk. Resistance determinants (RD) or only mutations were used in these studies.

Species	Database + Software	Genetic Feature	Reference
Campylobacter coli	Custom, ARDB, ResFinder + BLASTx, ClustalW	RD	(S. Zhao *et al.*, 2015)
Campylobacter jejuni	Custom, ARDB, ResFinder + BLASTx, ClustalW	RD	(S. Zhao *et al.*, 2015)
	ResFinder, PointFinder	RD	(Bortolaia *et al.*, 2020)
Campylobacter spp.	AMRFinder	RD	(Feldgarden *et al.*, 2019)
Enterococcus faecalis, Enterococcus faecium	ResFinder	RD	(Zankari *et al.*, 2013)
	ResFinder, PointFinder	RD	(Bortolaia *et al.*, 2020)
	ResFinder, NCBI Pathogens Database, BLASTn	RD	(Tyson, Sabo, Rice-Trujillo, Hernandez, & McDermott, 2018)
Escherichia coli	Custom + BLASTx, ClustalW	RD	(Tyson *et al.*, 2015)
	ResFinder	RD	(Zankari *et al.*, 2013)
	Custom + BLASTn	RD	(Stoesser *et al.*, 2013)
	AMRFinder	RD	(Feldgarden *et al.*, 2019)
	ResFinder, PointFinder	RD	(Bortolaia *et al.*, 2020)
	ResFinder, PointFinder	RD	(Aytan-Aktug, Clausen, Bortolaia, Aarestrup, & Lund, 2020)
Helicobacter pylori	CLC genomic Workbench	RD	(Tuan *et al.*, 2019)
Klebsiella pneumoniae	Custom + BLASTn	RD	(Stoesser *et al.*, 2013)
Mycobacterium tuberculosis	PhyResSE + Stampy	RD	(Pankhurst *et al.*, 2016)
	TBDReaMDB, MUBII-TB-DB minus phylogenetic SNPs + TB Profiler	Mutations	(Coll *et al.*, 2015)
	NR	Mutations	(Miotto *et al.*, 2017)
	Custom + SAMtools, mpileup, Cortex	RD	(Walker *et al.*, 2015)
	Hain, Cepheid, AIDb assays, literature + Mykrobe	RD	(Bradley *et al.*, 2015)
	Stampy, Platypus	RD	(M. L. Chen *et al.*, 2019)

	ResFinder, PointFinder	RD	(Aytan-Aktug *et al.*, 2020)
	Custom	RD	(Yang *et al.*, 2018)
	Mykrobe	RD	(Quan *et al.*, 2018)
Neisseria gonorrhoeae	Custom, BLAST, BWA mem	RD	(Eyre *et al.*, 2017)
Pseudomonas aeruginosa	Custom + NR	RD	(Kos *et al.*, 2015)
Salmonella enterica	ResFinder, PointFinder	RD	(Bortolaia *et al.*, 2020)
	ResFinder, PointFinder	RD	(Aytan-Aktug *et al.*, 2020)
	AMRFinder	RD	(Feldgarden *et al.*, 2019)
Salmonella enterica serovar Typhi	CARD, ResFinder, literature + GeneFinder	RD	(Day *et al.*, 2018)
	ARIBA, ResFinder, PointFinder	RD	(Mensah *et al.*, 2019)
Non-typhoidal *Salmonella enterica*	Custom + **BLASTx**, ClustalW	RD	(McDermott *et al.*, 2016)
	CARD, ResFinder + GeneFinder	RD	(Neuert *et al.*, 2018)
	CARD, RGI	RD	(Maguire, Rehman, Carrillo, Diarra, & Beiko, 2019)
Salmonella Typhimurium	ResFinder	RD	(Zankari *et al.*, 2013)
Shigella sonnei	CARD, ResFinder + GeneFinder	RD	(Sadouki *et al.*, 2017)
Staphylococcus aureus	Custom + BLASTn, tBLASTn	RD	(Gordon *et al.*, 2014)
	Custom + BLASTn	RD	(Aanensen *et al.*, 2016)
	Custom + Mykrobe	RD	(Bradley *et al.*, 2015)
	Custom + Mykrobe, GeneFinder, Typewriter	RD	(Mason *et al.*, 2018)
	ResFinder, PointFinder	RD	(Bortolaia *et al.*, 2020)
	ResFinder, PointFinder	RD	(Aytan-Aktug *et al.*, 2020)
Streptococcus pneumoniae	SRST2	RD	(Deng *et al.*, 2016)

Table 1-2. AMR prediction models using machine learning algorithms. Machine learning algorithms include, gradient boosted trees (GBT), random forest (RF), class-conditional Bernoulli mixture model (CBMM), neural networks (NN) AdaBoost (AdaB), support vector machine (SVM), linear regression (LinR), XGBoost (XGB), logistic regression (LogR), set covering machine (SCM), classification and regression trees (CART). Studies that also or only predicted MICs are highlighted in yellow.

Species	ML algorithm	Genetic Feature	Reference
Acinetobacter baumannii	AdaB	*k*-mers (genome)	(Davis *et al.*, 2016)
	RF	*k*-mers (genome)	(Santerre *et al.*, 2016)
	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)
	XGB	RD	(Kim *et al.*, 2019)
	SCM, RF	*k*-mers (genome)	(Hicks *et al.*, 2019)
Actinobacillus pleuropneumoniae	SVM, SCM	*k*-mers	(Z. Liu *et al.*, 2020)
Enterobacter aerogenes	LogR	RD (ResFams)	(Pesesky *et al.*, 2016)
Enterococcus faecium	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)
Escherichia coli	LogR	RD (ResFams)	(Pesesky *et al.*, 2016)
	GBT	Population structure, isolation year, gene content	(Moradigaravand *et al.*, 2018)
	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)
	NN, RF	RD (ResFinder+PointFinder)	(Aytan-Aktug *et al.*, 2020)
	RF, LinR	Mutations (raw reads) + RD (ResFinder)	(Pataki *et al.*, 2020)
	SVM	Pangenome	(Hyun, Kavvas, Monk, & Palsson, 2020)
	XGB	RD	(Kim *et al.*, 2019)
Klebsiella pneumoniae	LogR	RD (ResFams)	(Pesesky *et al.*, 2016)
	AdaB	*k*-mers (genome)	(Long *et al.*, 2017)
	XGB	*k*-mers (genome)	(Nguyen *et al.*, 2018)
	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)
	XGB	RD	(Kim *et al.*, 2019)
	SCM, RF	*k*-mers (genome)	(Hicks *et al.*, 2019)
	XGB	Conserved genes (non-AMR)	(Nguyen, Olson, Shukla, VanOeffelen, & Davis, 2020)
	RF	Partial genome alignments	(Aytan-Aktug *et al.*, 2021)
Mycobacterium tuberculosis	AdaB	*k*-mers (genome)	(Davis *et al.*, 2016)
	RF	*k*-mers (genome)	(Santerre *et al.*, 2016)
	LogR, RF,CBMM, SVM	RD	(Yang *et al.*, 2018)

	SVM	Pangenome	(Kavvas *et al.*, 2018)
	NN	RD	(M. L. Chen *et al.*, 2019)
	XGB	Conserved genes (non-AMR)	(Nguyen *et al.*, 2020)
	NN, RF	RD (ResFinder+PointFinder)	(Aytan-Aktug *et al.*, 2020)
	RF	Partial genome alignments	(Aytan-Aktug *et al.*, 2021)
	XGB	Conserved genes (non-AMR)	(Nguyen *et al.*, 2020)
Neisseria gonorrhoeae	LinR	RD (custom)	(Demczuk *et al.*, 2016)
	LinR	RD (custom)	(Eyre *et al.*, 2017)
	LinR	RD (custom)	(Eyre, Golparian, & Unemo, 2019)
	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)
	SCM, RF	*k*-mers (genome)	(Hicks *et al.*, 2019)
	LinR	RD (custom)	(Demczuk *et al.*, 2020)
Peptoclostridium difficile	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)
Pseudomonas aeruginosa	SVM	Pangenome +RD	(Hyun *et al.*, 2020)
	XGB	RD	(Kim *et al.*, 2019)
Salmonella enterica	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)
	XGB	RD	(Kim *et al.*, 2019)
	XGB	Conserved genes (non-AMR)	(Nguyen *et al.*, 2020)
	NN	RD (ResFinder+PointFinder)	(Aytan-Aktug *et al.*, 2020)
	RF	Partial genome alignments	(Aytan-Aktug *et al.*, 2021)
Nontyphoidal *Salmonella enterica*	XGB	*k*-mers (genome)	(Nguyen *et al.*, 2019)
	LogR, SCM	RD	(Maguire *et al.*, 2019)
Staphylococcus aureus	RF	Assembly (Custom database)	(Alam *et al.*, 2014)
	RF	*k*-mers (genome)	(Santerre *et al.*, 2016)
	AdaB	*k*-mers (genome)	(Davis *et al.*, 2016)
	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)
	XGB	RD	(Kim *et al.*, 2019)
	XGB	Conserved genes (non-AMR)	(Nguyen *et al.*, 2020)
	SVM	Pangenome +RD	(Hyun *et al.*, 2020)
	NN	RD (ResFinder+PointFinder)	(Aytan-Aktug *et al.*, 2020)
Staphylococcus haemolyticus	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)
Streptococcus pneumoniae	AdaB	*k*-mers (genome)	(Davis *et al.*, 2016)
	RF	*k*-mers (genome)	(Santerre *et al.*, 2016)
	RF	Penicillin-binding proteins	(Y. Li *et al.*, 2017)
	CART, SCM	*k*-mers (genome)	(Drouin *et al.*, 2019)

RESEARCH GOALS

My overarching research goal was to improve AMR genotype-phenotype prediction, examining the impact of different data, feature, algorithm, and evaluation choices. In the new and active field of genomics-based diagnostic microbiology, broadly testing new and existing methods is the only way to reveal knowledge and, eventually, achieve a gold standard clinical tool.

Only one published study uses CARD and RGI to predict AMR phenotypes. Even for database-independent methods, there are only a few studies that use known resistance determinants as features in machine learning, whereas most others use *k*-mers. For database-independent methods, we have applied different machine learning methods to identify which result in highest accuracy, yet also have interpretable models. Interpretable machine learning models are essential to understanding which genetic determinants are driving phenotypic resistance and to expand our knowledge of the underlying resistance mechanisms. I explored machine learning using different genetic features as a basis for predicting resistance, such as known resistance determinants versus all observed mutations (base substitutions, insertions, and deletions relative to a reference genome). Thus, a major goal of this thesis was to compare database dependent and independent methods for prediction of resistance phenotypes for clinical isolates and to elucidate the underlying AMR genotype-phenotype relationships.

While it is often sufficient to predict ‘resistant’ and ‘susceptible’ classifications, this method can be imprecise as these classifications are determined based on the minimum inhibitory concentration (MIC) value relative to a clinical breakpoint that is

different depending on the chosen guideline. For example, the piperacillin resistant MIC breakpoint for *Enterobacteriaceae* is different for the EUCAST guideline (>16μg/mL) compared to the CLSI guideline (≥128μg/mL). Another limitation is the lack of curation clarity as any resistance determinant that exemplifies an elevated MIC value compared to control can be curated into CARD, but the value of elevation is not recorded. Thus, detecting a CARD resistance determinant assumes the isolate is resistant; however, the MIC elevation in the literature may not surpass a clinical breakpoint value of 'resistant.' As a result, the second major aim of this thesis was to predict MIC values using machine learning, while maintaining the ability to learn genotype-MIC associations.

CHAPTER TWO: Identifying novel β-lactamase substrate activity through *in silico* prediction of antimicrobial resistance

CHAPTER TWO PREFACE

Portions of this work presented in this chapter have been published as:

Tsang KK, Maguire F, Zubyk HL, Chou S, Edalatmand A, Wright GD, Beiko RG, McArthur AG. Identifying novel β-lactamase substrate activity through *in silico* prediction of antimicrobial resistance. *Microbial Genomics*. 2021 Jan 8:000500.

Author contributions: KKT and AGM conceived the project. KKT and FM designed bioinformatics experiments. KKT, HLZ, and SC designed and performed the gene expression experiments. AGM curated clinical isolate phenotypes. KKT performed the analysis. KKT and AGM wrote the manuscript. All authors revised the manuscript. KKT wrote this chapter.

Acknowledgements

This research was funded by the Canadian Institutes of Health Research (PJT-156214 to A. G. M., MT-14981 to G. D. W.), the Ontario Research Fund (to G. D. W.), Genome Canada (to R. G. B.), a Canada Research Chair to G. D. W. and a Cisco Research Chair in Bioinformatics to A. G. M., supported by Cisco Systems Canada, Inc. K. K. T. was supported by an Ontario Graduate Scholarship, McMaster University's MacDATA Institute Graduate Fellowship and Michael G. DeGroote Institute for Infectious Disease Research Michael Kamin Hart Memorial Scholarship. F. M. was supported by a Donald Hill Family Fellowship in Computer Science. Computer resources were supplied by the McMaster Service Lab and Repository computing cluster, funded in part by grants to A. G. M. from the Canadian Foundation for Innovation (34531).

ABSTRACT

Diagnosing antimicrobial resistance (AMR) in the clinic is based on empirical evidence and current gold standard laboratory phenotypic methods. Genotypic methods have the potential advantages of being faster and cheaper, and having improved mechanistic resolution over phenotypic methods. We generated and applied rule-based and logistic regression models to predict the AMR phenotype from *Escherichia coli* and *Pseudomonas aeruginosa* multidrug-resistant clinical isolate genomes. By inspecting and evaluating these models, we identified previously unknown β-lactamase substrate activities. In total, 22 unknown β-lactamase substrate activities were experimentally validated using targeted gene expression studies. Our results demonstrate that generating and analysing predictive models can help guide researchers to the mechanisms driving resistance and improve annotation of AMR genes and phenotypic prediction, and suggest that we cannot solely rely on curated knowledge to predict resistance phenotypes.

INTRODUCTION

Antimicrobial resistance (AMR) is a global health crisis accelerated by overuse and misuse of antimicrobials. Amongst Gram-negative pathogens, AMR *Escherichia coli* and *Pseudomonas aeruginosa* are of urgent and critical concern. The World Health Organization has reported high resistance to fluoroquinolones and third-generation cephalosporins when treating urinary tract *E. coli* infections, leading to reliance on carbapenems as a last-resort treatment option (World Health Organization, 2014), while the US Centers for Disease Control and Prevention estimates nearly 32600 antibiotic-resistant *P. aeruginosa* infection-related hospitalizations in the USA alone in 2017, to which 2700 deaths were attributed (CDC, 2019).

Currently, the gold standards for diagnosing antibiotic resistance are culture-based phenotypic methods. However, the turnaround time for antibiotic susceptibility tests often surpasses the optimal time for life-threatening infection treatment (Maugeri, Lychko, Sobral, & Roque, 2019; Maurer, Christner, Hentschke, & Rohde, 2017) Furthermore, phenotypic tests do not reveal the genetic underpinnings of resistance. As such, genotypic methods that exploit high-throughput DNA sequencing technology combined with bioinformatics resources have the potential to be faster and more accurate and informative than the current phenotypic paradigm (Chan, 2016). There is growing momentum toward whole-genome sequencing of clinical infections, but there is a lag in the development of bioinformatic platforms that can accurately predict phenotypes such as virulence and AMR, which is essential for the full application of rapid pathogen sequencing as a robust diagnostic tool. Most sequencing pipelines rely on an AMR sequence database to predict

functional AMR genes from DNA sequences (Crofts, Gasparrini, & Dantas, 2017), of which there are many. For example, the Comprehensive Antibiotic Resistance Database (CARD) is an ontology-driven genomics database used by the Resistance Gene Identifier (RGI) software to predict intrinsic and acquired resistance determinants in genome sequences (Alcock *et al.*, 2020). The Antibiotic Resistance Gene-ANNOTation database (Gupta *et al.*, 2014) and Pathosystems Resource Integration Center (Wattam *et al.*, 2017) store a similar breadth of resistance determinants to CARD and also use blast-based tools for resistome annotations. Antibiotic Resistance Genes Online (Gupta *et al.*, 2014) only catalogues β-lactam and vancomycin resistance determinants, in comparison to ResFinder (Zankari *et al.*, 2012), which primarily annotates acquired resistance genes using BLASTN, while ResFams (Gibson *et al.*, 2015) is a database of protein domain hidden Markov models associated with AMR function.

Despite our dependence upon curated AMR databases for genotype analysis and prediction of phenotypes, maintaining and developing AMR databases and tools are challenging due to the ever-evolving AMR genetic landscape, inconsistencies in AMR gene nomenclature, sparsity of phenotypic data and lack of funding for biocuration (McArthur & Tsang, 2017; van Belkum *et al.*, 2019). Without comprehensiveness in phenotypic testing, such as antibiotic susceptibility testing using a broad panel of antibiotics, all of these databases will inherently be missing the full range of a resistance determinant's substrate specificity. Yet, as β-lactams are the most commonly used antibiotic (Cantu, Huang, & Palzkill, 1997), there is strong motivation in the AMR field to identify the substrate specificity of clinically prevalent β-lactamases (Cantu *et al.*,

1997; Chiou, Leung, & Chen, 2014; Jacquier *et al.*, 2013; Khan, Sallum, Zheng, Nau, & Hasan, 2014; Majiduddin & Palzkill, 2005), particularly with regard to β-lactams new to the marketplace. Despite the development of gene-based antibiotic susceptibility testing tools such as the Antibiotic Resistance Platform (G. Cox *et al.*, 2017), when novel β-lactamases emerge in clinical settings they are often only characterized using a limited selection of β-lactams, or are assumed to have similar substrate activity to a related β-lactamase. This leads to knowledge gaps in AMR databases for β-lactamase substrate specificity. In the face of missing experimental data, the prediction of novel substrate specificities for known β-lactamases can be performed using statistical modelling and machine learning methods (Davis *et al.*, 2016; Drouin *et al.*, 2016; Pesesky *et al.*, 2016). While these statistical models can be used to discover novel genotype–phenotype relationships, they often require large and diverse datasets to be effective. Previous studies have used rule-based and statistical models to predict antibiotic resistance phenotypes from genotypes, but only a few studies provide genotype–phenotype associations (Davis *et al.*, 2016; Drouin *et al.*, 2016; Pesesky *et al.*, 2016).

Here we report the *in silico* prediction of genotype–phenotype associations and substrate specificities for AMR determinants from multidrug-resistant *E. coli* and *P. aeruginosa* clinical isolates using two computational approaches (rules-based and logistic regression) based upon CARD's RGI (Alcock *et al.*, 2020). The rules-based method uses new software (the Efflux Pump Identifier) to account for overexpressed multi-component efflux pumps as well as hand-curated knowledge encoded by CARD's Antibiotic Resistance Ontology (ARO). This method helped identify gaps in CARD's curated

knowledge of β-lactam substrate activity that contributed to poor β-lactam resistance phenotype prediction. We then performed logistic regression on the same data, observing higher prediction accuracy across most antibiotic resistance phenotypes. We were then able to experimentally validate the predicted genotype–phenotype relationships (i.e., learned weights) used by logistic regression to identify previously unknown β-lactamase substrate activities.

METHODS

Bacterial isolates, antibiotic susceptibility testing, and DNA extraction

Clinical bacterial isolates were obtained from the IIDR Clinical Isolate Collection, which consists of isolates from the core clinical laboratory at Hamilton Health Sciences, Hamilton, Ontario. Samples were collected between 2015 and 2018 and were resistant to 3 or more antibiotics based on antimicrobial susceptibility to 18 and 17 antibiotics for *E. coli* and *P. aeruginosa*, respectively. As ertapenem lacks activity against *P. aeruginosa* (Livermore, Sefton, & Scott, 2003), it was not included in *P. aeruginosa* antibiotic susceptibility tests. Initial culture and antibiotic susceptibility testing (AST) were performed by Hamilton General Hospital General Microbiology Laboratory using a VITEK 2 Automated System and its Advanced Expert System (BioMérieux, Marcy-l′Étoile, France), compliant with the Clinical and Laboratory Standards Institute (CLSI) (CLSI, 2018) antibiotic susceptibility testing formulations, reporting CLSI breakpoint-determined susceptible (S), intermediate (I), or resistant (R). For DNA extraction, isolates were provided on blood agar plates and single colonies were restreaked onto brain heart infusion (BHI) agar. After overnight incubation, single colonies of each isolate were used to inoculate Luria–Bertani (LB) broth. Overnight broth cultures were used to prepare glycerol stocks for long-term storage at −80 °C. One millilitre of the same overnight cultures was centrifuged, the supernatant was removed and the pellet was stored at −80 °C for genomic DNA extraction. The Invitrogen Pure Link Genomic DNA Mini kit (K182002) was used for DNA extraction from pellets. DNA was eluted with water and stored at 4 °C.

Whole-genome sequencing, assembly and species identification

DNA sequencing library construction (Illumina Nextera XT DNA Library Preparation kit or NEBNext Ultra II DNA Library Preparation kit) and all sequencing runs were performed at the Farncombe Metagenomics Facility at McMaster University using 2×150 bp paired-end sequencing on an Illumina HiSeq 1500 platform (*E. coli* n=115, *P. aeruginosa* n=92) or 2×250 bp paired-end sequencing on an Illumina MiSeq v3 platform (*P. aeruginosa* n=10). Paired sequencing reads were trimmed using Trimmomatic (v0.36) (Bolger *et al.*, 2014), checked for quality using fastqc (v0.11.8, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Andrews, 2010) and de novo assembled using SPAdes (v3.9.0) (Bankevich *et al.*, 2012). The Livermore Metagenomics Analysis Toolkit (lmat, v1.2.6) (Ames *et al.*, 2013) was used to confirm bacterial species and screen for contamination or mixed culture. For *E. coli*, after quality trimming of the sequencing reads by Trimmomatic, sequencing isolate read coverage averaged 207.5-fold, assembly size averaged ~5,163,879 bp and N50s averaged 231,879 bp. For *P. aeruginosa*, quality-trimmed sequencing read coverage averaged 100.6-fold, assembly sizes averaged 6,680,703 bp and assembly N50s averaged 260,849 bp. Diversity of isolates for both *E. coli* and *P. aeruginosa* was assessed by multilocus sequence typing (MLST) via comparison to the reference sequences available at pubMLST (https://github.com/agmcarthur/pubMLST) (Jolley, Bray, & Maiden, 2018).

Curation of CARD

At minimum, CARD requires the curation of a '*confers_resistance_to_drug_class*' relationship between an AMR gene family and a drug class in the ARO. However, to predict specific drug resistance phenotypes we needed curation of a '*confers_resistance_to_antibiotic*' relationship between an individual resistance gene or mutation and a specific antibiotic. The curation of '*confers_resistance_to_antibiotic*' relationships is incomplete in CARD and is determined by experimental evidence of an elevation of MIC in the published literature (Alcock *et al.*, 2020). Using extensive literature review, we curated '*confers_resistance_to_antibiotic*' relationships for all resistance determinants identified as RGI Perfect or Strict RGI hits for our *E. coli* and *P. aeruginosa* isolates: an additional 250 '*confers_resistance_to_antibiotic*' relationships (152 *E. coli* and 98 *P. aeruginosa*) were added to CARD (available as of v2.0.2). During the curation process we also identified two errors in CARD curation. These included incorrect inclusion of mutation Y45C in the *E. coli* protein NfsA as conferring resistance to nitrofurantoin and the β-lactamase gene SHV-1 as conferring resistance to cefazolin. In both cases, the original publications lacked clear experimental support for these claims.

To additionally improve efflux pump prediction and facilitate the functionality of the Efflux Pump Identifier (EPI), *E. coli* and *P. aeruginosa* efflux meta-models (a combination of individual models) were curated into CARD v1.1.9, based on review of the literature. Efflux meta-models comprise protein homologue and/or protein overexpression models to represent a known efflux pump complex and its regulatory

network. For example, the AcrAB-TolC efflux system (ARO:3000384) is encoded along with its regulatory network: *marR, marA, acrR, sdiA, soxS, soxR,* and *rob*. In this meta-model, each component is a protein homologue model with the exception of *marR, acrR,* and *soxS,* which are protein overexpression models. We curated 21 *P. aeruginosa* efflux pump meta-models, 10 *E. coli* efflux pump meta-models and 2 plasmid-borne efflux pump meta-models known to confer resistance to the 18 antibiotics tested in this study for analysis by EPI.

Rules-based prediction of antibiotic susceptibility phenotypes

Isolate genomes were analysed using the Comprehensive Antibiotic Resistance Database (v2.0.2) and Resistance Gene Identifier (v4.1.0) (Alcock *et al.*, 2020), plus the new EPI (v1.0.0) software developed by KKT, to predict resistance determinants. The EPI predicts multi-component efflux pumps and their regulatory networks using the efflux meta-models curated in CARD (https://git.io/JJFhT). RGI and EPI results were filtered to only include RGI Perfect and Strict hits, and EPI Perfect and Partial hits, respectively. Antibiotic susceptibility phenotypes were predicted by traversing CARD's Antibiotic Resistance Ontology (ARO) to identify the antibiotic(s) each detected resistance determinant confers resistance to, based on peer-reviewed literature. In this rules-based method, the detection of a resistant determinant by RGI or EPI that had a '*confers_resistance_to_antibiotic*' relationship to an antibiotic in the ARO resulted in a 'resistant' phenotype prediction, otherwise a 'susceptible' phenotype was predicted. Computational antibiotic susceptibility predictions were then compared to clinical ASTs.

As AST 'intermediate' resistances were rare (2.2% of *P. aeruginosa* resistance phenotypes and 3.6% of *E. coli* resistance phenotypes), we treated them as 'resistant' in our analyses.

Using logistic regression to predict antibiotic resistance phenotypes

To prepare the datasets, all RGI results for each species were collated into count matrices Xij where i represents each genome of that species and j represents a specific AMR determinant detected by RGI at either Strict or Perfect cut-offs. There were 189 and 133 resistance determinants in the *E. coli* and *P. aeruginosa* matrix, respectively. The most appropriate algorithm for phenotype prediction was determined using the *E. coli* data, as these comprised the more balanced dataset. For each antibiotic, the resampled training data were used to fit four interpretable binary classification models: logistic regression, multinomial naïve Bayes, decision tree and random forest classifiers (Pedregosa, Varoquaux, Gramfort, & Michel, 2011). For each model the hyperparameters were then tuned using a threefold stratified shuffle split cross-validation scheme and evaluated using a negative log loss scoring function (Pedregosa *et al.*, 2011), as negative log loss considers prediction uncertainty in relation to the divergence of the predicted probabilities and the actual AMR phenotype. Logistic regression and random forest classifiers had the highest performance of all tested modelling methods, so we chose logistic regression, a simpler algorithm, as our classification paradigm under the principle of parsimony. To predict each antibiotic resistance phenotype, antibiotic-specific LR models were trained, optimized via cross-validation and tested separately for each species

dataset. To determine whether each species and antibiotic dataset was phenotypically balanced enough for LR, the relative proportion of resistant predictions to susceptible predictions was evaluated. If the less frequent phenotype represented <10% of all genomes it was considered inappropriate to train and properly test a model due to extreme class imbalance and low signal. For these antibiotics an 'unbalanced classifier' was trained and evaluated using all genomes of that species. Some antibiotics displayed an even more extreme case of imbalance where only a single phenotype was observed. For these, a 'dummy' model was used that only returned the observed phenotype (i.e. all observed isolates were resistant to an antibiotic and therefore the model always predicts resistance). For the remaining species-antibiotics combinations with greater label balance, 20% of the genomes were randomly selected with stratification (i.e. maintaining the relative proportion of susceptible to resistant) and withheld as a test set. The training set was then rebalanced using the synthetic minority over-sampling technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) as implemented in imbalanced-learn (v0.3.3) (Lemaître, Nogueira, Learning, & 2017, 2017) to generate a training set with equal proportions of susceptible and resistant genomes. After training of the *E. coli* models, the *P. aeruginosa* training data were used to fit and optimize logistic regression models via the same threefold stratified cross-validation scheme.

The individual trained antibiotic–species logistic regression models (including unbalanced and dummy classifiers) were evaluated against the test set to see if they could predict AMR phenotypes, with evaluation using precision–recall curves (summarized as average precision) and the receiver operating characteristic (summarized as area under the

curve) (Supplementary Figure 2-1 to 2-3) (Saito & Rehmsmeier, 2015). A test with perfect discrimination between resistance and susceptible phenotypes would have a receiver operating characteristic curve that passes through the upper-left corner (Supplementary Figure 2-1, 2-2). For each species the number of true positives, true negatives, false positives and false negatives was tallied and plotted for each antibiotic. To evaluate which resistance determinants within each classifier were important for predicting resistance phenotypes, we considered the estimated coefficients (scikit-learn's coef_attribute) as the 'weight of importance' for each resistance determinant. Thus, given two resistance determinants, each with an estimated coefficient value, the resistant determinant with a larger estimated coefficient value was interpreted as more important for predicting a particular resistance phenotype. The five most highly weighted predictors of each resistance phenotype were examined (Supplementary Figure 2-4,2-5), but all feature weights of importance and their P-values were inspected and are listed in Supplementary Tables S2–S5 available online: https://doi.org/10.1099/mgen.0.000500.

Antibiotic susceptibility testing (AST) using the Antibiotic Resistance Platform

In cases where we wished to perform AST for individual resistance genes, we cloned these genes into pGDP1/pGDP3 from the Antibiotic Resistance Platform (G. Cox *et al.*, 2017) and transformed into wild-type *E. coli* BW25113. AST was performed for *E. coli* BW25113 using the microdilution broth method, with the inoculum prepared using the growth method following CLSI guidelines (CLSI, 2018). Plates were sealed in a bag

and incubated for 18 h at 37 °C, 250 r.p.m. before the optical density at 600 nm was measured using the Spectramax microplate reader.

Software availability

CARD data and RGI software are available at the CARD website, http://card.mcmaster.ca. CARD (v2.0.2) and RGI (v4.1.0) were used for all resistome prediction, and RGI (v.5.1.0) was used for creating the heatmaps. The EPI software is available at https://github.com/karatsang/rulesbased_logisticregression/tree/v1.0.0/rulesbased/EffluxPumpIdentifier. LR and dataset partitioning were performed using scikit-learn (v0.20.0) (Pedregosa *et al.*, 2011) with data otherwise manipulated using numpy (v1.17.2) (Oliphant, 2006) and pandas (v0.25.1) (McKinney, 2010). For both datasets, the code, conda environments (using python v3.7.2 (van Rossum & Drake, 2003)), and intermediate data files required to generate this analysis are available: https://github.com/karatsang/rulesbased_logisticregression, https://doi.org/10.5281/zenodo.3988480.

RESULTS

Bacterial isolates, antibiotic susceptibility testing (AST), and whole-genome sequencing

In total, 115 *E. coli* and 102 *P. aeruginosa* putative multidrug-resistant clinical isolates were obtained from Hamilton Health Sciences hospitals (Hamilton, Ontario, Canada) and submitted for both genome sequencing and AST, i.e. categorized as 'resistant' or 'susceptible' for 18 antibiotics under Clinical and Laboratory Standards Institute (CLSI) guidelines. Among the isolates, 20 *E. coli* had no resistance to any of the tested antibiotics and all of the *P. aeruginosa* strains were resistant to at least 1 drug. Seventy-four *E. coli* and 101 *P. aeruginosa* isolates were resistant to 3 or more antibiotics. The antibiotics tested and the full AST results are summarized in https://github.com/karatsang/rulesbased_logisticregression/tree/v1.0.0/AST. In the *E. coli* dataset there were 30 unique multilocus sequence types (MLSTs) and 5 isolates with unresolved MLST allele(s). The 2 most prevalent *E. coli* MLSTs in the dataset were ST131 and ST1193, which 39 and 10 clinical isolates belonged to, respectively. Notably, ST131 is known to be a major cause of multidrug-resistant *E. coli* infections in the USA (Johnson, Johnston, Clabots, Kuskowski, & Castanheira, 2010) and a globally dominant clone (Pitout & DeVinney, 2017) associated with CTX-M β-lactamases, while ST1193 is a newer multidrug-resistant *E. coli* clonal group (2017–2019) associated with both CTX-M β-lactamases, plasmid-borne TEM-1 and aminoglycoside acetyltransferases (AACs) (Tchesnokova *et al.*, 2019; Wu, Lan, Lu, He, & Li, 2017; Xia *et al.*, 2017). In the *P. aeruginosa* dataset there were 59 unique MLSTs (43 known and 16 novel MLSTs) and 3

isolates with unresolved MLST allele(s). The three most prevalent MLSTs, ST244, ST235 and ST253, were identified in five *P. aeruginosa* isolates each. *P. aeruginosa* ST244 is an international clone, many isolates of which are multidrug-resistant (Y. Chen, Sun, Wang, Lu, & Yan, 2014; Empel *et al.*, 2007), ST235 is amongst the most prevalent of international clones originating from Europe, with regional acquisition of AMR genes (Treepong *et al.*, 2018), and ST253 a less common clone associated with multidrug resistance in Spain and Greece (Koutsogiannou *et al.*, 2013). The full MLST results are summarized in https://github.com/karatsang/rulesbased_logisticregression/tree/v1.0.0/MLST. Raw Illumina DNA sequencing reads for each isolate are available through National Center for Biotechnology Information (NCBI) BioProject PRJNA532924.

Rules-based interpretation leads to poor β-lactam phenotype prediction

Our rules-based algorithm relies on the resistome predictions of CARD's RGI and the genotype–phenotype relationships curated in CARD's ARO. RGI uses four bioinformatics models to predict the resistome of a clinical isolate, which are the protein homology, protein variant, rRNA variant and protein overexpression models (detailed at https://github.com/arpcard/rgi). The protein homology model detects a protein sequence based on its similarity to a curated reference sequence in CARD. The protein variant model builds on the protein homologue model to identify curated mutations that are shown to confer resistance in antibiotic targets, while the rRNA variant model performs the same function for mutations conferring resistance to antibiotics targeting ribosomal

RNAs. The protein overexpression model identifies proteins with or without mutations which reflects regulatory proteins that are functional without a mutation, but confer overexpression of their targets with a mutation. As CARD's RGI software is unable to predict multi-component efflux pump systems important for AMR, we developed the Efflux Pump Identifier (EPI) software to interpret RGI results for the prediction of overexpressed efflux pump systems, classifying them into three categories: Perfect, Partial and Putative. The Perfect category identifies sequence matches to CARD for all components of a predicted efflux multi-component system. The Partial category identifies all components of an efflux multi-component system, but at least one component is a sequence variant of CARD's reference sequence. The Putative category predicts potential efflux multi-component systems with missing components or otherwise entirely composed of previously uncharacterized sequence variants.

For our analyses we used the above models and RGI's Perfect and Strict criteria, supplemented with the EPI's interpretation of efflux complexes, to predict resistomes from isolate genome sequences. RGI's Perfect criterion requires that a query protein sequence be identical to a curated reference sequence in CARD, while Strict detects variants of known resistance determinants that pass a curated bit-score cut-off (protein homologue model) or a known AMR-conferring mutation (protein variant model) that can be found curated within CARD (card.mcmaster.ca). The predicted resistomes of the individual *P. aeruginosa* and *E. coli* isolates were generally unique and contained a large diversity of resistance determinants (Table 2-1, also see https://git.io/JJFh3), with the

exceptions being two groups of three *P. aeruginosa* isolates and five *E. coli* isolates that had the same predicted resistome, respectively.

Table 2-1. The prevalence of Perfect and Strict resistance determinants detected by the Resistance Gene Identifier, organized by the Antibiotic Resistance Ontology (ARO) drug class designations. Columns show number and percentage of sampled isolates having at least one AMR determinant associated with resistance to each drug class, broken down as harbouring efflux, non-efflux determinants, or both. For example, 98% of all *P. aeruginosa* isolates had a least one resistance gene for rifamycin resistance, with 99 isolates predicted to have only efflux gene(s) conferring resistance to rifamycin and a single isolate predicted to have only a non-efflux determinant of rifamycin resistance. The total number of *E. coli* and *P. aeruginosa* isolates is 115 and 102, respectively.

ARO Drug Class	# of *E. coli* isolates (non-efflux + efflux + both)	% of *E. coli* isolates	# of *P. aeruginosa* isolates (non-efflux + efflux + both)	% of *P. aeruginosa* isolates
acridine dye	0 + 115 + 0	100.0%	0 + 102 + 0	100.0%
aminocoumarin antibiotic	0 + 114 + 1	100.0%	0 + 101 + 1	100.0%
aminoglycoside antibiotic	0 + 44 + 71	100.0%	0 + 0 + 102	100.0%
benzalkonium chloride	0 + 115 + 0	100.0%	0 + 1 + 0	1.0%
bicyclomycin	0 + 1 + 0	0.9%	0 + 102 + 0	100.0%
carbapenem	0 + 0 + 115	100.0%	0 + 0 + 102	100.0%
cephalosporin	0 + 0 + 115	100.0%	0 + 0 + 102	100.0%
cephamycin	0 + 0 + 115	100.0%	0 + 101 + 1	100.0%
diaminopyrimidine antibiotic	50 + 1 + 3	47.0%	0 + 101 + 1	100.0%
elfamycin antibiotic	115 + 0 + 0	100.0%	2 + 0 + 0	2.0%
fluoroquinolone antibiotic	0 + 42 + 73	100.0%	0 + 67 +35	100.0%
fosfomycin	0 + 111 + 4	100.0%	102 + 0 + 0	100.0%
fusidic acid	0 + 1 + 0	0.9%	0 + 0 + 0	0.0%
glycopeptide antibiotic	0 + 111 + 4	3.5%	2 + 0 + 0	2.0%
glycylcycline	0 + 115 + 0	100.0%	0 + 100 + 0	98.0%
lincosamide antibiotic	4 + 68 + 3	65.2%	3 + 1 + 0	3.9%
macrolide antibiotic	0 + 60 + 55	100.0%	0 + 0 + 102	100.0%
monobactam	0 + 0 + 115	100.0%	0 + 0 + 102	100.0%
mupirocin	0 + 0 + 0	0.0%	1 + 0 + 0	1.0%
nitrofuran antibiotic	115 + 0 + 0	100.0%	0 + 2 + 0	2.0%
nitroimidazole antibiotic	0 + 115 + 0	100.0%	0 + 0 + 0	0.0%

nucleoside antibiotic	0 + 112 + 3	100.0%	0 + 1 + 0	1.0%
nybomycin	72 + 0 + 0	62.6%	21 + 0 + 0	20.6%
oxazolidinone antibiotic	0 + 0 + 0	0.0%	1 + 0 + 0	1.0%
penam	0 + 0 + 115	100.0%	0 + 0 + 102	100.0%
penem	0 + 65 + 50	100.0%	0 + 99 + 3	100.0%
peptide antibiotic	0 + 0 + 115	100.0%	0 + 0 + 0	100.0%
phenicol antibiotic	0 + 91 + 24	100.0%	0 + 1 + 101	100.0%
pleuromutilin antibiotic	39 + 0 + 0	33.9%	1 + 0 + 0	1.0%
rhodamine	0 + 115 + 0	100.0%	0 + 1 + 1	1.0%
rifamycin antibiotic	0 + 115 + 0	100.0%	0 + 99 + 1	98.0%
streptogramin antibiotic	42 + 0 + 0	36.5%	3 + 0 + 0	2.9%
sulfonamide antibiotic	67 + 0 + 0	58.3%	0 + 94 + 8	100.0%
sulfone antibiotic	67 + 0 + 0	58.3%	8 + 0 + 0	7.8%
tetracycline antibiotic	0 + 112 + 3	100.0%	0 + 99 + 3	100.0%
triclosan	0 + 114 + 1	100.0%	0 + 102 + 0	100.0%

In the *P. aeruginosa* clinical isolate dataset, RGI detected 4 Perfect and 38 Strict, non-efflux, unique resistance genes (protein homologue models) across 34 of CARD's drug classes, plus 4 unique, non-efflux mutations (protein variant models) known to confer resistance to particular antibiotics (ParE A473V, GyrA T83I, BasR L71R and EF-Tu R234F). In the *E. coli* dataset, RGI detected 31 Perfect and 59 Strict non-efflux, unique resistance genes (protein homologue models), plus 15 unique, non-efflux mutations or combinations of mutations (protein variant models) known to confer resistance to particular antibiotics (UhpT E350Q; ParC S80I, E84G; EF-Tu R234F; PBP3 D350N, S357N; GlpT E448K; GyrB S464Y; GyrA D87Y, D87G, D87N, S83L; CyaA S352T; PtsI V25I; NfsA Y45C). For efflux, in *P. aeruginosa* there were 2 unique Perfect and 14 Strict and in *E. coli* there were 11 unique Perfect and 34 Strict protein homologue models representing single-component efflux resistance genes. EPI additionally detected

one Perfect or Partial efflux complex with an overexpression mutation (*E. coli* AcrAB-TolC with MarR mutation Y137H conferring resistance to ciprofloxacin and tetracycline) in two different *E. coli* isolates; otherwise, EPI identified six unique Partial efflux pump complexes without an overexpression mutation among the *E. coli* isolates. In contrast, EPI did not identify any Perfect efflux pump complexes among *P. aeruginosa* isolates; however, three unique Partial efflux pump complexes with an overexpression mutation were identified in three different clinical isolates (MexEF-OprN with MexS F253L,V73A; MexAB-OprM with MexR R91C; MexAB-OprM with NalC S209R, G71E, A186T). Supplementary information and citations for all variants predicted by RGI/EPI can be found at CARD.

Comparing the above RGI and EPI resistome predictions, phenotypically classified by CARD's ARO, to the laboratory ASTs, we observed instances of true-positive, true-negative, false-positive and false-negative predictions of AMR phenotype for both *E. coli* and *P. aeruginosa* (Figure 2-1 and 2-2).

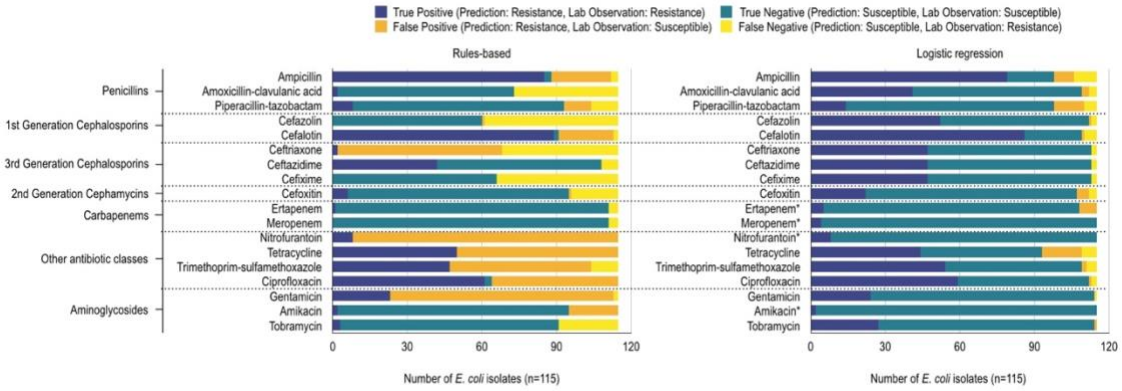


Figure 2-1. True positive, true negative, false positive, and false negative predictions of *E. coli* resistance phenotype using a rules-based (left) and logistic regression method (right). Antibiotic susceptibility tests used eighteen antibiotics organized into their respective drug classes. True positives (dark blue) and true negatives (teal) indicate the classifier predicted resistance and susceptibility correctly. False positives (orange) indicate classifier prediction of resistant but an AST of susceptible. Similarly, false negatives (yellow) indicate classifier prediction of susceptible but an AST of resistant. The rules-based method uses RGI, EPI, and the Antibiotic Resistance Ontology to predict resistance phenotypes. Logistic regression classifiers use RGI detected AMR determinants to predict resistance phenotypes. Logistic regression models for antibiotics for which <10% of a species' isolates displayed susceptible or resistant phenotypes could not be properly validated and tested and as such were trained using all the data (indicated by an asterisk).

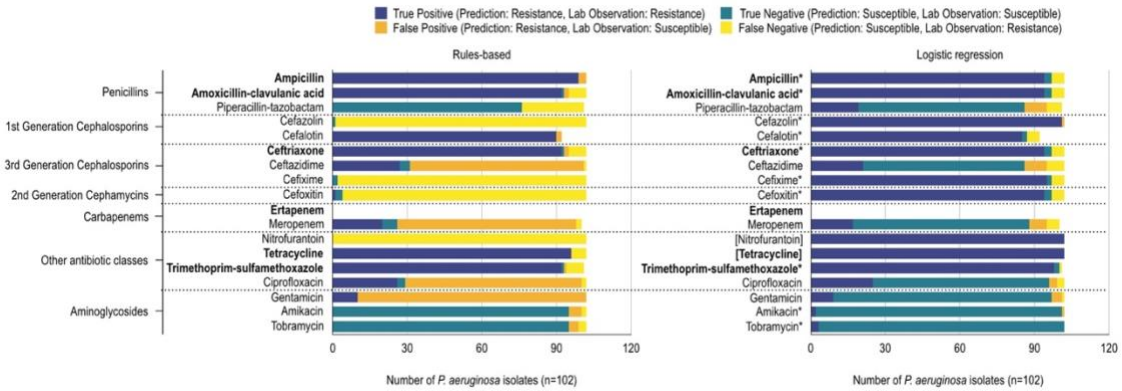


Figure 2-2. True positive, true negative, false positive, and false negative predictions of *P. aeruginosa* resistance phenotype using a rules-based (left) and logistic regression method (right). Antibiotic susceptibility tests used seventeen antibiotics (ertapenem was not tested in *P. aeruginosa*) organized into their respective drug classes. Prediction performances for antibiotic logistic regression classifiers using RGI detected AMR determinants to predict resistance phenotypes for *E. coli* and *P. aeruginosa*. True positives (dark blue) and true negatives (teal) indicates the classifier predicted resistance and susceptibility correctly. False positives (orange) indicate a classifier prediction of resistant but an AST of susceptible. Similarly, false negatives (yellow) indicate a classifier prediction of susceptible but an AST of resistant. The rules-based method uses RGI, EPI, and the Antibiotic Resistance Ontology to predict resistance phenotypes. Logistic regression classifiers use RGI detected AMR determinants to predict resistance phenotypes. Logistic regression models for antibiotics for which <10% of a species' isolates displayed susceptible or resistant phenotypes could not be properly validated and tested and as such were trained using all the data (indicated by an asterisk). Similarly, when all isolates were resistant or susceptible a 'dummy' model was used which always returns the relevant label (placed in square brackets). The bolded antibiotics represent antibiotics that *P. aeruginosa* confer intrinsic resistance towards, according to the Clinical & Laboratory Standards Institute (CLSI). The total of *P. aeruginosa* phenotype predictions does not always equal the total number of isolates (n=102) because not all isolates were tested against every antibiotic.

No antibiotic resistance phenotypes were predicted with 100% accuracy (defined as the percentage of correctly classified phenotypes). Most of the penicillin and cephalosporin (amoxicillin/clavulanic acid, piperacillin/tazobactam, cefazolin, ceftriaxone, ceftazidime, cefixime and meropenem) resistance phenotype predictions resulted in false negatives for both *E. coli* and *P. aeruginosa* (i.e. we failed to predict the observed resistance based on genome sequence). In particular, the prediction of both cefazolin and cefixime resistance phenotypes was less than 2% accurate in the *P. aeruginosa* dataset and less than 57% accurate in the *E. coli* dataset. In addition, for *E. coli* the rules-based algorithm failed to predict any of the observed cefazolin and cefixime resistance based on genome sequence (i.e. not a single true-positive result was obtained).

Logistic regression improves AMR phenotype prediction accuracy

A limitation of the rules-based method is that it only uses known and curated information to predict resistance and is thus inherently blind to any unknown AMR genotype–phenotype relationships. To overcome this limitation, we used logistic regression (LR) to independently identify patterns between RGI-predicted AMR determinants and observed AMR phenotypes. For the *E. coli* dataset (n=115) it was possible to train LR classification models, optimized via 3-fold cross-validation, and test them on a set of withheld isolates for 14 out of 18 antibiotics (Figure 2-1). Due to the relative imbalance of resistant versus susceptible isolates for amikacin, ertapenem, meropenem and nitrofurantoin, models trained for these antibiotics required the use of all isolates, preventing the evaluation of model generalizability on a held-out test set. In the

P. aeruginosa dataset, piperacillin/tazobactam, ceftazidime, meropenem, ciprofloxacin and gentamicin resistance prediction models were trained and tested on separate isolates, while nitrofurantoin and tetracycline required use of 'dummy' models (i.e. all isolates were intrinsically resistant) and the remainder of the AMR prediction models were trained on all isolates due to unbalanced sampling of resistant and susceptible isolates (Figure 2-2).

We evaluated model performance using test set average precision (i.e. trapezoidal area under the precision–recall curve) and a model was categorized as very precise if the test set average precision was ≥0.85, relative to previous studies. Generally, our models were very precise with our *E. coli* data, with a test set average precision of ≥0.85 for all antibiotics except amoxicillin/clavulanic acid (0.811), piperacillin/tazobactam (0.435) and cefoxitin (0.385). In contrast, the *P. aeruginosa* dataset was particularly problematic for LR, with the majority of resistance phenotypes being either ubiquitous (tetracycline and nitrofurantoin) or the less-frequent phenotype representing fewer than 10% of isolates (10/17 antibiotics; ertapenem was not evaluated for these isolates) (Figure 2-2). Only five antibiotics had properly fitted and evaluated models for *P. aeruginosa* : ceftazidime, ciprofloxacin, gentamicin, meropenem and piperacillin/tazobactam. These models had either moderate (ciprofloxacin:~0.650), poor (ceftazidime, piperacillin/tazobactam: 0.512, 0.403), or extremely poor (meropenem: 0.227, gentamicin C: 0.196) test set average precision.

Overall, using LR reduced problems of false-positive and false-negative prediction of AMR phenotypes (Figure 2-1 and 2-2). For *P. aeruginosa* cefazolin and

cefixime resistance phenotypes, where the rules-based approach had very few accurate predictions, LR was able to improve accuracy by 92 and 98%, respectively. Similarly, the rules-based method could not predict any true-positive *E. coli* cefazolin and cefixime resistance phenotypes, whereas LR improved accuracy by 45 and 41%, respectively. In both *P. aeruginosa* and *E. coli* datasets, LR reduced the number of false positives in most tested antibiotic resistance phenotypes compared to the rules-based method. Even in the antibiotic resistance phenotypes where the number of false positives increased, prediction accuracy still improved, e.g. *P. aeruginosa* piperacillin/tazobactam resistance and *E. coli* tobramycin resistance (Figure 2-1 and 2-2).

LR models predict novel β-lactamase activity

For every antibiotic resistance phenotype, LR assigns every resistance determinant a weight to estimate its relative contribution to the prediction. We investigated the five most highly weighted predictors for each antibiotic and pathogen to examine the predicted AMR genotype–phenotype relationships. LR weights that confirmed a known relationship (i.e. supported by the published literature and already curated in CARD) for *E. coli* included CTX-M-15 for ceftazidime resistance, *tet(C)* for tetracycline resistance, *aac(3)-IIb* for gentamicin and tobramycin resistance, *dfrA17* for trimethoprim/sulfamethoxazole resistance, and *gyrA* for ciprofloxacin resistance (Figure 2-3a–j) and for *P. aeruginosa* included *mexD* for amoxicillin/clavulanic acid, ceftriaxone, and cefoxitin resistance, *gyrA* for ciprofloxacin resistance, and *mexB* for amikacin resistance (Figure 2-3k–o).

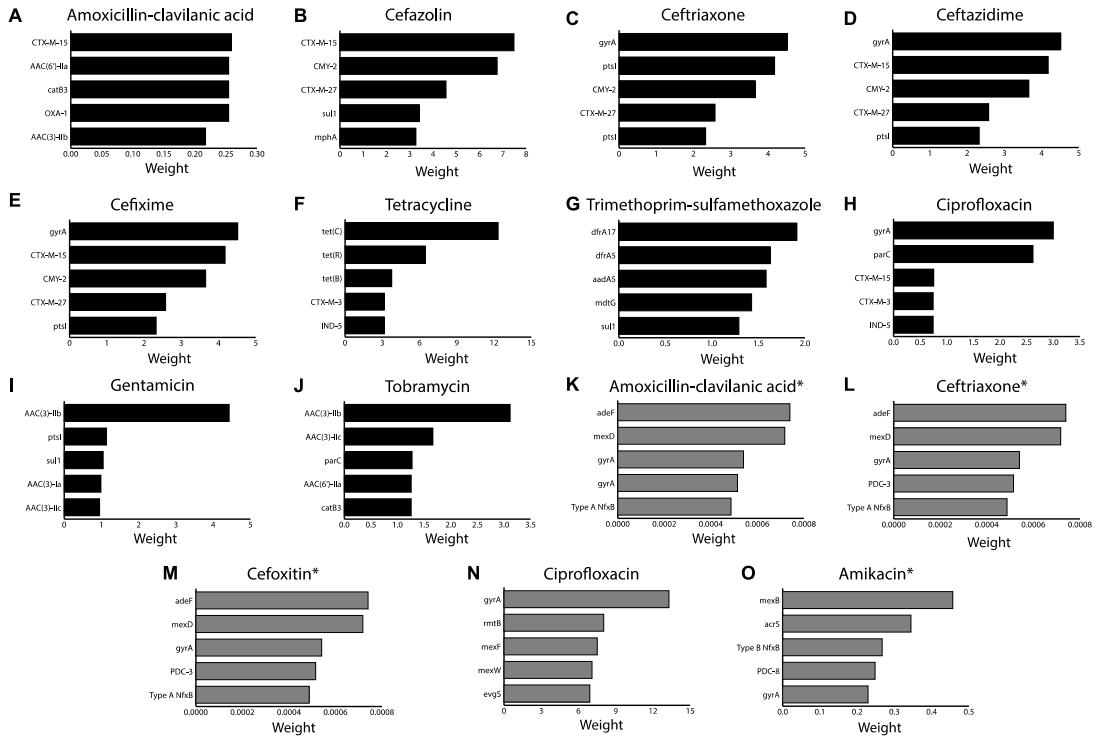


Figure 2-3. Logistic regression and RGI identify resistance determinants for predicting *E. coli* and *P. aeruginosa* resistance phenotypes that are supported by the literature. The x-axes indicate assigned logistic regression weights for individual AMR phenotype predictions, while the y-axes list the top five weighted AMR determinants. Black and grey bars represent *E. coli* and *P. aeruginosa* resistance phenotypes, respectively. An asterisk indicates that <10% of a species' isolates displayed a susceptible or resistant phenotype to amikacin and therefore could not be properly validated and tested, so were trained using all of the data. Models identifying resistance determinants inconsistent with the literature are shown in Supplementary Figures 2-4 and 2-5.

A number of the most highly weighted predictors suggested a previously undocumented substrate specificity for a known β-lactamase, most notably CMY-2 conferring resistance to amoxicillin/clavulanic acid and cefazolin, along with CTX-M-15 conferring resistance to cefixime. To independently test these highly weighted associations, we tested the substrate activity of 11 resistance genes predicted in either the *E. coli* isolates (*aac(6')-Ib-cr, CMY-2, CTX-M-15, CTX-M-3, CTX-M-27, OXA-1, OXA-50, TEM-1* and *TEM-30*) or *P. aeruginosa* isolates (*PDC-3* and *PDC-5*) using the Antibiotic Resistance Platform (ARP) (G. Cox *et al.*, 2017), concluding clinical resistance based on a ≥2-fold elevation in minimum inhibitory concentration (MIC) compared to control that also passed the CLSI Resistant MIC breakpoint value. In total, 22 previously unknown activities between 7 AMR genes and an antibiotic were experimentally validated as clinically relevant in at least 1 pathogen using the ARP and CLSI breakpoints (Table 2-2). These included new knowledge for resistance to ampicillin (CMY-2, CTX-M-3, CTX-M-27, OXA-1 and TEM-30), amoxicillin/clavulanic acid (CMY-2, CTX-M-3, OXA-1 and TEM-1), cefazolin (CMY-2, CTX-M-3, CTX-M-15, CTX-M-27 and TEM-1), cefixime (CMY-2 and CTX-M-3), ceftazidime (CMY-2, CTX-M-3 and CTX-M-27), ertapenem (CTX-M-27) and ceftriaxone (CMY-2 and CTX-M-3). However, none of the tested resistance genes explained the observed resistance to meropenem and an additional four genes only confirmed previous knowledge: AAC(6′)-Ib-cr conferring resistance to tobramycin (Robicsek *et al.*, 2006), TEM-1 conferring resistance to ampicillin (Sutcliffe, 1978), TEM-30 conferring resistance to amoxicillin/clavulanic acid (Belaaouaj *et al.*, 1994) and CTX-M-15 conferring resistance to ceftriaxone (Supplementary Table 2-1)

(Poirel, Gniadkowski, & Nordmann, 2002). ASTs also invalidated some predictions, e.g. CTX-M-15 conferring clinically relevant resistance towards cefixime and ceftazidime. Notably, while OXA-50 is reported to elevate the MIC towards ampicillin and cefotaxime when cloned into a multicopy plasmid and expressed in *P. aeruginosa*, like others (Girlich, Naas, & Nordmann, 2004), we did not observe any appreciable elevation in MIC compared to control in *E. coli* (data not shown). Overall, LR combined with AST validation provided a wealth of new knowledge on antibiotic specificities for β-lactamases appearing in clinical isolates. Interestingly, incorporation of these results into the rules-based algorithm improved resistance prediction in *E. coli* for cefazolin (75% improvement in true-positive results) and cefixime (31% improvement in true-positive results) (Figure 2-4) plus in *P. aeruginosa* for cefixime (34% improvement in true-positive results) and cefoxitin (35% improvement in true-positive results) (Supplementary Figure 2-6), illustrating the sensitivity of rules-based methods to available knowledge. Yet, even with this new knowledge, the rules-based algorithm was still outperformed by the LR approach.

Table 2-2. Antibiotic susceptibility testing (AST) of known resistance genes predicted to have previously undescribed activity. As per the Antibiotic Resistance Platform, AMR genes were cloned into the pGDP plasmid series and transformed into two strains of *E. coli*: wild-type *E. coli* BW25113, which is representative of a clinical isolate. AST was performed for each construct using the microdilution broth method, with the inoculum prepared using the growth method following CLSI guidelines. Dashes indicate lack of CLSI breakpoint for *P. aeruginosa* due to intrinsic resistance. NR: not relevant as *PDC-3* and *PDC-5* were only identified in *P. aeruginosa*.

Antibiotic	Resistance gene	Plasmid	MIC (µg/mL) wild-type *E. coli* BW25113	CLSI Resistant MIC (µg/mL) breakpoint for *Enterobacteriaceae*	CLSI Resistant MIC (µg/mL) breakpoint for *Pseudomonas aeruginosa*
ampicillin	None	None	64	≥32	-
	CMY-2	pGDP1	>256	≥32	-
	CTX-M-3	pGDP1	>256	≥32	-
	CTX-M-27	pGDP1	>256	≥32	-
	OXA-1	pGDP1	>256	≥32	-
	TEM-30	pGDP1	>256	≥32	-
amoxicillin-clavulanic acid	None	None	8-16	≥32/16	-
	CMY-2	pGDP1	256	≥32/16	-
	CTX-M-3	pGDP1	64	≥32/16	-
	CTX-M-15	pGDP1	16	≥32/16	-
	OXA-1	pGDP1	64	≥32/16	-
	TEM-1	pGDP1	128	≥32/16	-
cefazolin	None	None	4	≥8/≥32 (Urine only)	-
	CMY-2	pGDP1	>256	≥8/≥32 (Urine only)	-
	CTX-M-3	pGDP1	>256	≥8/≥32 (Urine only)	-
	CTX-M-27	pGDP1	>256	≥8/≥32 (Urine only)	-
	TEM-1	pGDP1	256	≥8/≥32 (Urine only)	-
cefixime	None	None	0.25	≥4	-
	CMY-2	pGDP1	>256	≥4	-
	CTX-M-3	pGDP1	32	≥4	-
ceftazidime	None	None	0.5	≥16	≥32
	CMY-2	pGDP1	256	≥16	NR
	CTX-M-3	pGDP1	16-32	≥16	NR

	CTX-M-27	pGDP1	128	≥16	NR
ertapenem	None	None	0.25	≥2	-
	CTX-M-27	pGDP1	128	≥2	-
ceftriaxone	None	None	0.25	≥4	-
	CMY-2	pGDP1	128	≥4	-
	CTX-M-3	pGDP1	>256	≥4	-

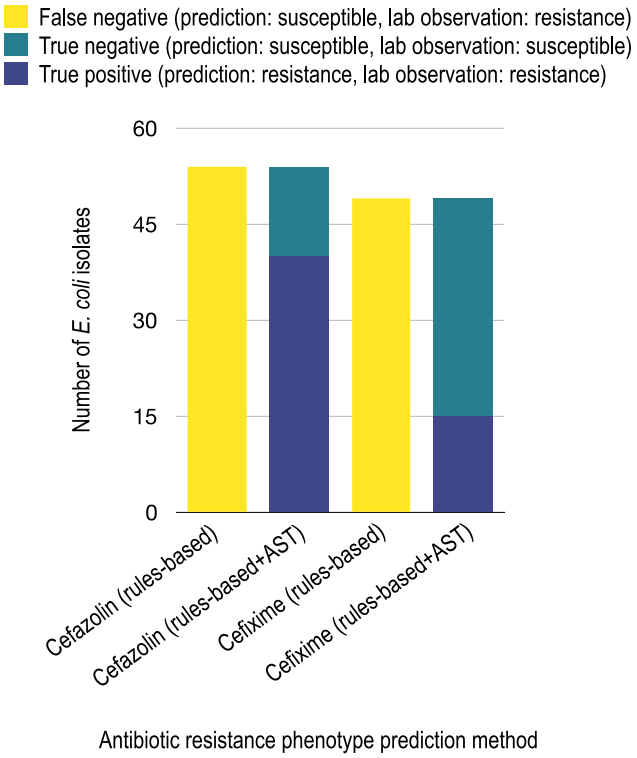


Figure 2-4. Improvement of *E. coli* cefazolin and cefixime resistance prediction using rules-based algorithm and substrate activity knowledge gained from antibiotic susceptibility testing (AST). Through antibiotic susceptibility testing, we observed CTX-M-3, CTX-M-27 and CMY-2 conferring clinically relevant resistance to cefazolin and cefixime. Curating this knowledge into CARD would improve cefazolin and cefixime true positive resistance prediction in *E. coli* by 74.1 and 30.6%, respectively.

DISCUSSION

Fast and accurate prediction of AMR phenotypes from genotypes would improve AMR surveillance, patient outcomes and antibiotic stewardship. Currently, our ability to diagnose bacterial infections is costly and slow, contributing to the misuse and overuse of antibiotics, as well as to poor clinical outcomes. Genotypic approaches using whole-genome sequencing paired with bioinformatics resources have the potential to be a faster and more accurate method. The goal of this study was to identify and elucidate β-lactamase substrate activity, a limiting factor in AMR phenotype prediction, by using two different *in silico* AMR phenotype prediction algorithms, subsequently validated using targeted gene expression experiments. In the rules-based method, we developed EPI to be used in combination with RGI to better identify overexpressed multi-component efflux pumps, while the LR method only used the resistance determinants predicted by RGI as its starting point. While naïve about the relative contribution of individual resistance determinants to overall resistance and sensitive to any gaps in knowledge for β-lactamase activity, the rules-based method nonetheless was able to accurately predict a number of resistance phenotypes when they involved well-characterized resistance determinants that confer resistance surpassing clinical breakpoints, e.g. AAC(6′)-Ib-cr for tobramycin. In terms of false-positive predictions using this approach, we hypothesize that CARD contains incorrect genotype–phenotype information, an environmental factor is altering the expression of a predicted resistance determinant, or that CARD has a knowledge gap regarding repressors. With the first scenario, removal of incorrect curation could decrease instances of false positives, highlighting one of the limitations of human biocuration for

AMR phenotype prediction. The second scenario, i.e. adaptive resistance, should not be a concern for our study, since our antibiotic susceptibility tests were standardized and automated, notwithstanding potential inconsistencies affecting gene expression (Fernández & Hancock, 2012). The third scenario suggests that there are gaps in the literature, as CARD only includes information published in peer-reviewed literature with clear experimental evidence of elevated resistance. Genetic determinants that decrease the expression or change the substrate profile of a resistance determinant, such as mutations within regulatory regions or active sites, would result in false-positive predictions. Alternatively, entirely unknown resistance genes or mutations could explain false-negative predictions of AMR phenotypes.

To identify relationships between known resistance genes and resistance phenotypes without relying on CARD's ARO for curated genotype–phenotype relationships, we used RGI in combination with LR. It is important to note that accurate and generalizable LR-based prediction of susceptibility or resistance to an antibiotic from detected AMR determinants is only feasible when there are relatively large numbers of genomes exemplifying each phenotype, which was not always the case in our data. Even with stratified sampling and methods, such as SMOTE (Chawla *et al.*, 2002), to resample datasets and improve balance (e.g. the relative proportion of susceptible and resistant isolates) there are limitations to what can be achieved with small datasets that are predominantly resistant or susceptible to a given antibiotic. Models that are not properly tested are likely to overfit to the data and are unlikely to generalize well for new data, in our case samples from outside the Hamilton, Ontario area. Additional validation of our

models using publicly available data is important for future studies; models may be dependent on feature selection, taxonomic distribution, resistance mechanism and algorithm choice. Yet, despite the models not being appropriately tested properly due to imbalance, LR proved a useful tool for improving prediction of resistance from genomic features, even without the rules-based algorithm's additional consideration of overexpressed, multi-component efflux pumps. LR substantially decreased instances of false positives or false negatives, and the poor performance for predicting particular resistance phenotypes (e.g. tetracycline resistance in *E. coli*, ceftazidime resistance in *P. aeruginosa* and piperacillin/tazobactam resistance in both species) could either represent a failure of the LR algorithm to capture the combination of resistance determinants required to predict resistance due to additive or synergistic resistance or to recognize undiscovered resistance determinants not in CARD and thus not predicted by RGI.

While bioinformatics tools such as breseq (Deatherage & Barrick, 2014) or k-mer approaches combined with LR could be used to potentially identify unknown mutations or functional gene loss (e.g. OprD loss is associated with imipenem, meropenem and doripenem resistance (Ocampo-Sosa *et al.*, 2012)), our prediction of CLSI (CLSI, 2018) 'resistant' and 'susceptible' resistance phenotypes places limits upon interpretation, as other clinical breakpoint guidelines exist, e.g. the European Committee on Antimicrobial Susceptibility Testing (EUCAST) (EUCAST, 2015) breakpoint guidelines are based on interpretation of quantitative MIC values, which unfortunately are not recorded in CARD or any other database for the breadth of known resistance genes and mutations. As such, detection of a CARD resistance determinant in a clinical isolate was interpreted as

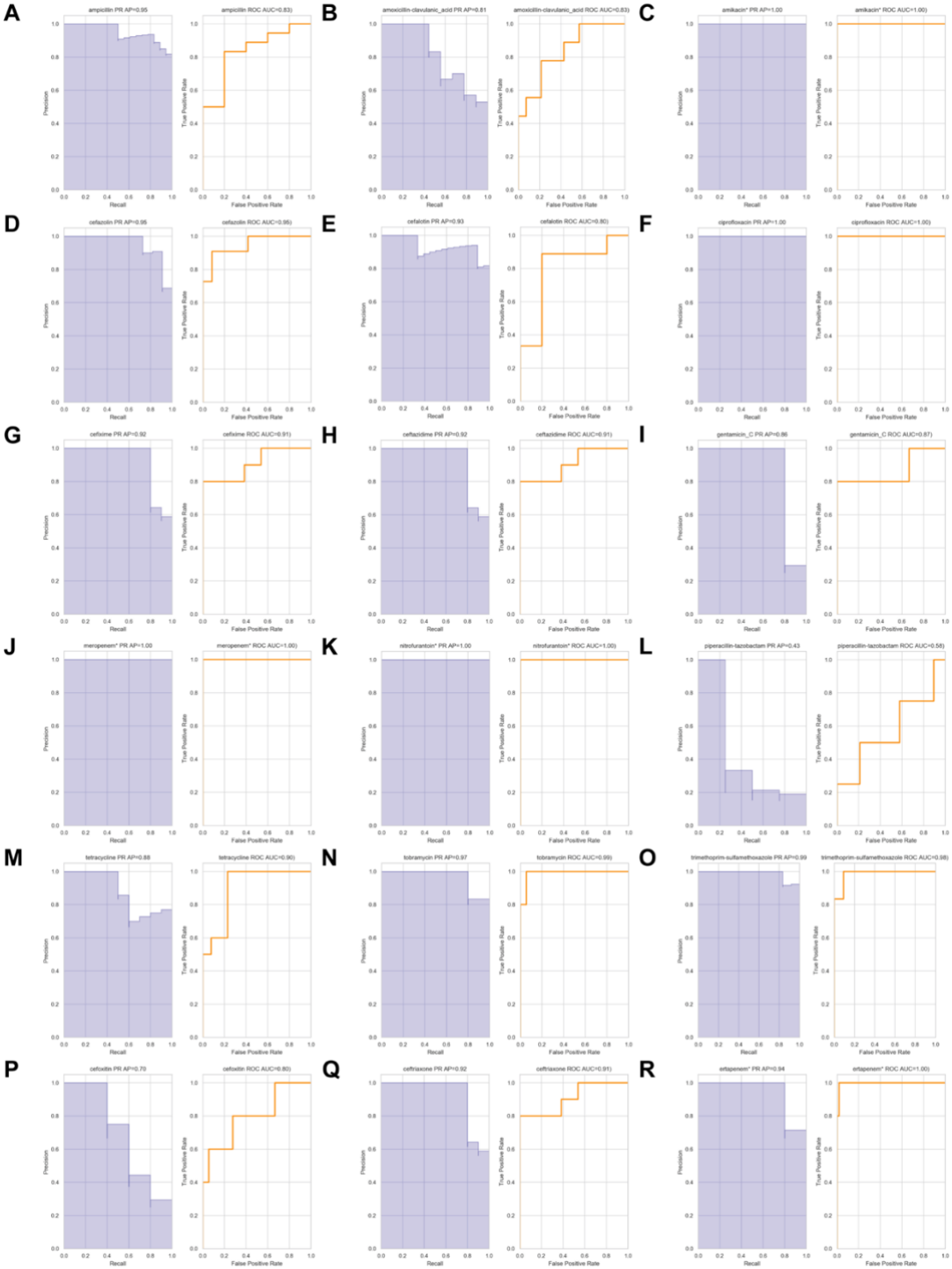
'resistant', even though in reality the MIC value generated by the gene may not have reached the CLSI or EUCAST breakpoints for resistant. Nonetheless, aligning with George E. P. Box's aphorism, 'all models are wrong, but some are useful' (Box, 1976), our goal was to identify the LR models with 'useful' or logical biological relevance with a focus on prevalent clinical β-lactamases. Prediction of genomic determinants responsible for resistance based on the feature weights of the LR only made biological sense in some cases based on the literature and knowledge. For example, *novA* was the highest weighted predictor for *P. aeruginosa* trimethoprim/sulfamethoxazole resistance, but is known to instead be involved in the transport of and resistance to novobiocin (Schmutz, Mühlenweg, Li, & Heide, 2003). Failure to predict logical determinants could be attributed to high levels of divergence from the canonical sequence or an unknown resistance determinant with prevalence correlated with NovA. In the balanced datasets, known relationships in CARD, such as Tet(C) conferring resistance to tetracycline in *E. coli* and *P. aeruginosa* GyrA mutation conferring resistance to ciprofloxacin, were predicted by both the rules-based and LR methods (Figure 2-3f, n). Beyond this, LR was additionally able to predict genotype–phenotype relationships that were useful in that they were new findings not predicted by the rules-based method and not published in the literature, yet consistent with known resistance mechanisms. Indeed, there is value in looking beyond the most highly weighted LR predictor, since analysis of a model can garner major insights into AMR genotype–phenotype relationships. We were able to experimentally validate many of the top five most highly weighted candidates, illustrating that systematic screening of a broad selection of antibiotics against known resistance

genes using molecular AST platforms such as the ARP (G. Cox *et al.*, 2017), perhaps guided by LR, or at minimum community adoption of standard panels of antibiotics for AST characterization of newly reported resistance genes, could be adopted to fill these gaps in the literature and improve antibiotic resistance phenotype prediction.

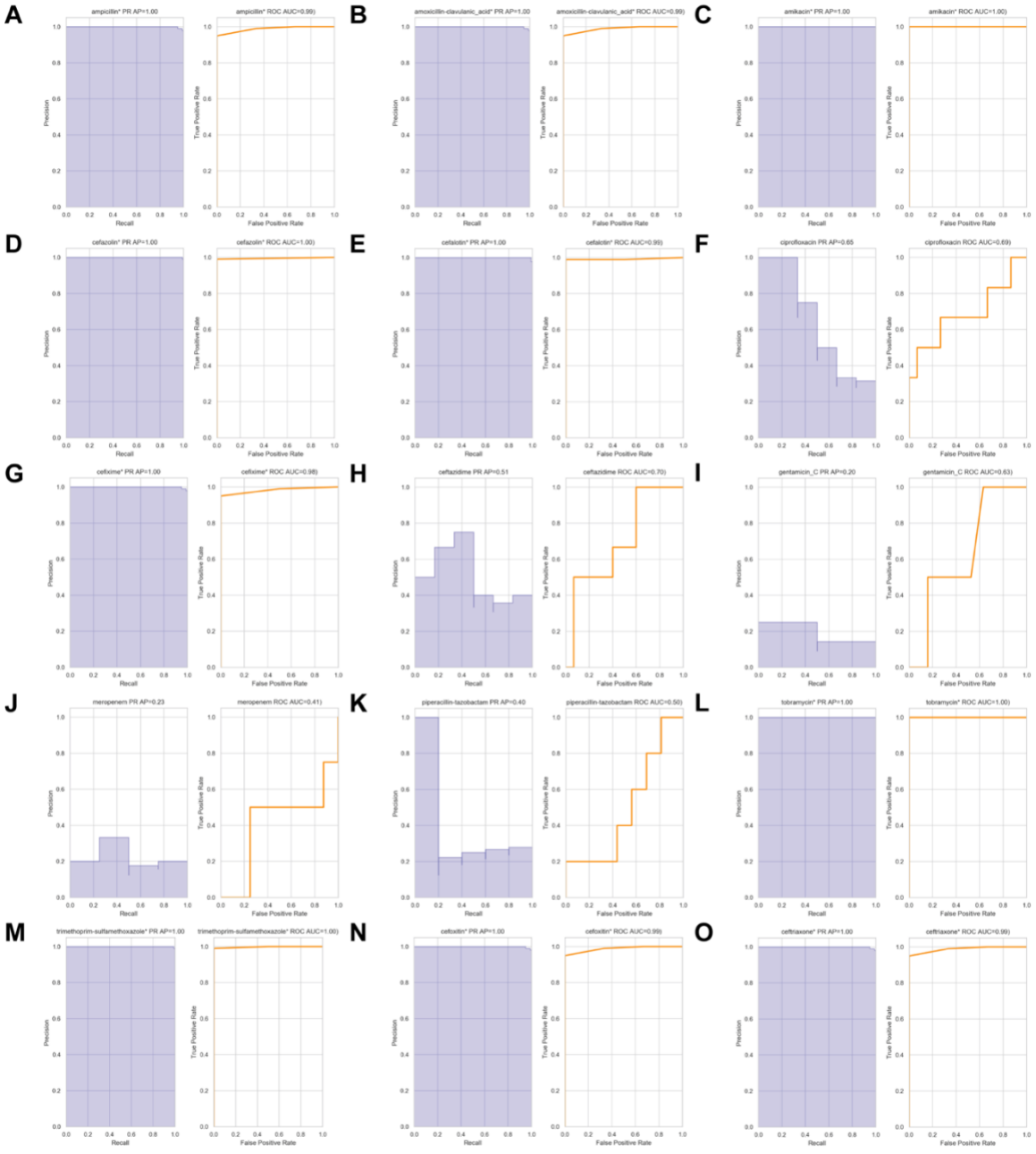
We have illustrated that completely accurate AMR phenotype prediction is not achievable using either rules-based or LR methods. There are likely unknown genomic determinants leading to both false-positive and false-negative prediction of resistance phenotypes, such as mutations in regulatory regions that change expression of a resistance gene. Overall, our results suggest that LR is capable of predicting resistance phenotypes and identifying substrate specificities of known resistance genes when there are sufficiently balanced datasets. Evaluating learned weights for each LR model led to novel hypotheses, illustrating the use of LR as an inductive approach to guide deductive research. Yet, our results also illustrate that full prediction of resistome and resistance phenotype will require careful examination of genome feature space and clinical breakpoints, plus broad and balanced sampling of diverse susceptible and resistant strains. It is our hope that collective advances in these methods will result in tools for clinical prediction of resistance, aiding antimicrobial stewardship and improving patient outcomes. Elucidating AMR genotype–phenotype relationships will reveal the genetic and mechanistic underpinnings of resistance to guide both public health surveillance and future drug discovery.

SUPPLEMENTARY MATERIAL

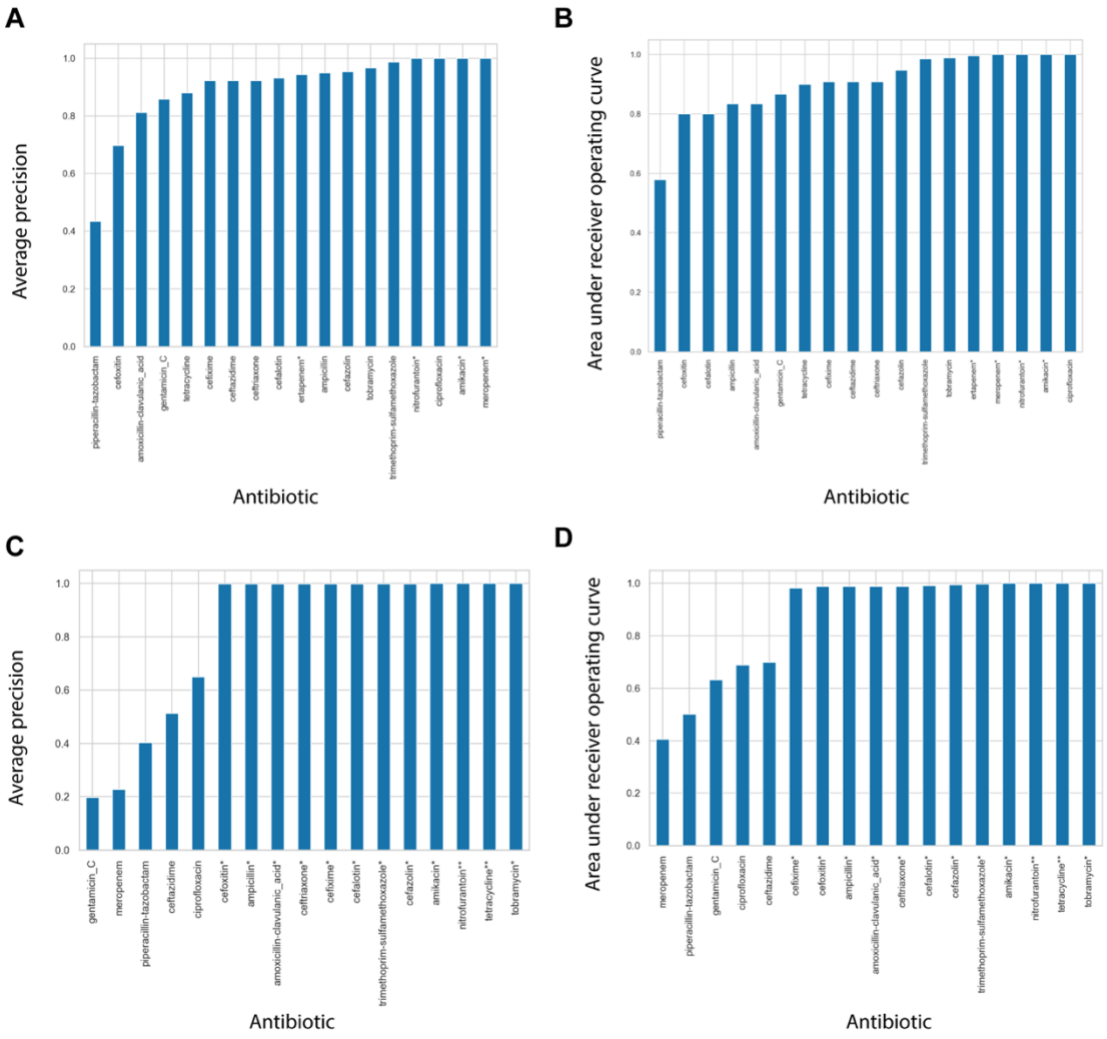
Supplementary Figures



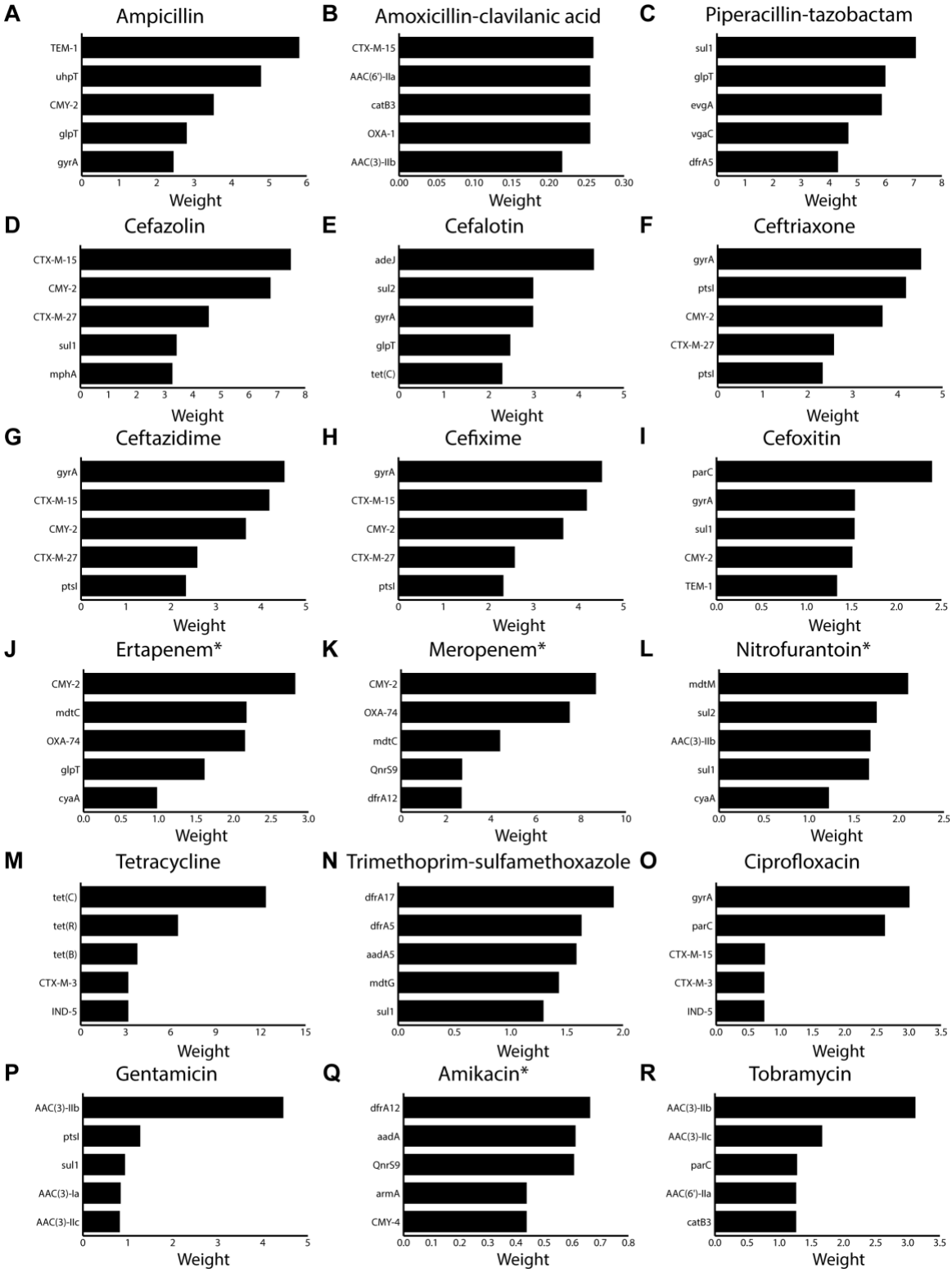
Supplementary Figure 2-1. Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves for logistic regression models developed for *E. coli* antibiotic resistance phenotype prediction. Models for (A) ampicillin, (B) amoxicillin-clavulanic acid, (C) amikacin, (D) cefazolin, (E) cefalotin, (F) ciprofloxacin, (G) cefixime, (H) ceftazidime, (I) gentamicin, (J) meropenem, (K) nitrofurantoin, (L) piperacillin-tazobactam, (M) tetracycline, (N) tobramycin, (O) trimethoprim-sulfamethoxazole, (P) cefoxitin, (Q) ceftriaxone, (R) ertapenem which <10% of a species' isolates displayed susceptible or resistant phenotypes could not be properly validated and tested (4 antibiotics for *E. coli*), so were trained using all the data (indicated by an asterisk).


A
ampicillin* PR AP=1.00
ampicillin* ROC AUC=0.99)
B
amoxicillin-clavulanic_acid* PR AP=1.00
amoxicillin-clavulanic_acid* ROC AUC=0.99)
C
amikacin* PR AP=1.00
amikacin* ROC AUC=1.00)
D
cefazolin* PR AP=1.00
cefazolin* ROC AUC=1.00)
E
cefalotin* PR AP=1.00
cefalotin* ROC AUC=0.99)
F
ciprofloxacin PR AP=0.65
ciprofloxacin ROC AUC=0.69)
G
cefixime* PR AP=1.00
cefixime* ROC AUC=0.98)
H
ceftazidime PR AP=0.51
ceftazidime ROC AUC=0.70)
I
gentamicin_C PR AP=0.20
gentamicin_C ROC AUC=0.63)
J
meropenem PR AP=0.23
meropenem ROC AUC=0.41)
K
piperacillin-tazobactam PR AP=0.40
piperacillin-tazobactam ROC AUC=0.50)
L
tobramycin* PR AP=1.00
tobramycin* ROC AUC=1.00)
M
trimethoprim-sulfamethoxazole* PR AP=1.00
trimethoprim-sulfamethoxazole* ROC AUC=1.00)
N
cefoxitin* PR AP=1.00
cefoxitin* ROC AUC=0.99)
O
ceftriaxone* PR AP=1.00
ceftriaxone* ROC AUC=0.99)
Precision
Recall
True Positive Rate
False Positive Rate

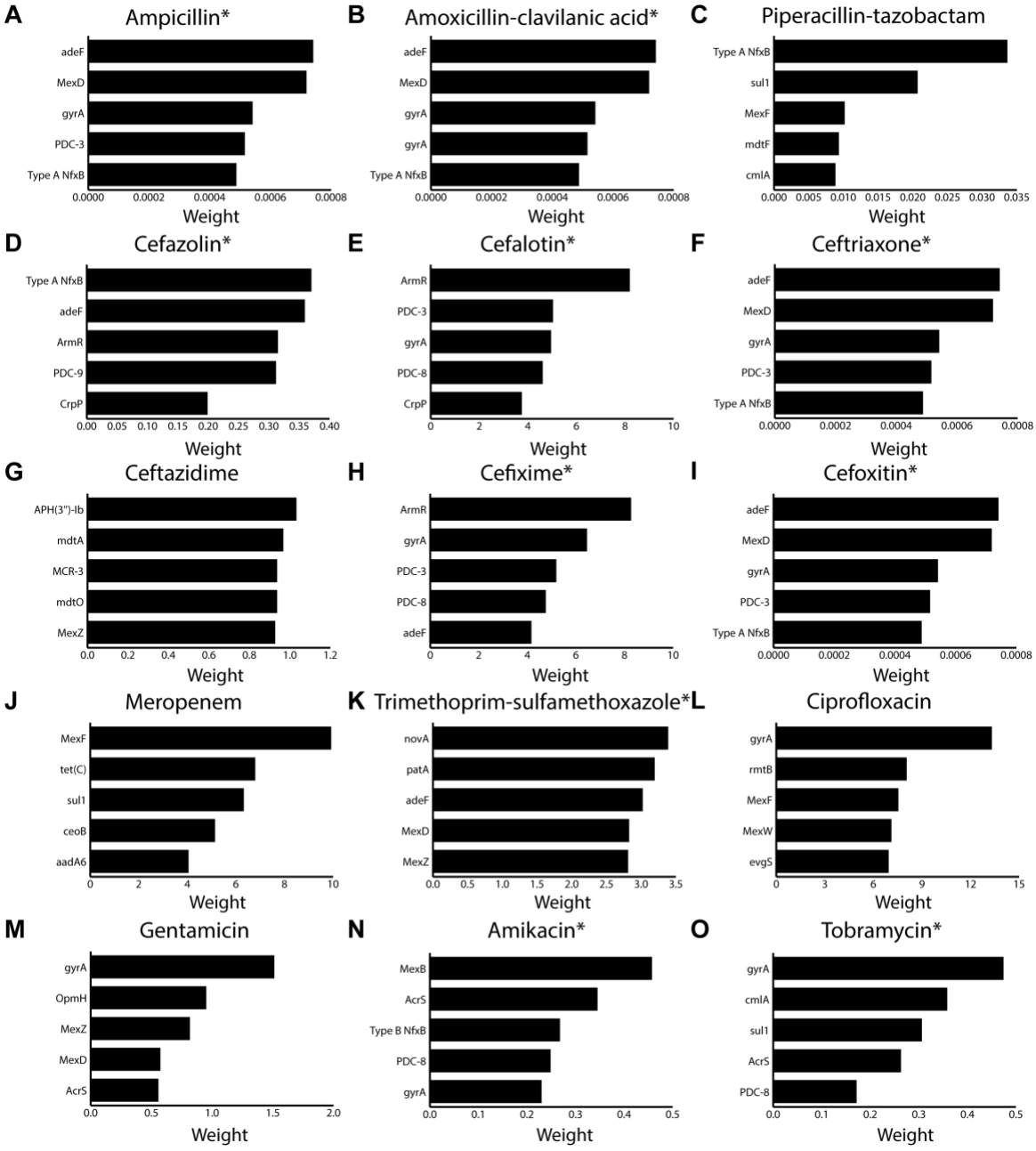
Supplementary Figure 2-2. Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves for logistic regression models developed for *P. aeruginosa* antibiotic resistance phenotype prediction. Models for (A) ampicillin, (B) amoxicillin-clavulanic acid, (C) amikacin, (D) cefazolin, (E) cefalotin, (F) ciprofloxacin, (G) cefixime, (H) ceftazidime, (I) gentamicin, (J) meropenem, (K) piperacillin-tazobactam, (L) tobramycin, (M) trimethoprim-sulfamethoxazole, (N) cefalotin, (O) ceftriaxone resistance which <10% of a species' isolates displayed susceptible or resistant phenotypes could not be properly validated and tested (10 antibiotics for *P. aeruginosa*), so were trained using all the data (indicated by an asterisk). Tetracycline, nitrofurantoin, and ertapenem resistance prediction models could not be developed for the following reasons. All isolates were resistant to tetracycline and nitrofurantoin, thus a 'dummy' model was used which always returns the relevant label. Ertapenem phenotypic AST was not performed for *P. aeruginosa.*



Supplementary Figure 2-3. Average precision and area under Receiver Operating Characteristic (ROC) graphs for (A, B) *E. coli* and (C, D) *P. aeruginosa* logistic regression models used for resistance phenotype prediction. X-axis indicates the antibiotic tested whereas the y-axis indicates the (A, C) average precision or the (B, D) area under the ROC curve for each logistic regression model. Models for antibiotics for which <10% of a species' isolates displayed susceptible or resistant phenotypes could not be properly validated and tested (10 antibiotics for *P. aeruginosa* and 4 antibiotics for *E. coli*), so were trained using all the data (indicated by an asterisk).

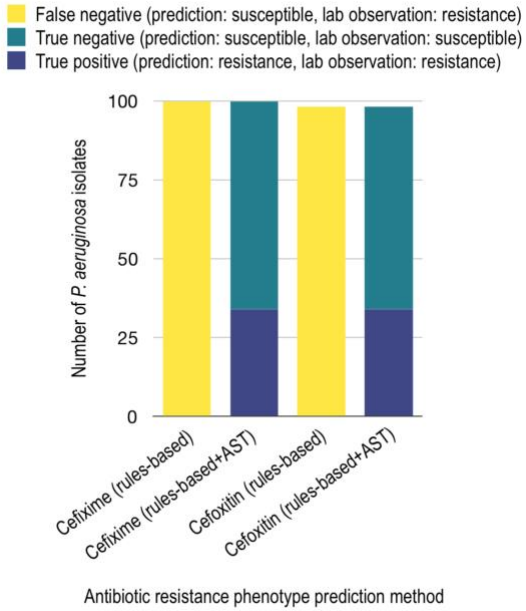

A
Ampicillin
TEM-1
uhpT
CMY-2
glpT
gyrA
Weight
B
Amoxicillin-clavilanic acid
CTX-M-15
AAC(6')-Ila
catB3
OXA-1
AAC(3)-IIb
Weight
C
Piperacillin-tazobactam
sul1
glpT
evgA
vgaC
dfrA5
Weight
D
Cefazolin
CTX-M-15
CMY-2
CTX-M-27
sul1
mphA
Weight
E
Cefalotin
adeJ
sul2
gyrA
glpT
tet(C)
Weight
F
Ceftriaxone
gyrA
ptsI
CMY-2
CTX-M-27
ptsI
Weight
G
Ceftazidime
gyrA
CTX-M-15
CMY-2
CTX-M-27
ptsI
Weight
H
Cefixime
gyrA
CTX-M-15
CMY-2
CTX-M-27
ptsI
Weight
I
Cefoxitin
parC
gyrA
sul1
CMY-2
TEM-1
Weight
J
Ertapenem*
CMY-2
mdtC
OXA-74
glpT
cyaA
Weight
K
Meropenem*
CMY-2
OXA-74
mdtC
QnrS9
dfrA12
Weight
L
Nitrofurantoin*
mdtM
sul2
AAC(3)-IIb
sul1
cyaA
Weight
M
Tetracycline
tet(C)
tet(R)
tet(B)
CTX-M-3
IND-5
Weight
N
Trimethoprim-sulfamethoxazole
dfrA17
dfrA5
aadA5
mdtG
sul1
Weight
O
Ciprofloxacin
gyrA
parC
CTX-M-15
CTX-M-3
IND-5
Weight
P
Gentamicin
AAC(3)-IIb
ptsI
sul1
AAC(3)-Ia
AAC(3)-IIc
Weight
Q
Amikacin*
dfrA12
aadA
QnrS9
armA
CMY-4
Weight
R
Tobramycin
AAC(3)-IIb
AAC(3)-IIc
parC
AAC(6')-Ila
catB3
Weight

Supplementary Figure 2-4. The top five highest weights of importance for *E. coli* antibiotic resistance phenotype prediction. The x-axis indicates assigned LR weights for individual antibiotics, while the y-axis list the top five weighted AMR determinants. Models for (A) ampicillin, (B) amoxicillin-clavulanic acid, (C) piperacillin-tazobactam, (D) cefazolin, (E) cefalotin, (F) ceftriaxone, (G) ceftazidime, (H) cefixime, (I) cefoxitin, (J) ertapenem, (K) meropenem, (L) nitrofurantoin, (M) tetracycline, (N) trimethoprim-sulfamethoxazole, (O) ciprofloxacin, (L) gentamicin (Q) amikacin, (R) tobramycin resistance which <10% of a species' isolates displayed susceptible or resistant phenotypes could not be properly validated and tested (4 antibiotics for *E. coli*), so were trained using all the data (indicated by an asterisk).



Supplementary Figure 2-5. The top five highest weights of importance for *P. aeruginosa* antibiotic resistance phenotype prediction. The x-axis indicates assigned LR weights for individual antibiotics, while the y-axis list the top five weighted AMR determinants. Models for (A) ampicillin, (B) amoxicillin-clavulanic acid, (C) piperacillin-tazobactam, (D) cefazolin, (E) cefalotin, (F) ceftriaxone, (G) ceftazidime, (H) cefixime, (I) cefoxitin, (J) meropenem, (K) trimethoprim-sulfamethoxazole, (L) ciprofloxacin, (M)

gentamicin, (N) amikacin, (O) tobramycin resistance which <10% of a species' isolates displayed susceptible or resistant phenotypes could not be properly validated and tested (10 antibiotics for *P. aeruginosa*), so were trained using all the data (indicated by an asterisk). Tetracycline, nitrofurantoin, and ertapenem resistance prediction models could not be developed for the following reasons. All isolates were resistant to tetracycline and nitrofurantoin, thus a 'dummy' model was used which always returns the relevant label. Ertapenem phenotypic AST was not performed for *P. aeruginosa.*



Supplementary Figure 2-6. Improvement of *P. aeruginosa* cefixime and cefoxitin resistance prediction using information gained from ASTs, RGI and ARO. Through antibiotic susceptibility testing (AST), we observed PDC-3 and PDC-5 conferring resistance to cefixime and cefoxitin. Curating this knowledge into CARD would improve cefixime and cefoxitin resistance true positive prediction in *P. aeruginosa* by 34.0% and 34.7%, respectively. However, there are no CLSI breakpoint guidelines for cefixime and cefoxitin because they are not used clinically to treat *P. aeruginosa* infections.

Supplementary Tables

Supplementary Table 2-1. Antibiotic susceptibility tests (AST) performed on known resistance genes. As per the Antibiotic Resistance Platform, AMR genes were cloned into the pGDP plasmid series and transformed into two strains of *E. coli*: wild-type *E. coli* BW25113, which is representative of a clinical isolate, and a hyperpermeable, efflux-deficient mutant strain, *E. coli* BW25113ΔbamBΔtolC. AST was performed for each construct using the microdilution broth method, with the inoculum prepared using the growth method following CLSI guidelines. Dashes indicate that there are no CLSI breakpoint guidelines for using that particular antibiotic against Enterobacteriaceae or *P. aeruginosa*. In red are the '*confers_resistance_to_antibiotic*' relationships considered to be clinically relevant and curated in CARD. NR: not relevant as some resistance genes are known to be intrinsic to *P. aeruginosa*.

Antibiotic	Resistance gene	Plasmid	MIC (µg/mL) wild-type *E. coli* BW25113	CLSI Resistant MIC (µg/mL) breakpoint for Enterobacteriaceae	CLSI Resistant MIC (µg/mL) breakpoint for *Pseudomonas aeruginosa*
ampicillin	None	None	64	≥32	-
	CMY-2	pGDP1	>256	≥32	-
	CTX-M-3	pGDP1	>256	≥32	-
	CTX-M-15	pGDP1	>256	≥32	-
	CTX-M-27	pGDP1	>256	≥32	-
	OXA-1	pGDP1	>256	≥32	-
	PDC-3	pGDP1	>256	≥32	-
	PDC-5	pGDP1	>256	≥32	-
	TEM-1	pGDP1	>256	≥32	-
	TEM-30	pGDP1	>256	≥32	-
amoxicillin-clavulanic acid	None	None	8-16	≥32/16	-
	CMY-2	pGDP1	256	≥32/16	-
	CTX-M-3	pGDP1	64	≥32/16	-
	CTX-M-15	pGDP1	16	≥32/16	-
	CTX-M-27	pGDP1	8-16	≥32/16	-
	OXA-1	pGDP1	64	≥32/16	-
	PDC-3	pGDP1	256	≥32/16	-
	PDC-5	pGDP1	256	≥32/16	-
	TEM-1	pGDP1	128	≥32/16	-

	TEM-30	pGDP1	128	≥32/16	-
piperacillin-tazobactam	None	None	4	≥128/4	≥128/4
	CMY-2	pGDP1	32-64	≥128/4	NR
	CTX-M-3	pGDP1	32	≥128/4	NR
	CTX-M-15	pGDP1	4	≥128/4	NR
	CTX-M-27	pGDP1	>256	≥128/4	NR
	OXA-1	pGDP1	32	≥128/4	NR
	PDC-3	pGDP1	>256	NR	≥128/4
	PDC-5	pGDP1	32	NR	≥128/4
	TEM-1	pGDP1	64-128	≥128/4	NR
	TEM-30	pGDP1	16	≥128/4	NR
cefazolin	None	None	4	≥8/≥32 (Urine only)	-
	CMY-2	pGDP1	>256	≥8/≥32 (Urine only)	-
	CTX-M-3	pGDP1	>256	≥8/≥32 (Urine only)	-
	CTX-M-15	pGDP1	128	≥8/≥32 (Urine only)	-
	CTX-M-27	pGDP1	>256	≥8/≥32 (Urine only)	-
	OXA-1	pGDP1	8	≥8/≥32 (Urine only)	-
	PDC-3	pGDP1	>256	NR	-
	PDC-5	pGDP1	>256	NR	-
	TEM-1	pGDP1	256	≥8/≥32 (Urine only)	-
	TEM-30	pGDP1	2	≥8/≥32 (Urine only)	-
cefalotin	None	None	16-32	-	-
	CMY-2	pGDP1	>256	-	-
	CTX-M-3	pGDP1	>256	-	-
	CTX-M-15	pGDP1	128	-	-
	CTX-M-27	pGDP1	16	-	-
	OXA-1	pGDP1	16-32	-	-
	PDC-3	pGDP1	>256	NR	-
	PDC-5	pGDP1	>256	NR	-
	TEM-1	pGDP1	256	-	-
	TEM-30	pGDP1	8	-	-
cefixime	None	None	0.25	≥4	-
	CMY-2	pGDP1	>256	≥4	-
	CTX-M-3	pGDP1	32	≥4	-
	CTX-M-15	pGDP1	0.5	≥4	-
	CTX-M-27	pGDP1	4	≥4	-

	OXA-1	pGDP1	0.25-0.5	≥4	-
	PDC-3	pGDP1	>256	NR	-
	PDC-5	pGDP1	>256	NR	-
	TEM-1	pGDP1	0.25-0.5	≥4	-
	TEM-30	pGDP1	0.5	≥4	-
ceftazidime	None	None	0.5	≥16	≥32
	CMY-2	pGDP1	256	≥16	NR
	CTX-M-3	pGDP1	16-32	≥16	NR
	CTX-M-15	pGDP1	0.5	≥16	NR
	CTX-M-27	pGDP1	128	≥16	NR
	OXA-1	pGDP1	0.5	≥16	NR
	PDC-3	pGDP1	32	NR	≥32
	PDC-5	pGDP1	32	NR	≥32
	TEM-1	pGDP1	1	≥16	NR
	TEM-30	pGDP1	0.25	≥16	NR
ertapenem	None	None	0.25	≥2	-
	CMY-2	pGDP1	1-2	≥2	-
	CTX-M-3	pGDP1	0.5	≥2	-
	CTX-M-15	pGDP1	0.25	≥2	-
	CTX-M-27	pGDP1	128	≥2	-
	OXA-1	pGDP1	0.25	≥2	-
	PDC-3	pGDP1	0.25	NR	-
	PDC-5	pGDP1	0.25	NR	-
	TEM-1	pGDP1	0.25-0.5	≥2	-
	TEM-30	pGDP1	0.25	≥2	-
meropenem	None	None	0.25	≥4	≥8
	CMY-2	pGDP1	0.25	≥4	NR
	CTX-M-3	pGDP1	0.25	≥4	NR
	CTX-M-15	pGDP1	0.25	≥4	NR
	CTX-M-27	pGDP1	0.5	≥4	NR
	OXA-1	pGDP1	0.25	≥4	NR
	PDC-3	pGDP1	0.25	NR	≥8
	PDC-5	pGDP1	0.25-64	NR	≥8
	TEM-1	pGDP1	0.25-0.5	≥4	NR
	TEM-30	pGDP1	0.25	≥4	NR
cefoxitin	None	None	8	≥32	-

	CMY-2	pGDP1	8	≥32	-
	CTX-M-3	pGDP1	32-256	≥32	-
	CTX-M-15	pGDP1	16	≥32	-
	CTX-M-27	pGDP1	0.5	≥32	-
	OXA-1	pGDP1	8	≥32	-
	PDC-3	pGDP1	>256	NR	-
	PDC-5	pGDP1	64	NR	-
	TEM-1	pGDP1	16	≥32	-
	TEM-30	pGDP1	1	≥32	-
ceftriaxone	None	None	0.25	≥4	-
	CMY-2	pGDP1	128	≥4	-
	CTX-M-3	pGDP1	>256	≥4	-
	CTX-M-15	pGDP1	128	≥4	-
	CTX-M-27	pGDP1	0.25	≥4	-
	OXA-1	pGDP1	0.25	≥4	-
	PDC-3	pGDP1	32	NR	-
	PDC-5	pGDP1	32	NR	-
	TEM-1	pGDP1	0.25	≥4	-
	TEM-30	pGDP1	0.25	≥4	-
gentamicin	None	None	0.5	>16	NR
	aac(6')-Ib-cr	pGDP3	0.5	>16	NR
amikacin	None	None	1-2	>64	NR
	aac(6')-Ib-cr	pGDP3	4-8	>64	NR
tobramycin	None	None	0.5-1	>16	NR
	aac(6')-Ib-cr	pGDP3	32-64	>16	NR

CHAPTER THREE: Antimicrobial resistance prediction model performance is dependent on dataset, algorithm, and evaluation metric

CHAPTER THREE PREFACE

Author contributions: KKT and AGM conceived the project and designed experiments. KKT performed the analysis. KKT and AGM wrote the manuscript. KKT wrote this chapter.

Acknowledgements

This research was funded by the Canadian Institutes of Health Research (PJT-156214 to A.G.M.), a David Braley Chair in Computational Biology to A.G.M., and an Ontario Graduate Scholarship, a McMaster University MacDATA Institute Graduate Fellowship, and a Michael G. DeGroote Institute for Infectious Disease Research Michael Kamin Hart Memorial Scholarship to K.K.T. Computer resources were supplied by the McMaster Service Lab and Repository computing cluster, funded in part by grants from the Canadian Foundation for Innovation (34531 to A.G.M.) and donation of hardware from Cisco Systems Canada, Inc. We are grateful to Dr. Gerard Wright (McMaster University, Canada) for sharing the EC1 and PA1 datasets, whose acquisition was funded by the Ontario Research Fund.

ABSTRACT

Antimicrobial resistance (AMR) is a global health problem that is exacerbated by antibiotic misuse and overuse. Our inability to determine the antibiotic susceptibility of infectious pathogens within an optimal treatment timeframe is driven by the difficulties of culturing. Using whole genome sequencing, researchers have developed accurate AMR prediction models for a number of pathogens. These previously published studies are difficult to compare since they use different datasets, algorithms, genetic features, and evaluation metrics. In addition, only a few studies have attempted to elucidate the effect of these parameters on AMR prediction model performance. In this study, we show a three-way dependency upon dataset, algorithm, and evaluation metric on AMR prediction model performance in *Escherichia coli, Neisseria gonorrhoeae,* and *Pseudomonas aeruginosa.* We used known resistance determinants and mutations as genetic features combined with feature filtering methods which can have a variety of effects on AMR prediction model performance. Using only plasmid borne resistance determinants generally produced poorer performing AMR prediction models compared to using resistance determinants encoded by both the chromosome and plasmid. We showed that representing and filtering genetic features can improve AMR prediction models that can address sequencing error, gene copy number, and physicochemical properties of amino acid substitutions. We observed how choosing an evaluation metric dictates which algorithm is selected and that generally naïve Bayes performs poorly for AMR phenotype prediction using most evaluation metrics. Lastly, we demonstrate how AMR prediction model performance is also specific to the pathogen and antibiotic of interest. Therefore,

since we show how AMR prediction model performance relies on a number of decisions that a researcher can make during model construction, e.g., evaluation metric selection, algorithm, feature selection, and dataset stratification, we suggest testing and incorporating non-genetic metadata to determine how to create the best prediction models for a given dataset. Furthermore, we highlight the importance of sampling broadly and deeply, while collecting metadata, such as site of infection, to include into AMR prediction models. Lastly, we argue that domain experts should consider shifting from developing broad generalizable models to narrow and more specific AMR prediction models. Altogether, it is of added value to understand the effects of these parameters on AMR prediction model performance to further build a foundation for machine learning diagnostics in clinical microbiology labs.

INTRODUCTION

In 2015, the World Health Organization published a global action plan on antimicrobial resistance (AMR) which highlights the need to improve awareness, strengthen knowledge, reduce infections, prioritize stewardship, and develop economic investment (World Health Organization, 2015). AMR has been estimated to cause 126,000-700,000 deaths per year globally (Naghavi *et al.*, 2017; O'Neill, 2016), however it is important to understand the limitations and assumptions of these estimates (de Kraker, Stewardson, & Harbarth, 2016; Limmathurotsakul *et al.*, 2019). In 2017, the World Health Organization published a critical priority pathogens list including both Gram-negative and Gram-positive bacteria, including carbapenem resistant *Pseudomonas aeruginosa* and *Enterobacteriaceae* in the critical priority list, as well as, cephalosporin and fluoroquinone resistant *Neisseria gonorrhoeae* in the second high priority list (World Health Organization, 2017b). In 2019, the Centres for Disease and Control estimated 3,200 cases of multi-drug resistant *Pseudomonas aeruginosa*, 550,000 drug-resistant *Neisseria gonorrhoeae* cases, and 210,500 extended spectrum β-lactamase producing or carbapenem resistant *Enterobacteriaceae (CDC, 2019)*. As rates of drug resistant infections increase, we are simultaneously faced with a shrinking antimicrobial drug discovery pipeline (Butler & Paterson, 2020; Hutchings, Truman, & Wilkinson, 2019; World Health Organization, 2021). As such, prevention, timely diagnosis, and appropriate treatment of drug resistant infections guides the principles of antibiotic stewardship programs.

The current gold standard of diagnosing AMR currently relies on culture-based, phenotypic methods, namely antibiotic susceptibility tests (e.g., disk diffusion and broth dilution). However, the turnaround time for these results can be longer than the optimal infection treatment window, particularly for fastidious organisms. In contrast, genotypic methods of diagnosing AMR include PCR based nucleic acid amplification tests and whole genome sequencing (WGS). Currently, the costs and time to results are some limiting factors for point of care application of WGS. Resolving the challenges of using WGS for AMR diagnostics is imperative for when sequencing technology and costs are realistically implemented in a clinical environment. Feasibility issues are confounded by the gap between genotype and phenotype prediction, as AMR phenotypes are produced by the interplay of multiple factors, including genetic elements and the environment of the infecting bacteria (Geisinger & Isberg, 2017). Understanding the genetic drivers of AMR phenotypes facilitates the development of novel and improved molecular diagnostics for AMR (World Health Organization, 2019). However, associations between genotype and phenotype are difficult to uncover because of the polygenic nature of AMR, which perpetrates the difficulty in developing and updating AMR databases and software.

There are many bioinformatics tools and databases available to predict AMR determinants from bacterial genomes. For example, the Comprehensive Antibiotic Resistance Database (CARD) (Alcock *et al.*, 2020) is an ontology-driven genomics database used by the Resistance Gene Identifier (RGI) software to predict intrinsic and acquired resistance determinants in genome sequences. The Antibiotic Resistance Gene-ANNOTation database (Gupta *et al.*, 2014) and Pathosystems Resource Integration

Center (Davis *et al.*, 2020) store a similar breadth of resistance determinants to CARD. ResFinder (Bortolaia *et al.*, 2020) focuses on acquired resistance genes, while ResFams (Gibson *et al.*, 2015) is a database of protein domain hidden Markov models for AMR function prediction. However, with the abundance of bioinformatics pipelines available comes discordant AMR genotype predictions across research groups, hospital laboratories, public health laboratories, and clinical diagnostic companies (Doyle *et al.*, 2020). If these AMR predictions were used to inform antibiotic treatment, at least one would recommend a different antibiotic and few would be based on a clear association between genotype and phenotype. Presence of an AMR gene or mutation does not guarantee expression and a subsequent drug resistant infection (Tsang *et al.*, 2021).

With advances in genome sequencing technology, researchers are increasingly examining the use of machine learning to predict AMR phenotype from genotype. Most previously published literature predicts ‘resistant’ or ‘susceptible’ phenotypes (Chowdhury, Call, & Broschat, 2019; Coelho *et al.*, 2013; Davis *et al.*, 2016; Drouin *et al.*, 2019; Z. Liu *et al.*, 2020; Pesesky *et al.*, 2016; Shi *et al.*, 2019; Tsang *et al.*, 2021; Yang *et al.*, 2018), while some (additionally) predict minimum inhibitory concentrations (MIC) (Demczuk *et al.*, 2016; Demczuk *et al.*, 2020; Eyre *et al.*, 2019; Eyre *et al.*, 2017; Golparian *et al.*, 2018; Hicks *et al.*, 2019; Nguyen *et al.*, 2018; Nguyen *et al.*, 2020; Nguyen *et al.*, 2019; Pataki *et al.*, 2020). The genetic features used in these prediction models include known resistance determinants, nucleotide sequence *k*-mers, and mutations to predict AMR phenotypes for *Klebsiella pneumoniae, Neisseria gonorrhoeae, Actinobacillus pleuropneumoniae, Mycobacterium tuberculosis,*

Escherichia coli, Pseudomonas aeruginosa, Salmonella enterica, and *Staphylococcus aureus*. Typically, the accuracy of AMR prediction models is greater than 90%, however some publications use different prediction model evaluation metrics, such as area under the curve (Cassini *et al.*, 2019) sensitivity (Yang *et al.*, 2018), and F1 score (Nguyen *et al.*, 2020). Previously published works also use a number of different algorithms, including logistic regression (Pesesky *et al.*, 2016; Tsang *et al.*, 2021), linear regression (Demczuk *et al.*, 2020; Eyre *et al.*, 2017), XGBoost (Nguyen *et al.*, 2018; Nguyen *et al.*, 2020; Nguyen *et al.*, 2019), set covering machine (Drouin *et al.*, 2016; Z. Liu *et al.*, 2020), support vector machine (Z. Liu *et al.*, 2020), random forest (Hicks *et al.*, 2019), and deep learning (Shi *et al.*, 2019). The inconsistency between the algorithm, evaluation metric and datasets used in previously published works makes AMR prediction models difficult to compare.

In our previous work, we illustrated the power of logistic regression for prediction of 'Resistant' or 'Susceptible' phenotype under CLSI guidelines based on genome sequences of *E. coli* and *P. aeruginosa* (Tsang *et al.*, 2021), yet there is a lack of consensus on the most appropriate evaluation metric to assess machine learning model performance for AMR. Typically, an evaluation metric (e.g., accuracy, balanced accuracy, negative log loss score, precision, recall, F1 score) is chosen to measure prediction model performance. Once an evaluation metric is chosen, different machine learning algorithms are then put to the test. The lack of standardization in evaluation metric choice in publications inherently challenges our ability to compare different algorithms and studies.

Feature selection is another criterion that is determined prior to generating prediction models, i.e. which aspects of the genome to use for construction of models. Genetic features can be derived from bacterial sequences by annotating known AMR determinants, identifying all mutations using a reference sequence, or generating *k*-mers (short nucleotide sequences) that span the entire genome sequence. For example, Hicks *et al.* have evaluated parameters affecting AMR *Neisseria gonorrhoeae*, *Klebsiella pneumoniae* and *Acinetobacter baumannii* phenotypic prediction model performance and reliability using set covering machine and random forest algorithms with *k*-mers derived from genome assemblies, illustrating accuracy varies by antibiotic, genomic diversity, and pathogen (Tsang *et al.*, 2021). Overall, too few genetic features restricts the machine learning algorithm's ability to learn, leading to an underfitted model that has more bias towards incorrect predictions. In contrast, too many genetic features can lead to overfitting, i.e. an overly specific model that is not generalizable to new data. Thus, genetic feature selection is important for generating useful AMR prediction models, particularly if interpretation of the mechanisms driving resistance is valued. Using known resistance determinants as features also allows for easier interpretation of AMR prediction models but limits the discovery of novel resistance determinants.

While in our previous work we evaluated a few different algorithms, we only used negative log loss as our evaluation metric, geographically limited datasets for *E. coli* and *P. aeruginosa,* and known resistance determinants as features (Tsang *et al.*, 2021). The purpose of this publication is to elucidate the effect of algorithm, evaluation parameter, and dataset on AMR prediction models. We use a number of genetic features (e.g.,

chromosome- and plasmid-borne resistance determinants, mutations) with different feature filtering methods to test four different algorithms (e.g., logistic regression, decision trees, naïve Bayes, and random forest). We show that while building AMR prediction models is intricate, understanding the genomic context and biology when making decisions about feature, evaluation metric, and algorithm selection are important to the quality and interpretability of AMR phenotype prediction. Finally, since we illustrate that AMR prediction models are also specific to pathogen, antibiotic, and data stratification, we argue that perhaps the AMR prediction field should deviate from developing universally generalizable models to more specific and localized AMR prediction models.

METHODS

Bacterial Isolates

For construction and testing of models, we used the *E. coli* (EC1) and *P. aeruginosa* (PA1) sample collections from our previous work (Tsang *et al.*, 2021), *E. coli* (EC2) (MacFadden *et al.*, 2019) and *P. aeruginosa* (PA2) (Davis *et al.*, 2020) collections from the PATRIC database, and two previously published *N. gonorrhoeae* collections: NG1 (Lee *et al.*, 2018) and NG2 (Eyre *et al.*, 2017). Unpublished phenotypic testing data from dataset EC2 are available on https://github.com/karatsang/DatasetAlgorithmEvaluation. Each of these genome data sets are associated with phenotypic estimates of Susceptible (S), Intermediate (I), or Resistant (R) for each antibiotic tested. Datasets EC1, PA1, and NG1 provided S and R categories for each dataset based on CLSI guidelines, whereas datasets EC2, PA2, and NG2 included minimum inhibitory concentrations which we interpreted into S and R categories also using CLSI guidelines (CLSI, 2018). Phenotypic testing of EC1, EC2, and PA1 were performed using the Vitek 2 system, while PA2 included methods including broth microdilution, Trek Sensititre custom plates, and Vitek systems. NG1 and NG2 phenotypic susceptibilities were performed using agar dilution. Any samples categorized as Intermediate (I) were re-encoded as R. Descriptions of these genomes and associated phenotypic measurements are presented in Table 3-1. For more information about the phenotypic measurements of each dataset, refer to their respective primary publication source (Davis *et al.*, 2020; Lee *et al.*, 2018; MacFadden *et al.*, 2019; Tsang *et al.*, 2021). All datasets were balanced (i.e., less frequent phenotype represented >10% of all

genomes), with the exception of: EC1 (amikacin, meropenem, nitrofurantoin, ertapenem), EC2 (cefazolin, ertapenem, nitrofurantoin), PA1 (ampicillin, amoxicillin-clavulanic acid, amikacin, cefazolin, cefixime, tobramycin, cefoxitin, ceftriaxone), NG1 (penicillin, spectinomycin), and NG2 (cefixime). All datasets had raw sequencing FASTQ data available, with the exception of dataset PA2 where only genome assemblies (FASTA) were available.

Table 3-1. Descriptions of the datasets. Not all antibiotics were tested against every clinical isolate.

Dataset	Pathogen	Number of isolates	Geography	Site(s) of infection	Antibiotics tested
EC1	*Escherichia coli*	115	Hamilton, Ontario, Canada	blood, urine, sputum, abdomen, rectal	ampicillin, amoxicillin-clavulanic acid, amikacin, cefazolin, cefalotin, ciprofloxacin, cefixime, ceftazidime, gentamicin C, meropenem, nitrofurantoin, piperacillin-tazobactam, tetracycline, tobramycin, trimethoprim-sulfamethoxazole, cefoxitin, ceftriaxone, ertapenem
EC2	*Escherichia coli*	1097	Ontario, Canada	blood, urine	ampicillin, cefazolin, cefotaxime, ciprofloxacin, ertapenem, gentamicin, meropenem, nitrofurantoin, trimethoprim-sulfamethoxazole
PA1	*Pseudomonas aeruginosa*	102	Hamilton, Ontario, Canada	blood, urine, sputum, abscess,	ampicillin, amoxicillin-

				eye, abdomen, arm, leg, chest, endotracheal tube, ulcer, catheter tip, foot	clavulanic acid, amikacin, cefazolin, cefalotin, ciprofloxacin, cefixime, ceftazidime, gentamicin C, meropenem, nitrofurantoin, piperacillin-tazobactam, tetracycline, tobramycin, trimethoprim-sulfamethoxazole, cefoxitin, ceftriaxone
PA2	*Pseudomonas aeruginosa*	533	worldwide	unknown	amikacin, cefixime, ceftazidime, ciprofloxacin, gentamicin, meropenem, piperacillin-tazobactam, tobramycin
NG1	*Neisseria gonorrhoeae*	398	New Zealand	cervix, vaginal, urethral, anorectal, penile, throat/pharyngeal	ciprofloxacin, azithromycin, penicillin, tetracycline, spectinomycin
NG2	*Neisseria gonorrhoeae*	660	Brighton (England), USA, & Canada	urethra, rectum, pharynx, cervix, eye	ciprofloxacin, azithromycin, penicillin, tetracycline, cefixime

Genetic feature generation

For each isolate, raw short read sequences are first trimmed using Trimmomatic (Bolger *et al.*, 2014) and then either used to identify mutations using breseq (v 0.35.3) (Deatherage & Barrick, 2014), assembled into chromosomal and plasmid DNA using SPAdes (Robertson & Nash, 2018), or assembled into plasmid DNA alone using HyAsP (Müller & Chauve, 2019) (v1.0.0). Mutations were identified using breseq with default parameters and the following reference sequences: *E. coli* O83:H1 str. NRG 857C

(ASM18334v1), *E. coli O157:H7 str. Sakai DNA* (NC_002695.2), *P. aeruginosa* PAO1 (NC_002516.2), *P. aeruginosa* UCBPP-PA14 (NC_008463.1), *N. gonorrhoeae* ATCC 49226, WHOF, WHOG, WHOK, WHOL, WHOM, WHON, WHOO, WHOP, WHOU, WHOV, WHOW, WHOX, WHOZ. Gdtools was used to annotate the breseq results (Deatherage & Barrick, 2014). Only genome assemblies (FASTA) were available for the PA2 dataset, thus mutation prediction was not performed. Resistance determinants were predicted in the chromosomal and plasmid DNA assemblies using the Resistance Gene Identifier (RGI, v 5.1.0) and Comprehensive Antibiotic Resistance Database (CARD, v 3.0.8) (Alcock *et al.*, 2020). RGI categorizes resistance determinants as 'Perfect' or 'Strict' if the predicted amino acid sequence is 100% identical to the reference sequence in CARD or if the predicted amino acid sequence passes a curated bitscore cutoff, respectively. Since RGI is dependent on CARD, RGI is unable to identify new resistance determinants, while breseq is CARD-independent, meaning that it is able to identify unknown mutations driving resistance.

Genetic feature filtering

We removed any mutations from breseq that were only observed in one isolate in a given dataset to reduce the potential misrepresentation of data, as it is difficult to differentiate between sequencing error, transcription error, and a *bona fide* mutation if it appears in a single isolate. In contrast, if a mutation was identified in multiple isolates, it is less likely to be a sequencing/transcription error. To remove potential spurious resistance determinant predictions by RGI, we applied a Grantham Score (Grantham,

1974) filter to categorize amino acid substitutions (relative to CARD reference) into classes of physicochemical dissimilarity: conservative (0-50), moderately conservative (51-100), moderately radical (101-150) or radical (≥151). We removed any RGI hits that had a Grantham Score greater than 151.

AMR prediction modelling

Genetic features for each dataset were collated into count matrices Xij where i represents each genome of that dataset and j represents a specific genetic feature. After matrices of genetic features were generated, we used four different machine learning algorithms (logistic regression, decision tree, random forest, and naïve Bayes) to build models for prediction of antimicrobial resistance phenotype from genotype for each antibiotic against each pathogen. Logistic regression (LR) is the simplest of all four algorithms and is an extension of linear regression (R. E. Wright, 1995). The decision tree (DT) algorithm creates a model by learning simple decision rules inferred from the features (Quinlan, 1986). Random forest (RF) is a combination of many decision trees, which makes it more difficult to interpret (Breiman, 2001). Lastly, Naïve Bayes (NB) assumes independence among features when developing a model (Lewis, 1998). For each dataset or combination of datasets, the hyperparameters were tuned using a threefold stratified shuffle split 3-fold cross-validation scheme and the training sets were evaluated using accuracy, balanced accuracy, average precision, precision, F1 score, negative log loss score, and recall for all AMR prediction models to determine which evaluation metric allows for greatest differentiation when evaluating model performance (Pedregosa

et al., 2011). Accuracy (and balanced accuracy), precision, F1, and recall can be calculated using values from a confusion matrix (e.g., true positives, true negatives, false positives, and false negatives) (Supplementary Table 3-1) (Ting, 2017). Accuracy often reported because of its simplicity, but can be strongly skewed when using imbalanced datasets (i.e., less representative phenotype is <10% prevalent in the dataset), which is what balanced accuracy can resolve. Average precision summarizes precision-recall curves and is also useful for imbalanced datasets. Lastly, log loss can also be used for imbalanced datasets and it considers prediction uncertainty in relation to the divergence of the predicted probabilities and the actual AMR phenotype. For more details regarding model development, we adhered to the methods in our previously published AMR prediction models (Tsang *et al.*, 2021).

Machine learning and dataset partitioning were performed using scikit-learn (Pedregosa *et al.*, 2011) (v0.20.0) with data otherwise manipulated using numpy (Oliphant, 2006) (v1.17.2) and pandas (McKinney, 2010) (v0.25.1). Heatmaps were generated using seaborn (v0.11.0). The code and conda environments (using python v3.7.2) and intermediate data files required to generate this analysis are available: https://github.com/karatsang/DatasetAlgorithmEvaluation

RESULTS

Genetic feature generation and filtering

Across each *Escherichia coli*, *Pseudomonas aeruginosa*, and *Neisseria gonorrhoeae* dataset (Table 3-1), we used SPAdes (Bankevich *et al.*, 2012) to assemble chromosomes and plasmids of each isolate (Table 3-2), while HyAsP (Müller & Chauve, 2019) was used to predict plasmid sequences alone. The only exception was the *Pseudomonas aeruginosa* PA2 dataset, for which raw sequencing reads were not available and we used the available assemblies, which represent both chromosome and plasmid sequences. In general, *E. coli* had 1.2-1.5 circular plasmids per sample, whereas *N. gonorrhoeae* had 1.0-1.2 circular plasmids per sample (Table 3-3). Across all 102 *P. aeruginosa* isolates, only one circular plasmid was identified.

Using the Comprehensive Antibiotic Resistance Database (CARD) and Resistance Gene Identifier (RGI) we predicted known AMR determinants in a given dataset. We first identified resistance determinants in SPAdes plasmid and chromosome assemblies, with larger datasets finding larger numbers of determinants due to the increase in genomic diversity (Supplementary Figure 3-1 to 3-6). For example, in the *E. coli* datasets, 184 and 448 resistance determinants were identified in EC1 (n=115) and EC2 (n=972), respectively. Overall, fewer resistance determinants were predicted for *N. gonorrhoeae* than for *E. coli* or *P. aeruginosa* (Supplementary Figure 3-1, 3-3, 3-6). Similar results were observed when using RGI to annotate the HyAsP assemblies (Supplementary Figure 3-7 to 3-10), including fewer for *N. gonorrhoeae* plasmids (9-11 AMR determinants) than *E. coli* plasmids (69-140 AMR determinants). In our analyses of all datasets, we

represented RGI results in two ways: only the presence of the resistance gene or the presence of the resistance gene combined with its RGI criteria (Perfect amino acid sequence match to CARD reference sequence or Strict variant of the CARD reference sequence). For example, if both a Perfect and Strict TEM-1 were identified in a sample, in the first representation there would only be one TEM-1 feature counted but in the second representation a TEM-1 Perfect feature and a TEM-1 Strict feature would be separately counted. Representing the resistance gene with its RGI criteria ('PS' representation) increased the number of features by 10-30%. We also filtered the RGI features in two ways. The first was selecting for resistance determinants that were found in at least two or more samples within the dataset to remove any uniquely identified and uninformative features. Typically, this reduced the number of features by 10-30%. Next, we created a Grantham score ('GS') filter to remove any predicted resistance genes that were considered to have radical amino acid changes, and thus likely representative of RGI false positives. The GS filter typically reduced the number of features by ~50%.

Table 3-2. Average length and N50 of SPAdes (chromosome and plasmid) assemblies. Standard deviation (SD) in brackets. As sequencing reads were unavailable for PA2, we used the provided genome assemblies.

Dataset	Number of samples	Average length in base pairs (SD)	Average N50 in base pairs (SD)
EC1	115	5,163,879 (182,559)	231,879 (80,960)
EC2	1097	5,226,698 (505,792)	142,140 (62,816)
PA1	102	6,677,430 (325,747)	257,809 (93,830)
PA2	533	6,690,535 (301,819)	655,035 (1,669,049)
NG1	398	2,121,269 (35,734)	39,070 (7,620)
NG2	675	2,135,000 (162,840)	57,727 (16,653)

Table 3-3. HyAsP predicted plasmid characteristics. Standard deviation (SD) is in brackets where applicable. As sequencing reads were unavailable for PA2, we were unable to perform HyAsP.

Dataset	Number of samples	Total number of putative plasmid contigs (circular + non-circular)	Number of circular plasmids	Average length of circular plasmids in base pairs (SD)	Average read depth of circular plasmids (SD)	Gene density of circular plasmids (SD)
EC1	115	813	178	112,707 (188,005)	27.08 (59.32)	0.31 (0.33)
EC2	1097	8707	1188	14,626 (26,577)	33.28 (70.42)	0.71 (0.14)
PA1	102	54	1	95,654	1.47	0.56
PA2	Sequencing reads not available					
NG1	398	548	413	6,467 (8,640)	32.17 (14.64)	0.79 (0.1)
NG2	675	800	797	9,405 (12,823)	23.99 (30.70)	0.83 (0.10)

Feature selection drives AMR prediction model performance for *E. coli, P. aeruginosa,* and *N. gonorrhoeae*

In the *E. coli* datasets, filtering resistance determinants had little effect on AMR prediction model performance (Figure 3-1, Supplementary Figure 3-11 to 3-15) while use of plasmid-borne resistance determinants exclusively instead of chromosome and plasmid-borne resistance determinants together decreased AMR prediction model performance using most evaluation metrics (Figure 3-2, Supplementary Figure 3-16). In the *P. aeruginosa* datasets, the PS or GS filter improved meropenem resistance prediction for both datasets (Figure 3-3), as well as improved ceftazidime, ciprofloxacin, and piperacillin-tazobactam resistance prediction for the PA2 dataset (Supplementary Figure 3-17 to 3-19). The PS filter improved amoxicillin-clavulanic acid, cefixime, cefoxitin, ceftriaxone, and trimethoprim-sulfamethoxazole resistance prediction in dataset PA1

(Supplementary Figure 3-20 to 3-24), while ciprofloxacin, azithromycin, penicillin, and tetracycline resistance prediction models were improved using chromosome and plasmid resistance determinants together combined with the PS filter or the GS filter (Figure 3-4, Supplementary Figure 3-25 to 3-27). Use of feature filtering for plasmid-borne resistance determinants had little effect on *N. gonorrhoeae* AMR prediction performance and plasmid-borne resistance determinants generally produced poorer models compared to using chromosome and plasmid resistance determinants, with a few exceptions that illustrated an interaction between feature selection and algorithm choice (Figure 3-5, Supplementary Figure 3-28). In the NG1 dataset, the accuracy, F1, recall, and log loss of penicillin resistance prediction is improved when using random forest and plasmid-borne resistance determinants (Figure 3-5), but the same effect was not observed for penicillin resistance in dataset NG2. Yet, the log loss of all cefixime resistance prediction models was improved using plasmid-borne resistance determinants (Supplementary Figure 3-28).

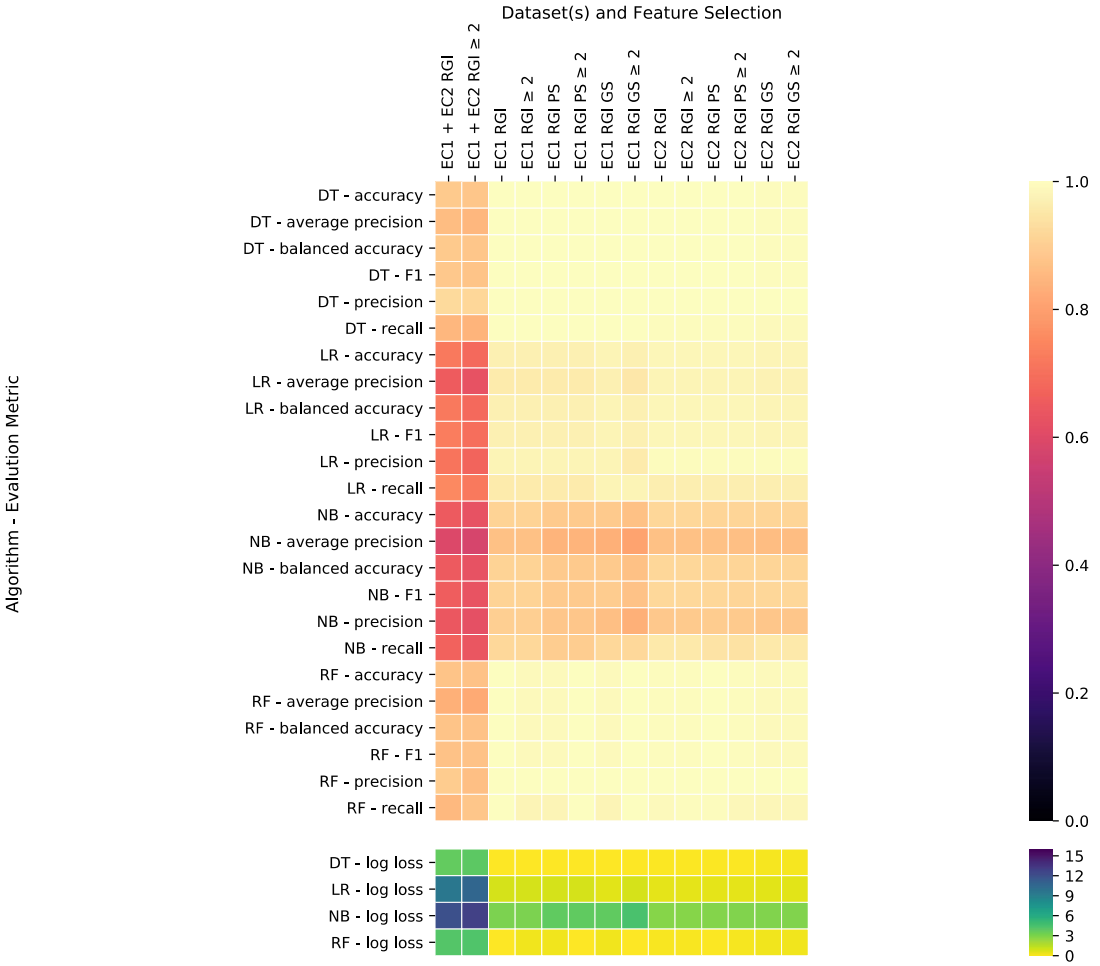


Figure 3-1. AMR prediction models for *E. coli* ciprofloxacin resistance across two datasets (EC1 and EC2). Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. The x-axis describes different feature filtering methods. For resistance genes, we represented them in two different ways: only the resistance gene or the resistance gene and its criteria (PS). For example, if a Perfect and Strict TEM-1 were identified in a sample, in the first representation, there would only be one TEM-1 feature, and in the second representation there would be a TEM-1 Perfect and TEM-1 Strict feature. Only using resistance determinants that were found in at least two or more samples within the dataset are represented as (≥2). A Grantham score (GS) filter to remove any resistance genes that were considered to have radical amino acid changes. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. The evaluation metrics on the top heatmap indicate qualities to be maximized, thus the closer to one (the more yellow), the better the prediction model. Whereas log loss is ideally zero (the more yellow) as it indicates a more probable, better prediction model.

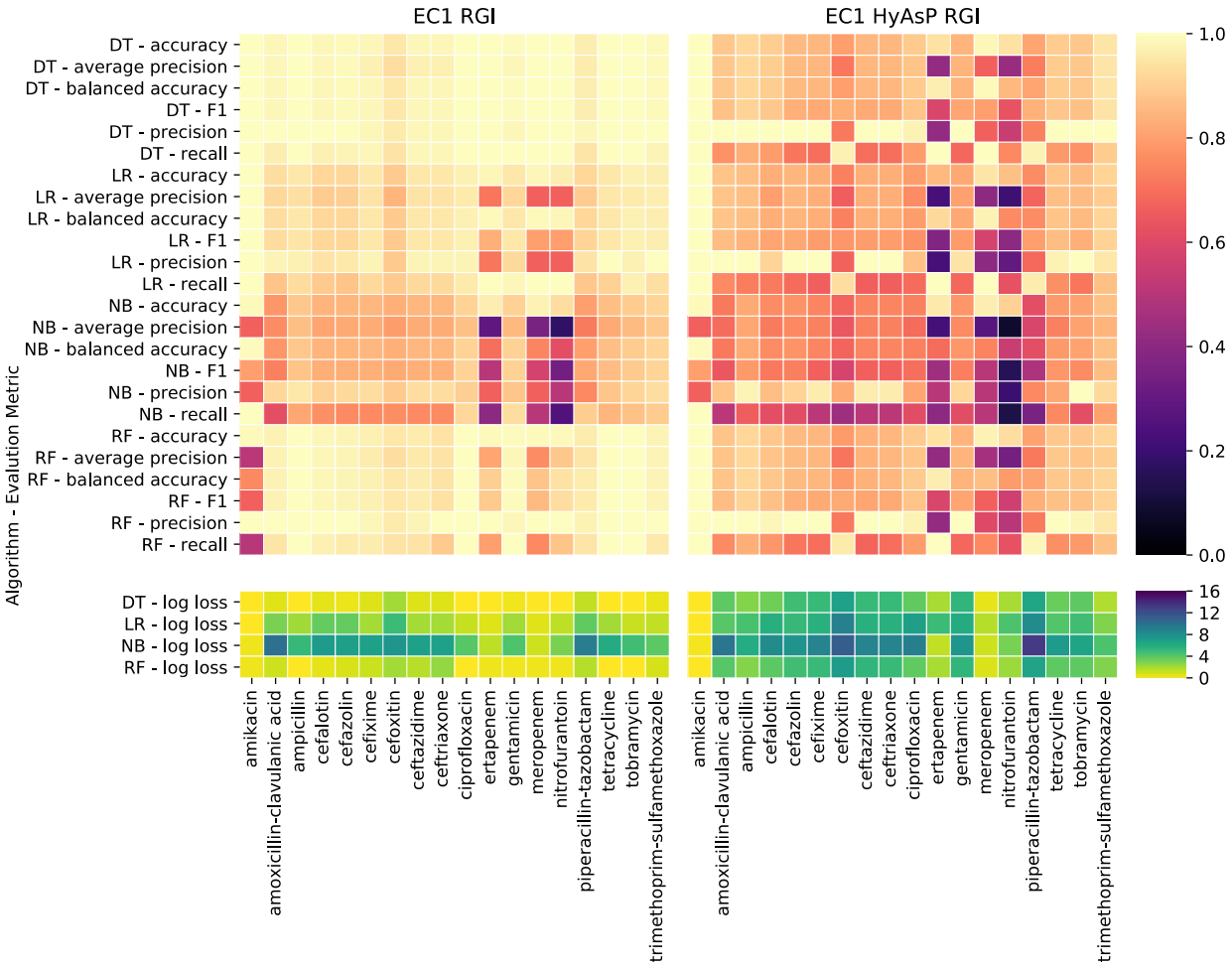


Figure 3-2. AMR prediction models for *E. coli* resistance to compare using SPAdes (chromosome + plasmid) or HyAsP (plasmid) assemblies in dataset EC1. Each square represents an AMR prediction model created using an algorithm, known resistance determinants (no filter), and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

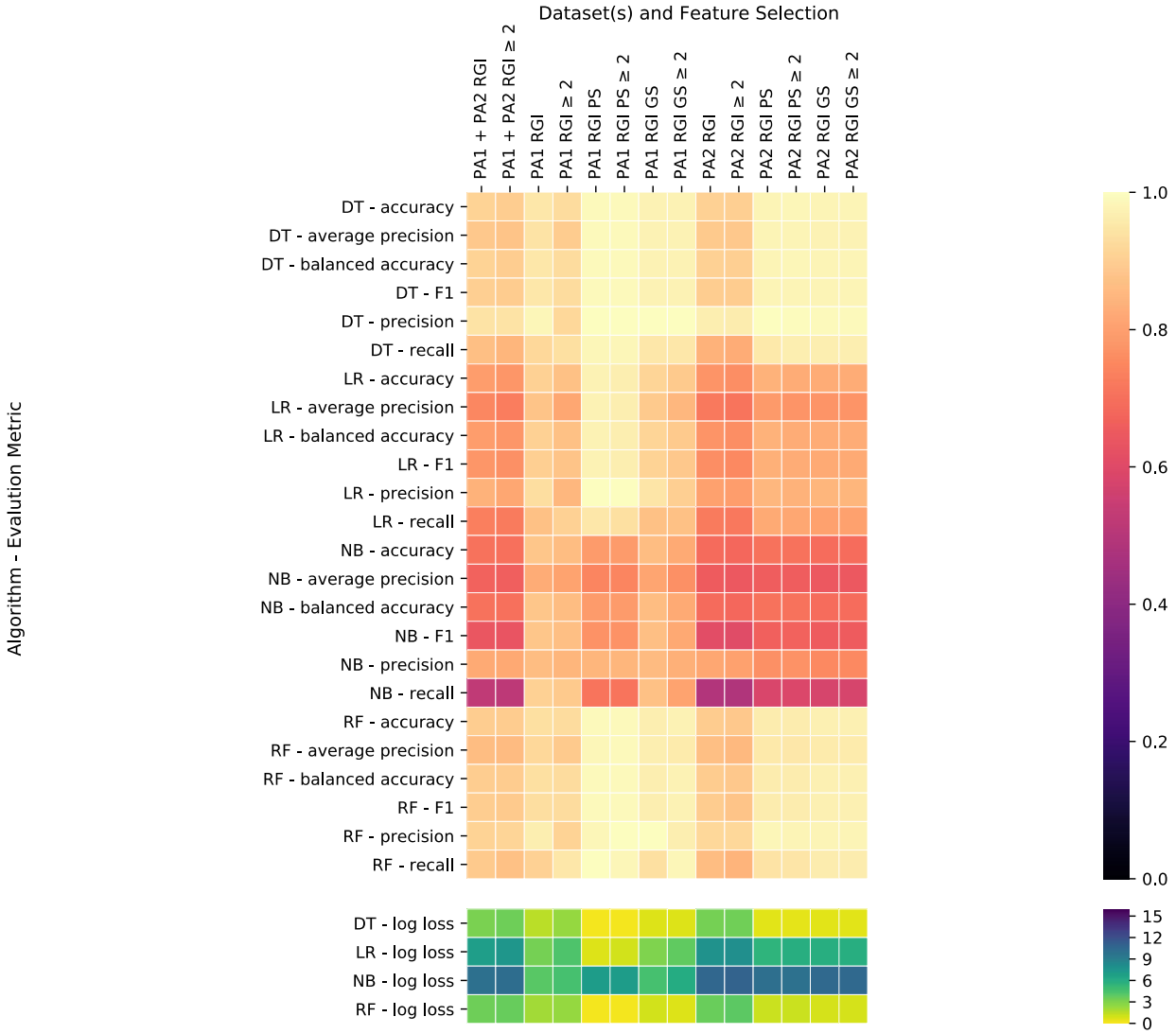


Figure 3-3. AMR prediction models for *P. aeruginosa* meropenem resistance across two datasets (PA1 and PA2). Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.

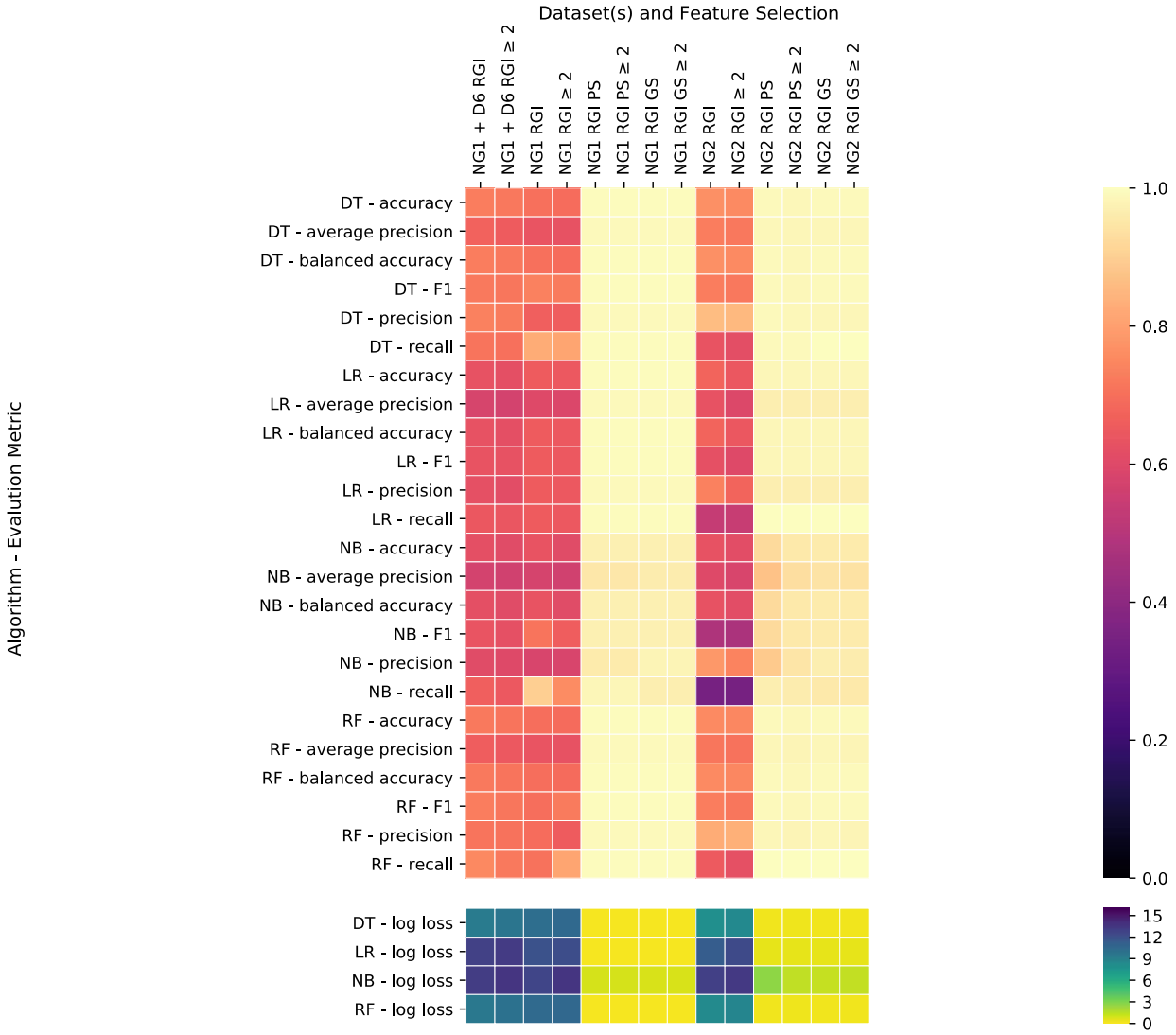


Figure 3-4. AMR prediction models for *N. gonorrhoeae* ciprofloxacin resistance across two datasets (NG1 and NG2). Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.

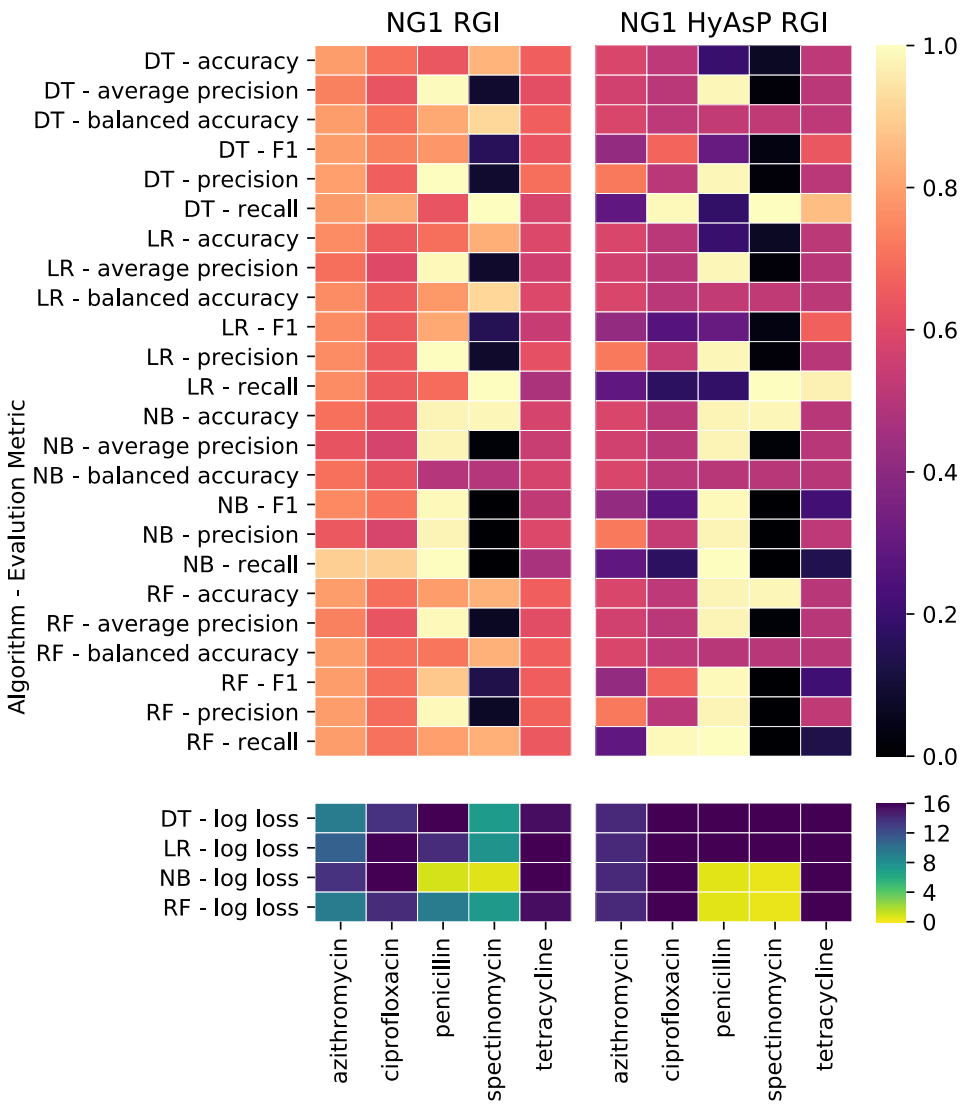


Figure 3-5. AMR prediction models for *N. gonorrhoeae* resistance to compare using SPAdes (chromosome + plasmid) or HyAsP (plasmid) assemblies in dataset NG1. Each square represents an AMR prediction model created using an algorithm, known resistance determinants (no filter), and assessed using an evaluation metric where its colour represents the performance. For more detail on performance interpretation, see Figure 3-1.

Mutation identification is dependent on reference sequence selection and can improve AMR prediction models

As an alternative to dependence upon known resistance determinants as predicted by RGI, breseq was used to identify all single nucleotide polymorphisms (SNPs) in each sample using a collection of annotated reference genome sequences. Using these SNPs as genetic features in model training and analogous to the RGI results, more mutations were identified within larger datasets (Supplementary Figure 3-29 to 3-33). For both *E. coli* datasets, using *E. coli* O157 str. Sakai as a reference identified more mutations than using *E. coli* O83:H1 str. NRG 857C (Supplementary Figure 3-29 to 3-30). Using *P. aeruginosa* UCBPP-PA14 as a reference generated more mutations than using *P. aeruginosa* PAO1 (Supplementary Figure 3-31). In contrast, in the *N. gonorrhoeae* datasets, using *N. gonorrhoeae* WHOF generated the most mutations (39,735) in dataset NG1 (n=398, Supplementary Figure 3-32) while *N. gonorrhoeae* WHOV identified the most mutations (70,578) in dataset NG2 (n=660, Supplementary Figure 3-33). In dataset NG1, use of 15 different references identified 36,000 to 40,000 mutations, whereas in dataset NG2 the number of mutations ranged from 35,000 to 70,000.

When using SNPs as features in model generation for *E. coli*, use of two different reference sequences did not have a large impact on AMR model prediction performance regardless of the dataset, algorithm, or evaluation metric (Figure 3-6, Supplementary Figure 3-34). We observed that using SNPs combined with naïve Bayes overall as well as SNPs combined with random forest for some antibiotics (e.g., nitrofurantoin in EC1 and meropenem in EC2) generated poor prediction models (Figure 3-6, Supplementary Figure

3-34). However, using logistic regression or decision trees created strong prediction models regardless of the evaluation metric. Using the mutations generated from EC1 and EC2 generally performed better than using known resistance genes predicted by RGI (Figure 3-6, Supplementary Figure 3-34,3-35). Similarly, using one *P. aeruginosa* dataset, PA1, to generate mutations using two different reference sequences created subtle differences in prediction model quality (e.g., using *P. aeruginosa* PAO1 performed slightly worse for ciprofloxacin resistance prediction models using naïve Bayes) (Figure 3-7). Furthermore, using *P. aeruginosa* mutations generated from reference sequences creates better AMR prediction models compared to using known resistance determinants predicted by RGI (Figure 3-7, Supplementary Figure 3-36). For the *N. gonorrhoeae* datasets, we used 15 reference sequences to generate mutations and there were subtle differences between each in the quality of the prediction model (Figure 3-8, Supplementary Figure 3-36 to 3-41), e.g., using *N. gonorrhoeae* WHOP and NG1 for penicillin resistance prediction (Figure 3-8). Naïve Bayes performed poorly for both datasets and every antibiotic using most evaluation parameters (Figure 3-7, Supplementary Figure 3-36), whereas random forest only performed poorly for spectinomycin and cefixime resistance prediction using all evaluation metrics except accuracy and precision (Supplementary Figure 3-40, 3-41).

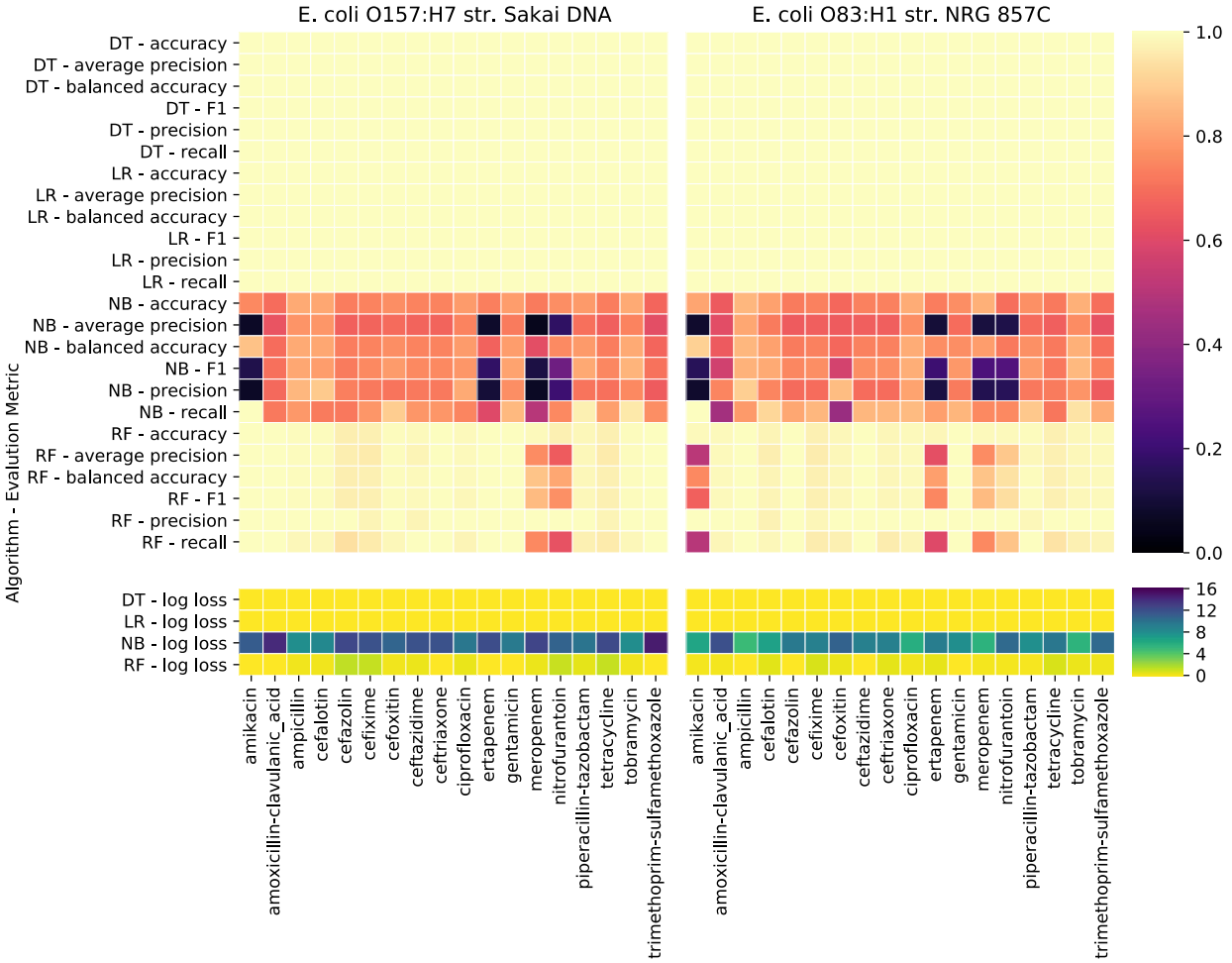


Figure 3-6. AMR prediction models for *E. coli* resistance to compare using two different reference strains for mutation generation in dataset EC1. Each square represents an AMR prediction model created using an algorithm, mutations generated using a reference sequence, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

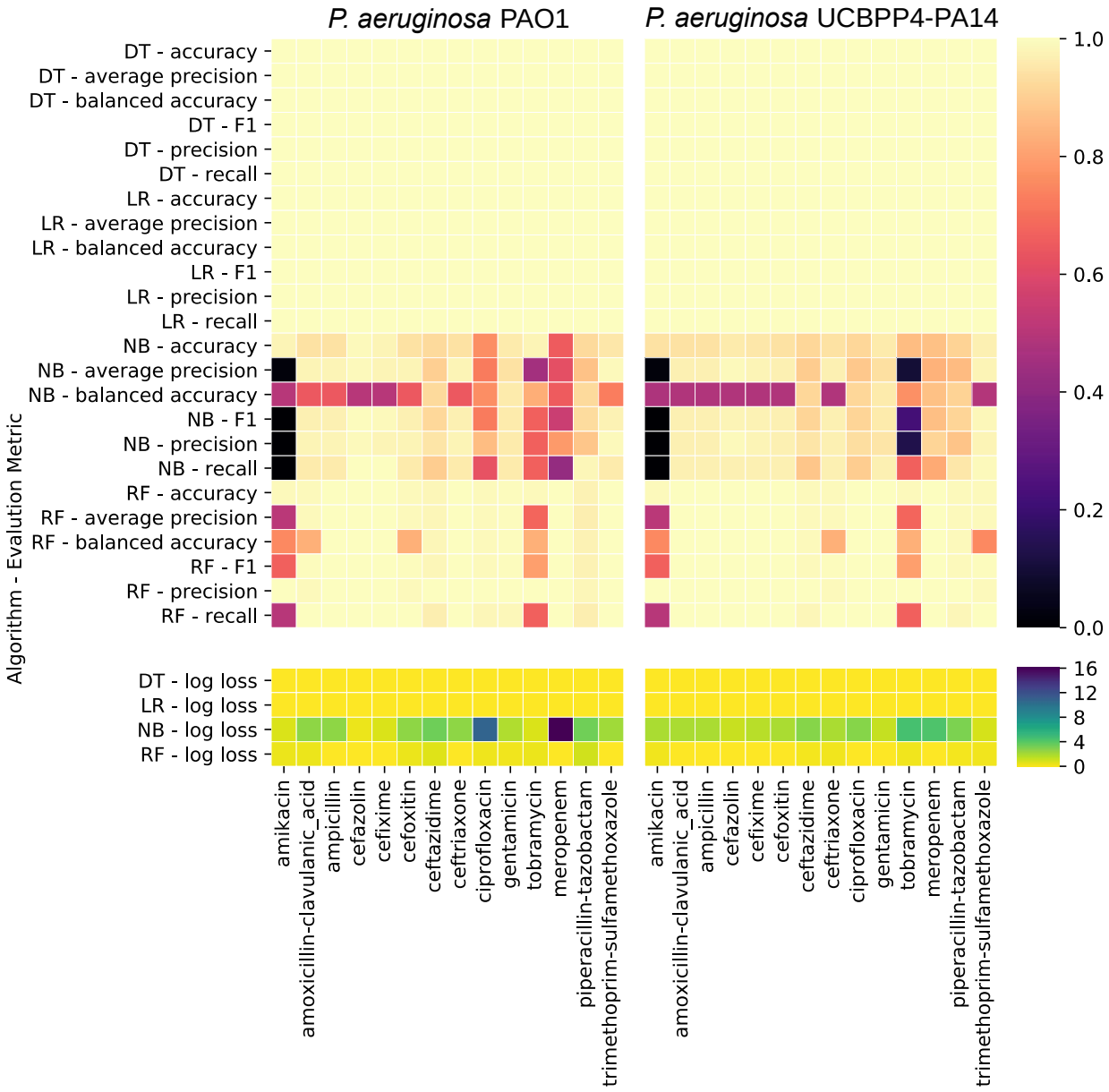


Figure 3-7. AMR prediction models for *P. aeruginosa* to compare using two different reference strains for mutation generation in dataset PA1. Each square represents an AMR prediction model created using an algorithm, mutations generated using a reference sequence, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

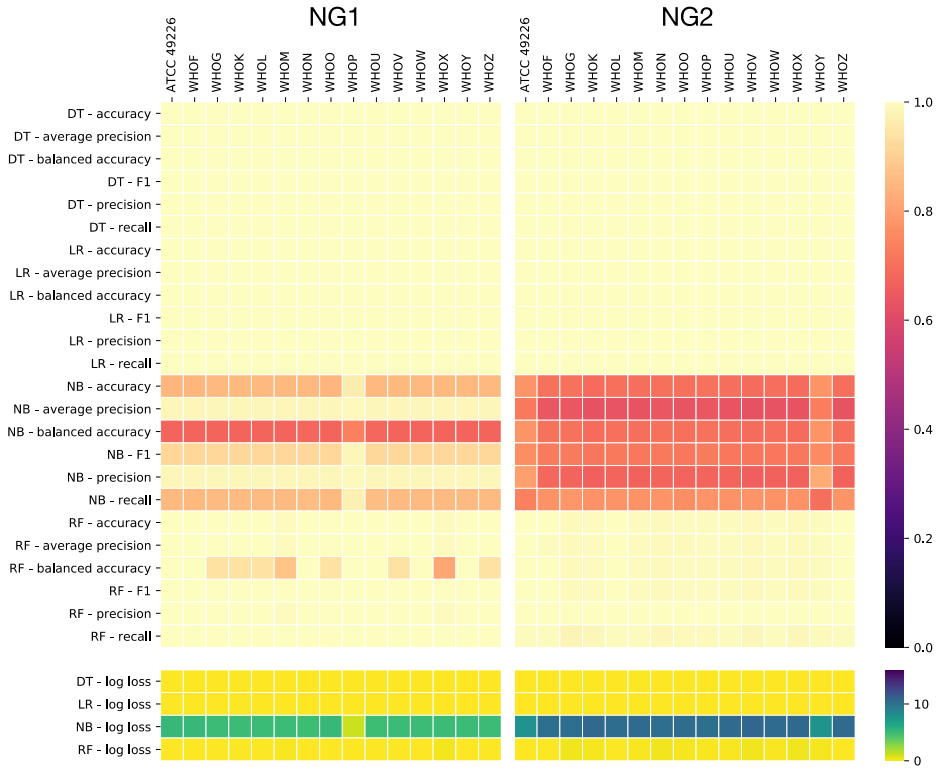


Figure 3-8. Penicillin resistant prediction models for *N. gonorrhoeae* to compare using 15 different reference strains for mutation generation in dataset NG1 and NG2. Each square represents an AMR prediction model created using an algorithm, mutations generated using a reference sequence, and assessed using an evaluation metric where its colour represents the performance. On the x-axis are the 15 different reference strains. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

Evaluation metric and algorithm choice are inter-related

To determine the best algorithm for predicting resistance to a particular antibiotic, one must first choose an evaluation metric to compare across all models. However, choosing different evaluation metrics can change the algorithm that will be chosen. To best illustrate the inter-relatedness of the evaluation metric and the algorithm choice, we compared ciprofloxacin resistance prediction models as ciprofloxacin resistance was the only phenotype that was included in every dataset. Generally, using resistance determinants and regardless of any evaluation metric chosen, Naïve Bayes performed poorer than decision tree, logistic regression, and random forest (Figure 3-9). However, the poor prediction quality of naïve Bayes is more evident in the *E. coli* and *P. aeruginosa* datasets than in the *N. gonorrhoeae* datasets (Figure 3-9). Even using mutations as features, use of Naïve Bayes is generally a poor algorithm choice (Figure 3-6 to 3-8), yet there are some exceptions, e.g., using average precision or precision to evaluate the Naïve Bayes prediction models derived from dataset NG1 (Figure 3-8). In addition, there are also differences in prediction model performance across the antibiotic resistance phenotypes, e.g., the random forest meropenem and nitrofurantoin resistance prediction models perform poorer than other antibiotic resistance phenotypes using a number of different evaluation metrics (Figure 3-6).

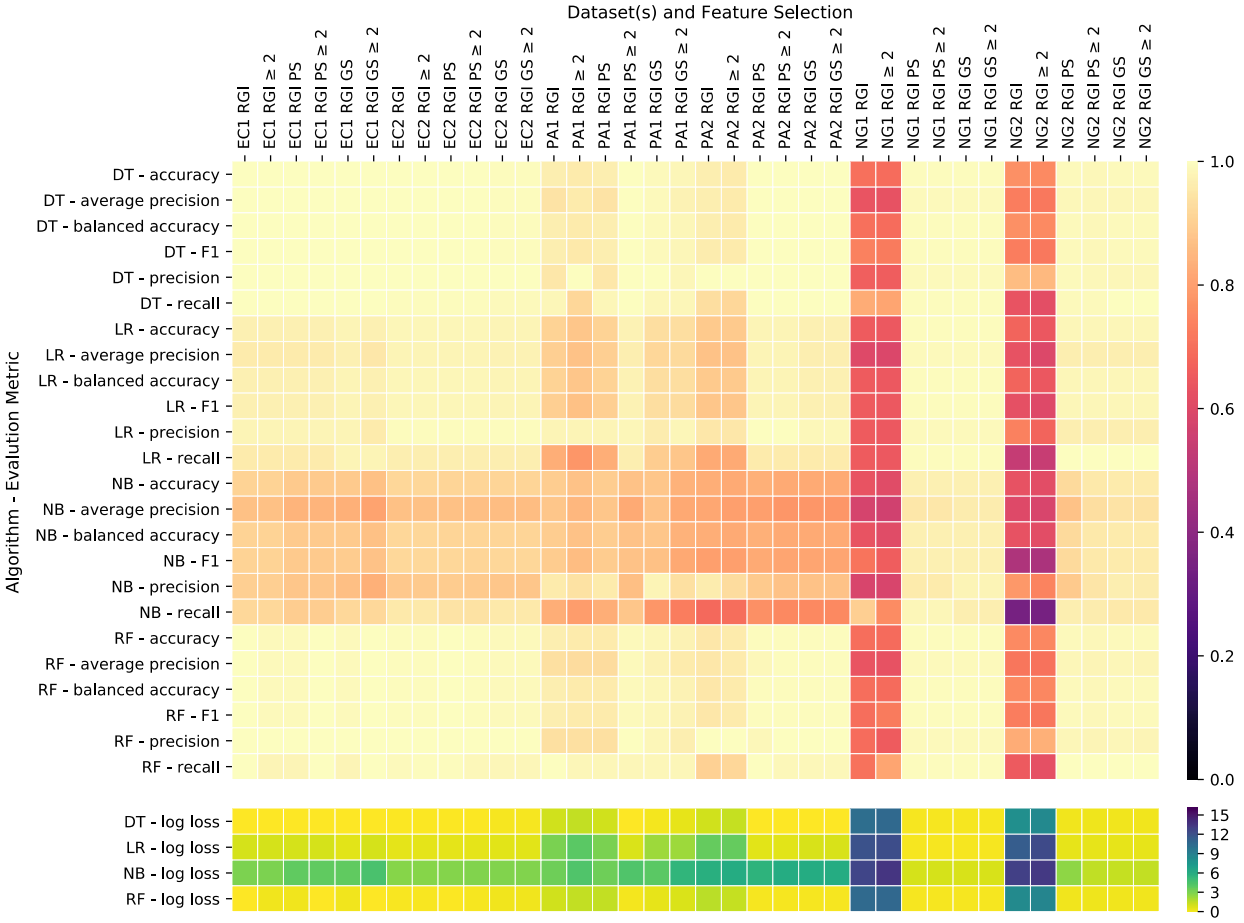


Figure 3-9. Ciprofloxacin prediction models for across all species and datasets with feature filtering. Each square represents an AMR prediction model created using an algorithm, resistance determinants, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.

Prediction model performance is specific to antibiotic

Regardless of whether known resistance determinants (RGI) or mutations derived from a reference sequence (breseq) were used as features, the quality of AMR prediction models depended upon the antibiotic examined. Across all datasets tested, we observed differences in prediction performance for every antibiotic. In the *E. coli* datasets, ertapenem, meropenem, and nitrofurantoin resistance were more difficult to predict using known resistance genes or mutations (Figure 3-2, 3-6), while with the *P. aeruginosa* datasets, ceftazidime and meropenem prediction models were not as strong as prediction models for other antibiotics (Figure 3-7, 3-10). In the *N. gonorrhoeae* datasets, using known resistance genes there are differences in prediction performance for each antibiotic (Figure 3-11), but the difference in prediction performance across antibiotics is more subtle when using mutations (Figure 3-8, Supplementary Figure 3-38 to 3-41). One observable difference is the poor prediction performance of using random forest for spectinomycin and cefixime resistance prediction (Supplementary Figure 3-40, 3-41).

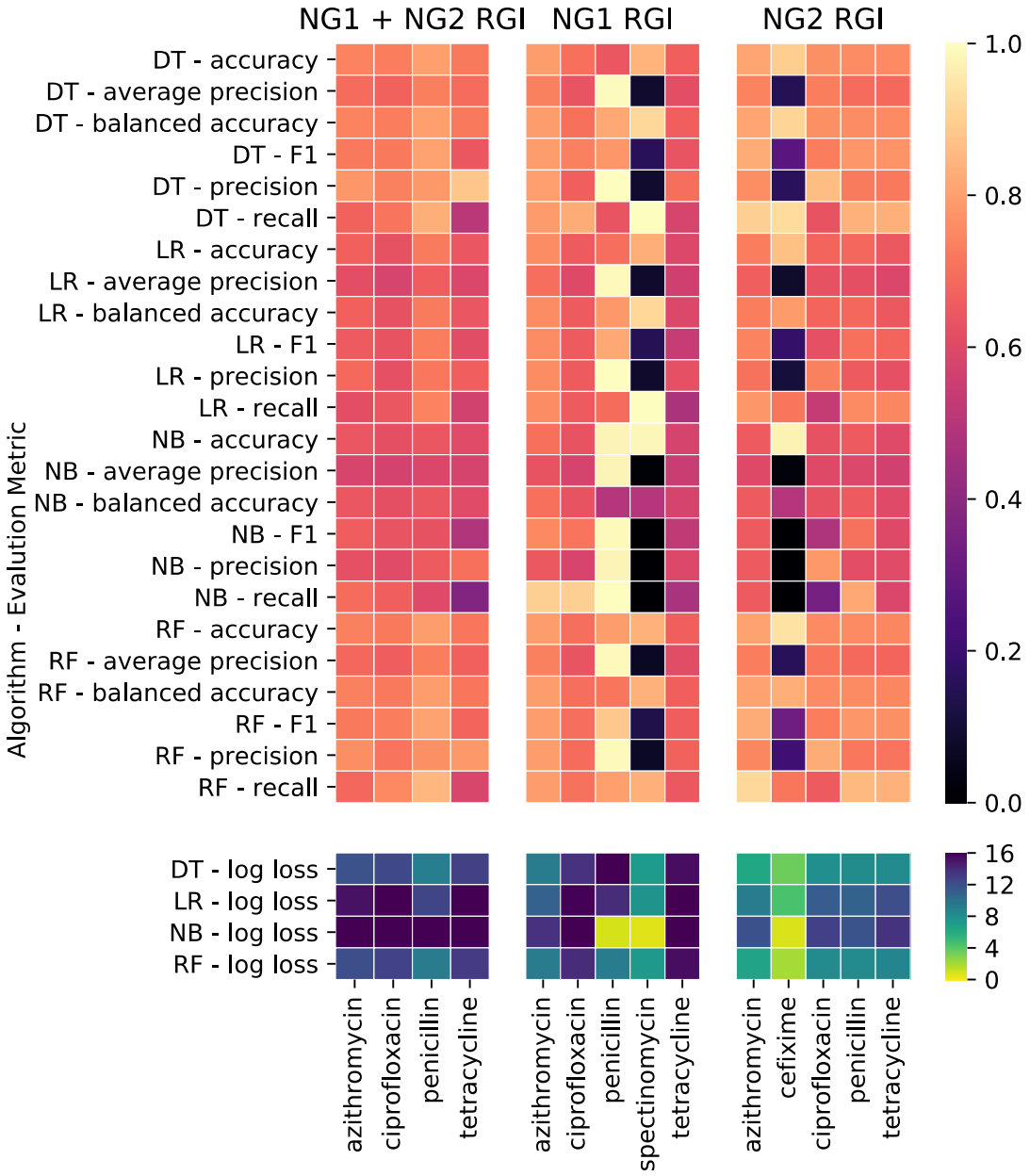


Figure 3-10. AMR prediction models using known resistance determinants and no filtering for *N. gonorrhoeae* across two datasets (NG1 and NG2). Each square represents an AMR prediction model created using an algorithm, known resistance determinants, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

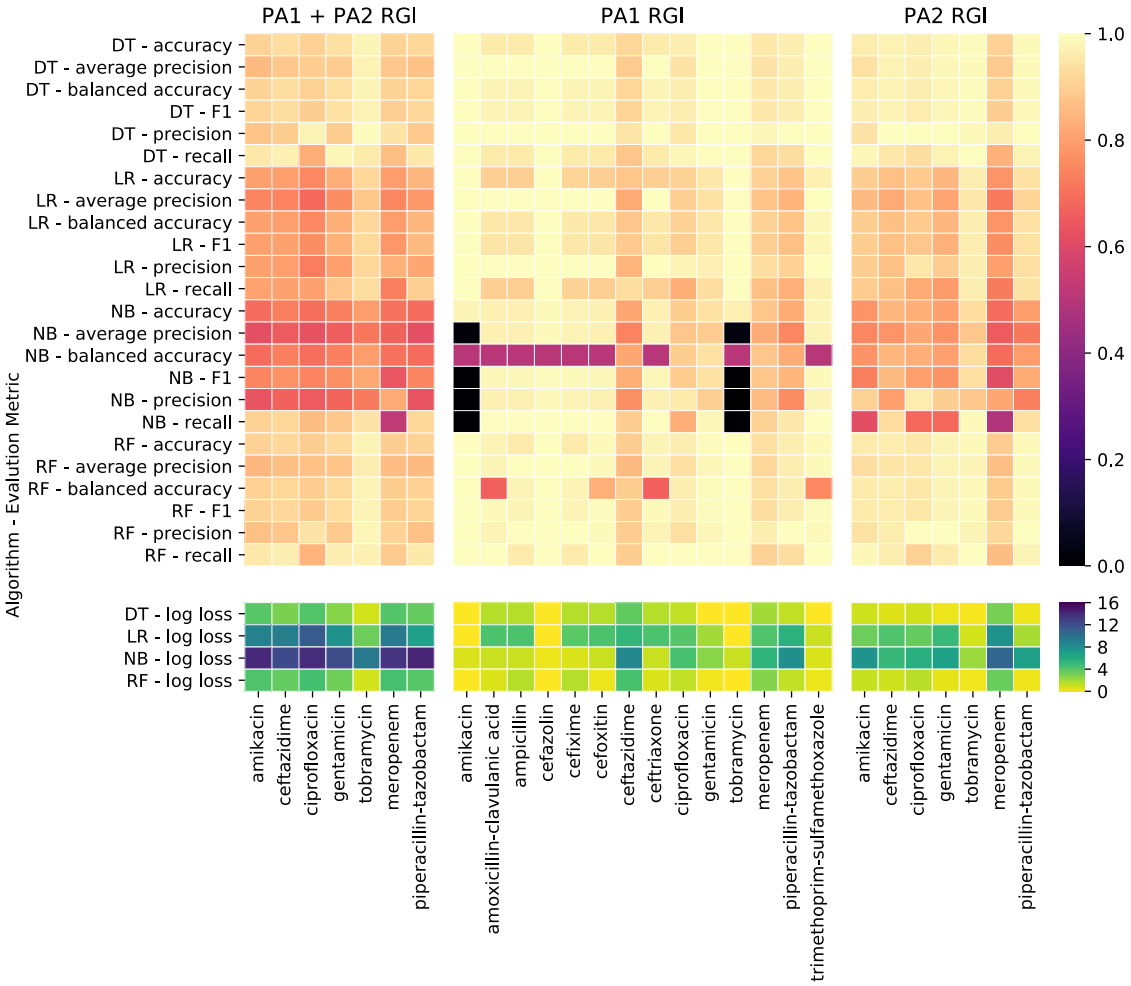


Figure 3-11. AMR prediction models using known resistance determinants and no filtering for *P. aeruginosa* across two datasets (PA1 and PA2). Each square represents an AMR prediction model created using an algorithm, known resistance determinants, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

Pathogen and dataset drive AMR prediction model quality

We used geographically different datasets of the same pathogen to create AMR prediction models. For *E. coli* ertapenem and nitrofurantoin resistance prediction, using EC2 creates prediction models that are generally worse than using EC1 (Supplementary Figure 3-35, Supplementary Figure 3-42, 3-43), while for *P. aeruginosa* amikacin resistance prediction using PA1 performed worse than using PA2 when using naïve Bayes and random forest (Supplementary Figure 3-44). In general, using PA2 for ceftazidime resistance prediction performed better than using PA1 (Supplementary Figure 3-45), but PA2 performed worse for meropenem resistance prediction (Supplementary Figure 3-46). For *P. aeruginosa* piperacillin-tazobactam resistance prediction, using PA1 performed slightly better than using PA2 (Supplementary Figure 3-47), whereas using naïve Bayes and PA2 or logistic regression and PA1 generated better tobramycin resistance prediction models (Supplementary Figure 3-48).

Dataset specificity is particularly evident when combining datasets of the same species. In both *E. coli* and *P. aeruginosa*, when we combined their respective datasets (EC1+EC2 and PA1+PA2) to create prediction models, the quality of the prediction models generally decreased (Supplementary Figure 3-35, 3-36), but decreasing prediction model quality when combining datasets is less evident using the *N. gonorrhoeae* datasets (Supplementary Figure 3-49).

While the decision tree or random forest algorithm would be chosen for ciprofloxacin resistance across all pathogens, the *E. coli* prediction models are better than the *P. aeruginosa* prediction models, which are better than those of *N. gonorrhoeae*

(Figure 3-9). Similarly, for *E. coli*, *P. aeruginosa*, and *N. gonorrhoeae* cefixime resistance prediction, the quality of prediction models is better in *P. aeruginosa* and *E. coli* in contrast to *N. gonorrhoeae*, particularly when using precision, average precision and F1 metrics (Supplementary Figure 3-50). Using naïve Bayes is generally poor for *E. coli* resistance prediction regardless of which evaluation metric (Supplementary Figure 3-51), whereas with *P. aeruginosa* only evaluating naïve Bayes with balanced accuracy equates to poor prediction models.

Stratification based on site of infection can improve AMR prediction models

For the *E. coli* datasets, we had enough compiled data to stratify by site of infection (e.g., blood or urine) before generating the AMR prediction models and stratification by urine improved cefazolin, ciprofloxacin, and trimethoprim-sulfamethoxazole resistance prediction models, but reduced the quality of nitrofurantoin resistance prediction models (Figure 3-12). When we excluded any known resistance genes only found in one sample, some of the nitrofurantoin resistance prediction models improved using random forest and stratification (Supplementary Figure 3-52).

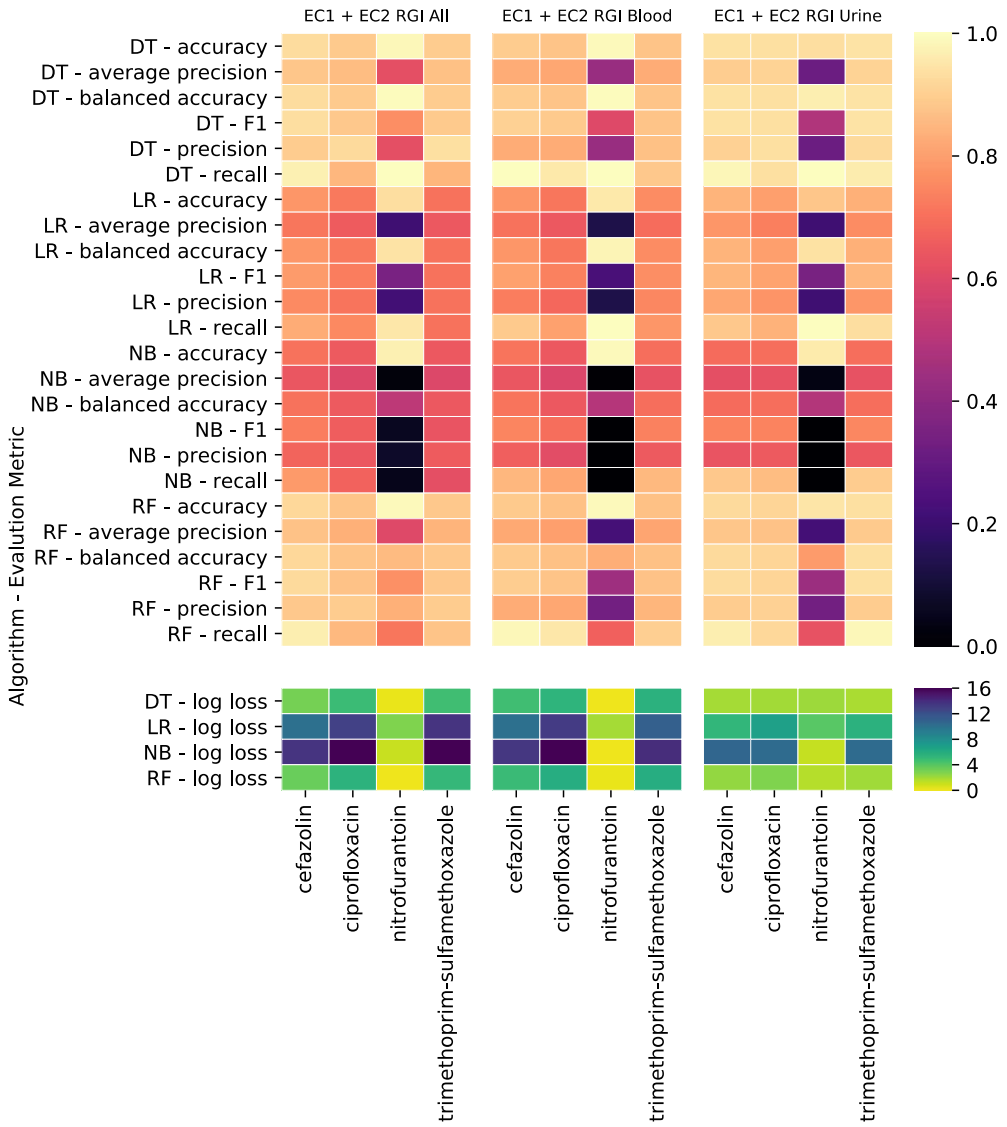


Figure 3-12. AMR prediction models using resistance determinants and no filtering for *E. coli* across two datasets (EC1 and EC2) stratified by site of infection. Each square represents an AMR prediction model created using an algorithm, known resistance determinants, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

DISCUSSION

Using machine learning to create AMR prediction models can create new or improve existing molecular diagnostics. Since it is currently difficult to compare AMR prediction publications due to the variety in datasets and parameters used, we show the effect of these parameters on AMR prediction model performance. We show that AMR prediction model performance depends on dataset, algorithm, and evaluation metric. Furthermore, we show that AMR prediction model performance also relies on antibiotic, pathogen, feature selection and stratification of datasets.

While AMR prediction models rely on a number of different parameters, many previously published AMR prediction studies have limited their use of such variables. One publication by Hicks *et al.* has evaluated parameters affecting AMR *Neisseria gonorrhoeae, Klebsiella pneumoniae* and *Acinetobacter baumannii* prediction model performance and reliability (Hicks *et al.*, 2019). The authors used set covering machine and random forest algorithms with *k*-mers derived from genome assemblies and showed that AMR prediction model accuracy varies by antibiotic, genomic diversity, and pathogen using balanced accuracy, sensitivity, and specificity as evaluation metrics. Our work builds upon their publication to show their similar findings in *Escherichia coli* and *Pseudomonas aeruginosa*. Furthermore, we use additional algorithms, evaluation metrics, and biologically relevant feature filtering methods to identify their effect on AMR prediction model performance.

To our knowledge, there has been one AMR prediction modeling publication that has compared the use of different biologically relevant genetic features and feature

filtering methods on AMR prediction model performance (Aytan-Aktug *et al.*, 2020). Using known resistance determinants compared to using mutations has implications on performance, interpretability, and generalizability of AMR prediction models. Typically, tens to hundreds of known resistance determinants compared to thousands when mutations are used as genetic features. The sheer number of features when using mutations requires more computational resources to build AMR prediction models compared to using known resistance determinants. In addition, having hundreds of thousands, if not millions, of mutations decreases our ability to interpret the AMR prediction model, e.g., the mutations important for AMR prediction. With the limitations of using mutations, it is nonetheless important to highlight that use of mutations can improve *E. coli*, *P. aeruginosa*, and *N. gonorrhoeae* AMR prediction performance for a number of different antibiotics. This is biologically reflective, as many *N. gonorrhoeae* resistance phenotypes are mediated or caused by mutations (Ng, Martin, Liu, & Bryden, 2002; Shafer & Folster, 2006; Unemo & Shafer, 2014). Thus, choosing resistance determinants or mutations implies trade-offs between interpretability and performance. We also compared the use of known resistance determinants identified in SPAdes (chromosome and plasmid) and HyAsP (plasmid) assemblies. Generally, using plasmid-borne resistance determinants decreased AMR prediction model performance in *E. coli* and *N. gonorrhoeae*, suggesting that AMR is driven by resistance determinants in both chromosome and plasmids, which supports current knowledge (Moradigaravand *et al.*, 2018; Unemo & Shafer, 2014). Lastly, as a part of feature selection we filtered genetic features for sequencing error, gene copy number, and amino acid physicochemistry by

including features only found in two or more samples, representing resistance determinants with their RGI Perfect or Strict criteria ('PS'), and using a Grantham Score filter ('G'), respectively. Again, the effect on AMR prediction model performance varies depending on the dataset, antibiotic, and pathogen. For example, amikacin resistance prediction is improved by using the PS representation and GS filter for only dataset PA2 and not PA1 (Supplementary Figure 3-44). In contrast, the PS representation and GS filters improve ciprofloxacin resistance prediction models (using resistance determinants) in both *N. gonorrhoeae* and *P. aeruginosa* datasets, but has limited impact for *E. coli* datasets. This suggests that copy numbers of quinolone resistance genes are important in *E. coli* prediction models (Minh *et al.*, 2012). However, in both *N. gonorrhoeae* and *P. aeruginosa* it has been established that amino acid substitutions predominantly cause ciprofloxacin resistance (Belland, Morrison, Ison, & Huang, 1994; L. Zhao, Wang, Li, He, & Jian, 2020) and thus the PS representation (which does not address mutation copy number) is perhaps determining associative, not causative, relationships between the genes and phenotypes. Alternatively, the GS filter is removing any resistance determinants that are predicted to have radical changes in amino acids disrupting function, thereby removing noise to improve ciprofloxacin resistance models. CARD's RGI software should consider systematic evaluation of Grantham score filters in addition to the current curated bitscore cutoffs.

Selecting an algorithm not only influences AMR prediction performance but also our ability to interpret the genetic features driving the model. Naïve Bayes generally built poor AMR prediction models and we hypothesize that it is because this algorithm

assumes independence among the genetic features, while we know that AMR can be multifactorial (Piddock, 2014). Logistic regression, decision trees, and random forest do not assume independence among genetic features and nearly always performed better than Naïve Bayes.

When choosing an evaluation metric to measure the performance of AMR prediction models, it is essential to understand what the metric is evaluating. For example, Brankin & Fowler argue that sensitivity and specificity are more important than accuracy and precision (Brankin & Fowler, 2019). To select appropriate treatment for the patient and for stewardship, minimizing false negative predictions is a priority, whereas when there are limited antibiotics available, false positives should be minimized (Brankin & Fowler, 2019). Selecting an evaluation metric thus not only has implications on what will be considered the best prediction model, but also the clinical applicability of the model.

Stratifying datasets essentially subgroups data based on a similarity, whether that be patient demographic, geographic location, timeframe, or site of infection. Hicks *et al.* showed variation in performance of prediction models developed from a variety of sampling frames, e.g., temporal, geographic and/or sampling approach (Hicks *et al.*, 2019), and we also similarly show stratifying datasets by site of infection has a variable effect on prediction model performance. Perhaps stratifying by urine as a site of isolation improved cefazolin, ciprofloxacin, and trimethoprim-sulfamethoxazole resistance prediction models because these antibiotics are treatments for urinary tract infections and thus our data reflect treatment at time of sampling (Alanazi, Alqahtani, & Aleanizy, 2018; Car, 2006; C. E. Cox, 1973; Uppala, King, & Patel, 2019). However, the nitrofurantoin

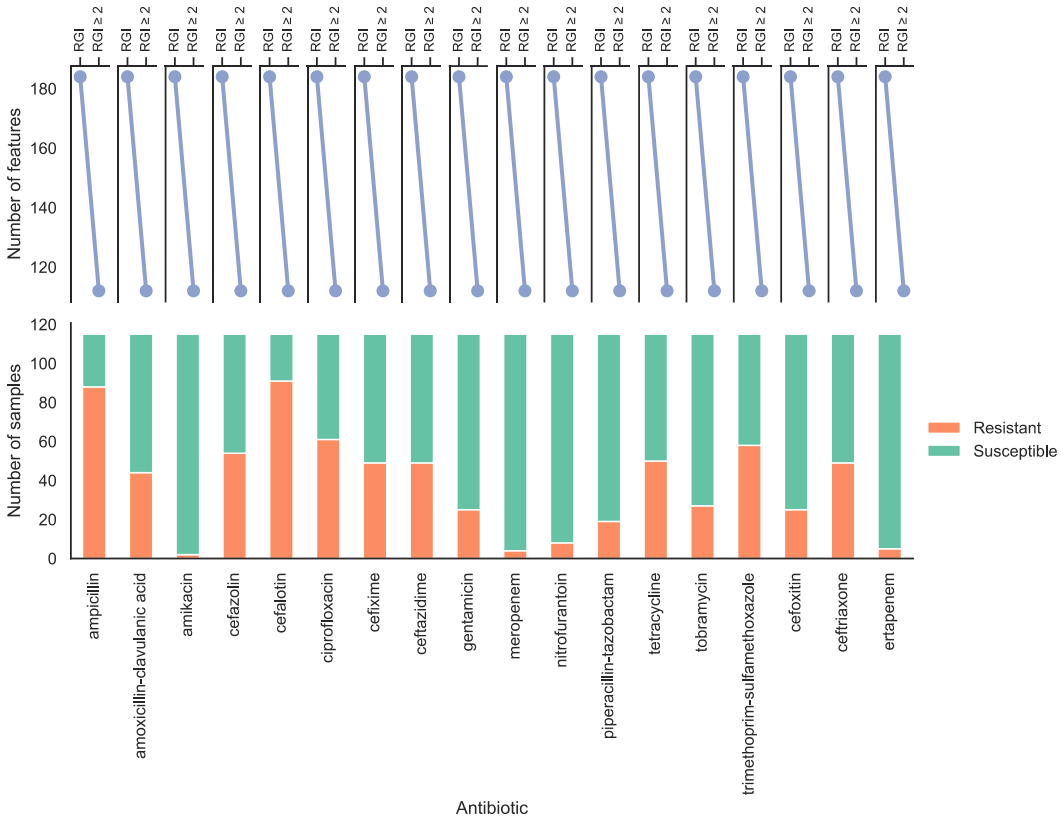
resistance prediction model decreased in performance even though nitrofurantoin is used to treat urinary tract infections (Kashanian *et al.*, 2008), yet this could possibly be attributed to the imbalanced training dataset. In addition, stratifying by site of infection may have inadvertently stratified by phylogeny. In contrast, combining datasets of the same pathogen decreased the performance of all AMR prediction models, which is comparable to previous findings where use of a global gonococcal dataset did not improve prediction accuracy (Hicks *et al.*, 2019). This suggests sampling for genomic and geographic diversity may not necessarily improve AMR prediction model performance and that increased sampling of similar population structures can improve performance. All AMR ML approaches may need to be local.

Resistant and susceptible categorizations are based on the minimum inhibitory concentration (MIC) being higher or lower than a breakpoint value, which are established by the Clinical & Laboratory Standards Institute (CLSI) (CLSI, 2018) and European Committee on Antimicrobial Susceptibility Testing (EUCAST) (EUCAST, 2015). One limitation of our work is that predicting resistant and susceptible categorizations using one guideline may not be generalizable towards the other. Furthermore, while predicting resistant and susceptible may be sufficient for managing individual patients, it provides less surveillance information than predicting MICs. With MIC prediction there is added value where pathogens have MICs approaching the breakpoint values, providing additional information on the evolution and spread of resistance. There are a number of publications that develop models for MIC prediction, sometimes in addition to resistant and susceptible categorizations (Demczuk *et al.*, 2020; Eyre *et al.*, 2019; Eyre *et al.*,

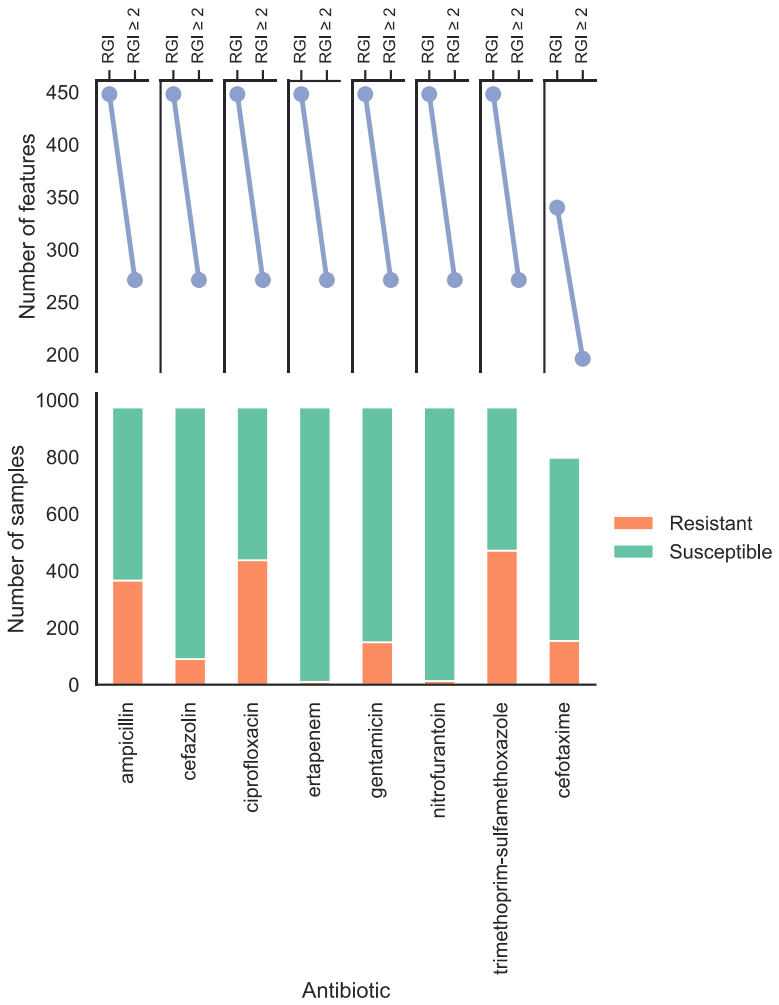
2017; Hicks *et al.*, 2019; Nguyen *et al.*, 2018; Nguyen *et al.*, 2020; Nguyen *et al.*, 2019; Pataki *et al.*, 2020), however to our knowledge, there has not been a publication that tests the effect of different parameters (e.g., features, algorithm, dataset) on MIC prediction model performance. Understanding the parameters that affect AMR and MIC prediction models are important to further elucidate the mechanistic drivers of resistance and acknowledge the models' applicability and limitations in a clinical microbiology lab for patient care and public health surveillance.

SUPPLEMENTARY MATERIAL

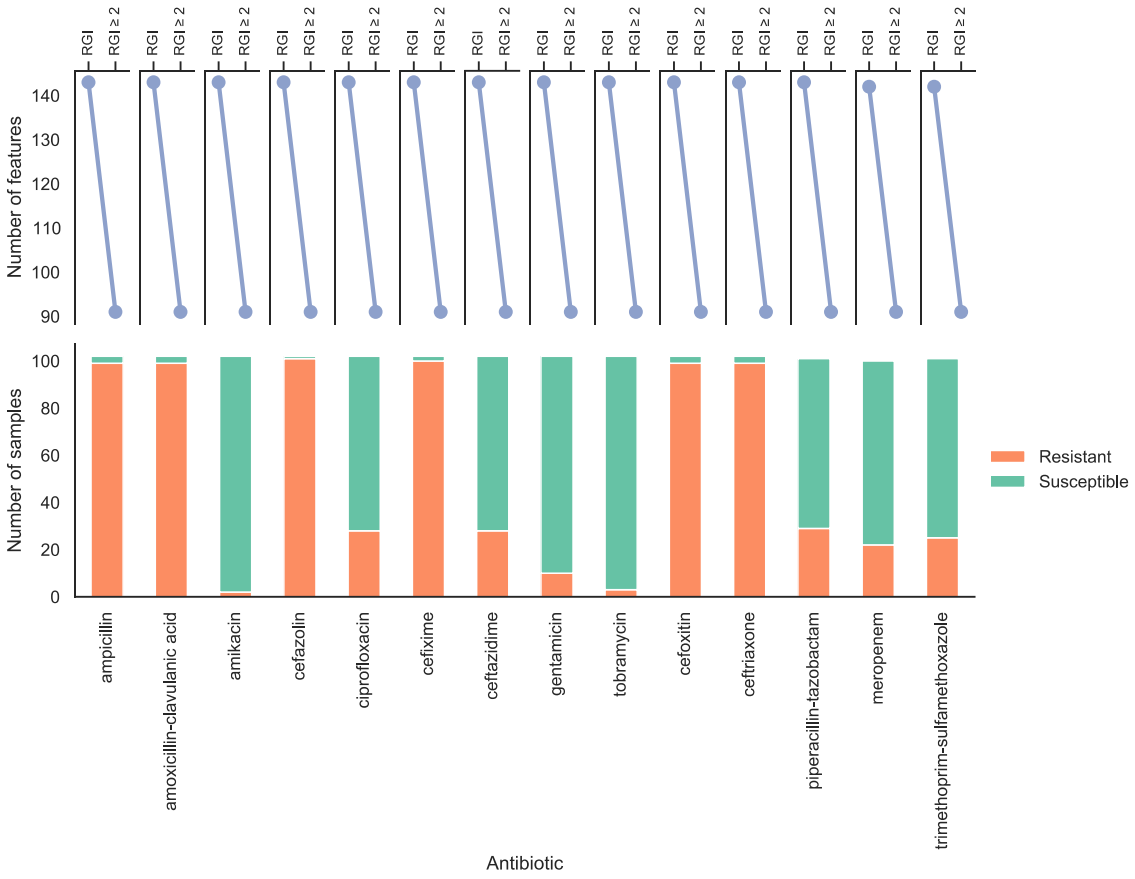
Supplementary Figures



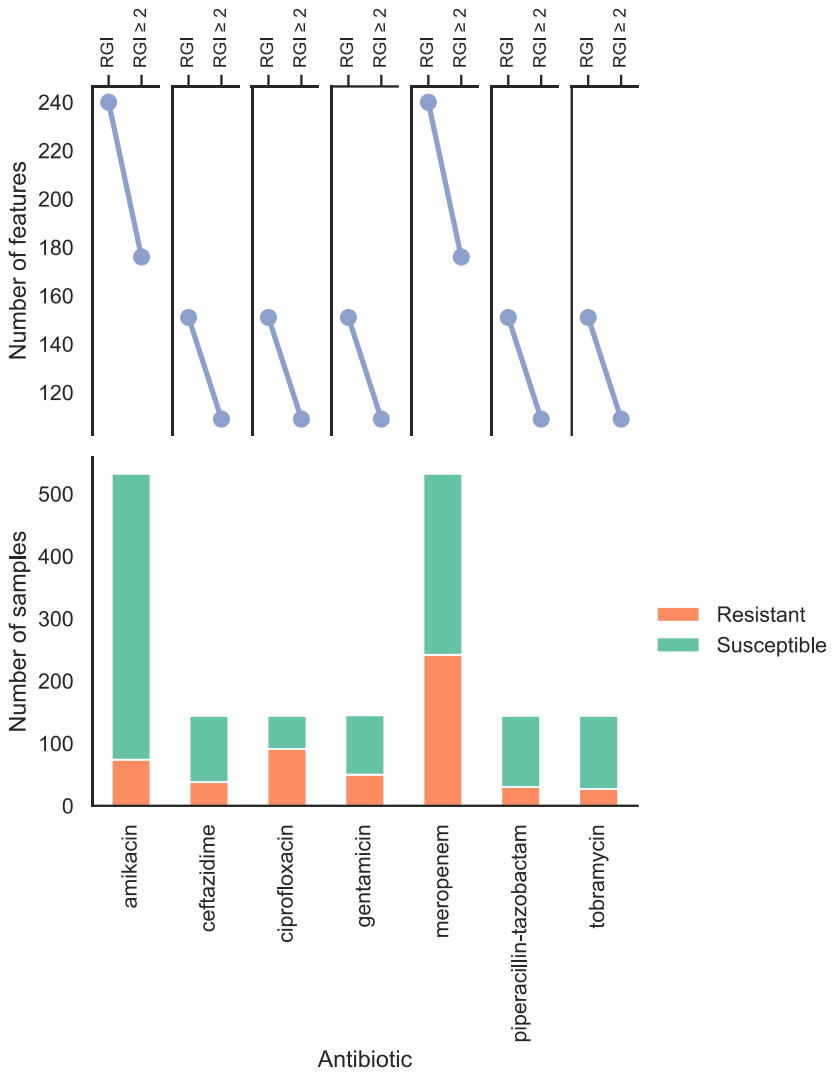
Supplementary Figure 3-1. Resistance determinant prediction and distribution of AMR phenotypes in *E. coli* SPAdes assemblies (Dataset EC1). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



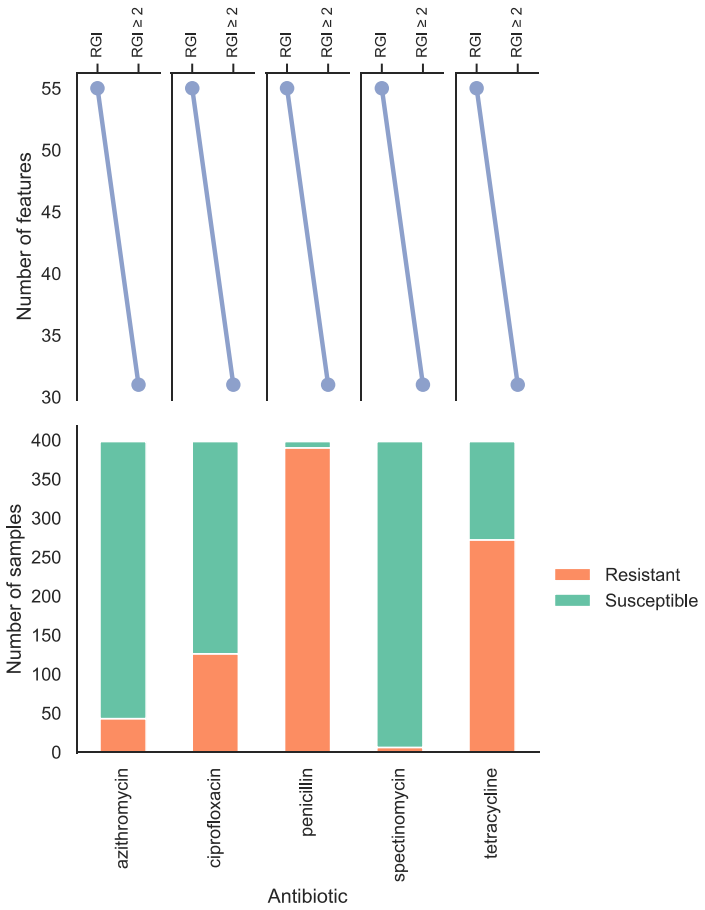
Supplementary Figure 3-2. Resistance determinant prediction and distribution AMR phenotypes in *E. coli* SPAdes assemblies (Dataset EC2). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



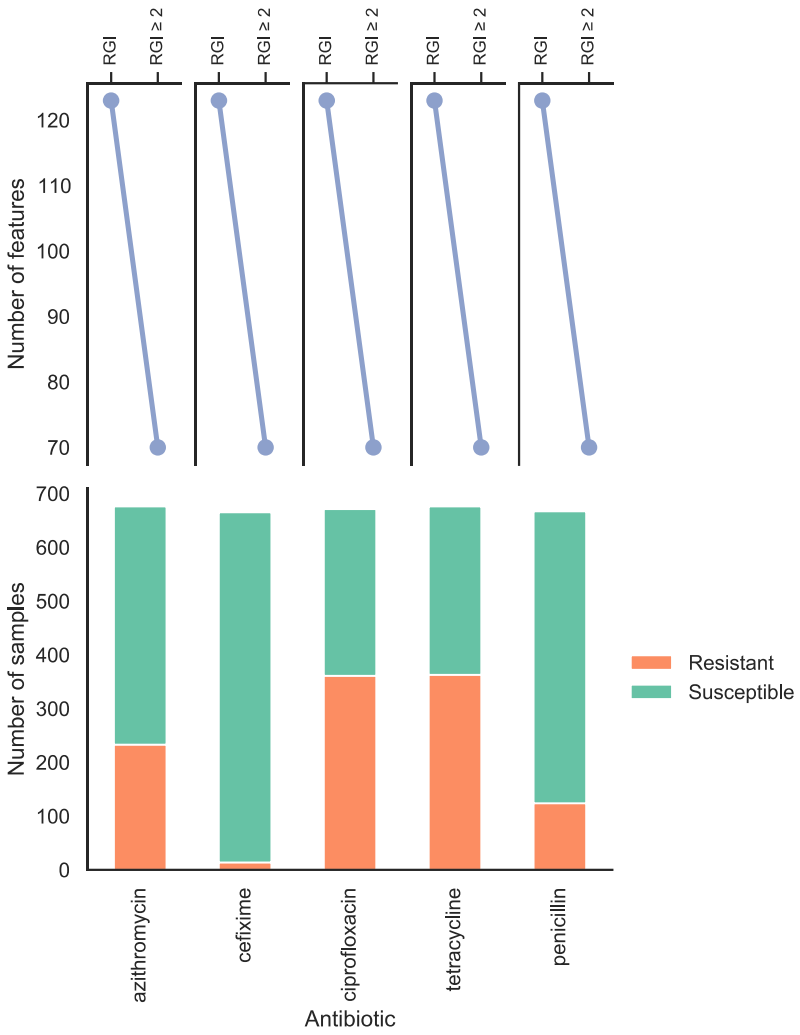
Supplementary Figure 3-3. Resistance determinant prediction and distribution AMR phenotypes in *P. aeruginosa* SPAdes assemblies (Dataset PA1). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



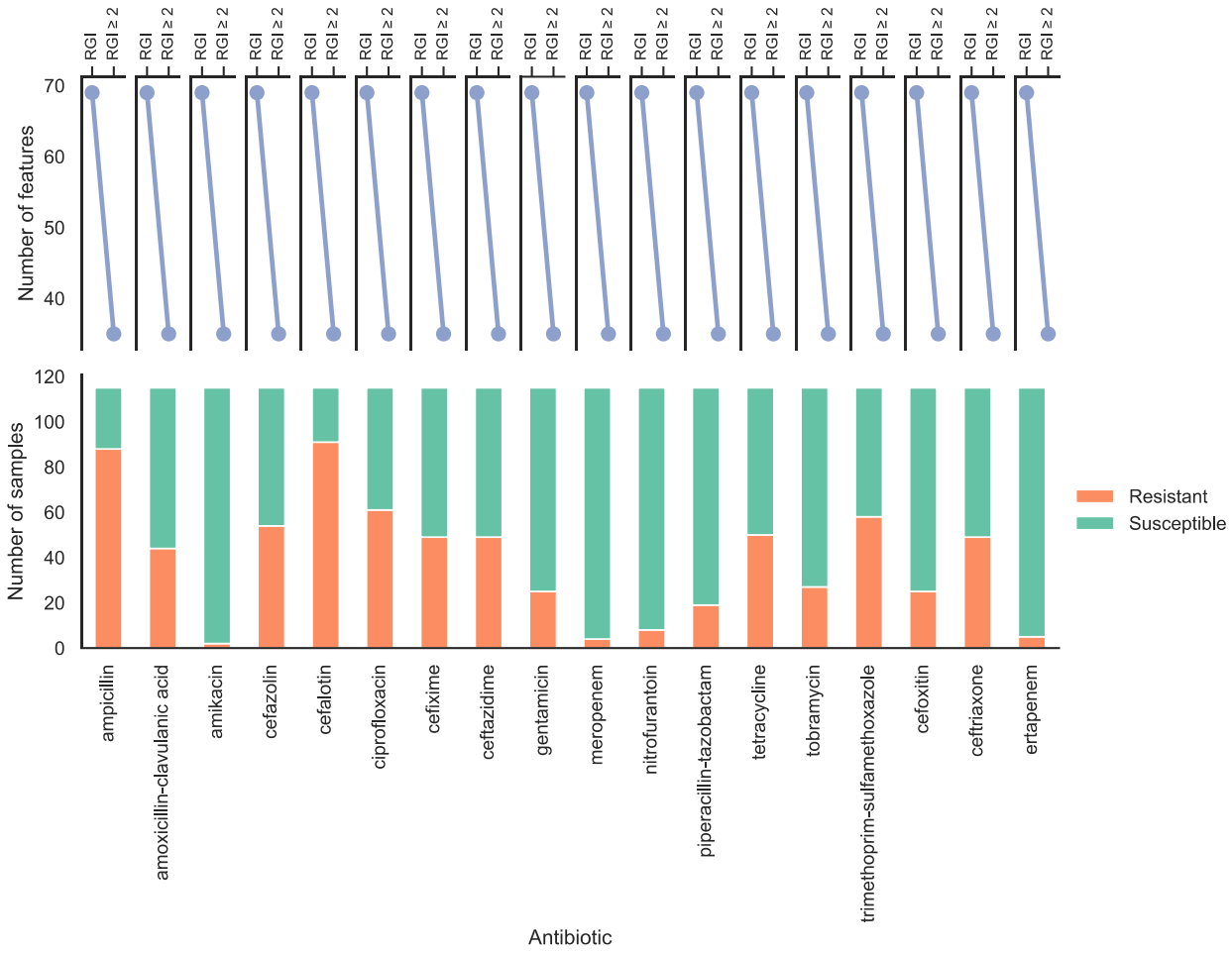
Supplementary Figure 3-4. Resistance determinant prediction and distribution AMR phenotypes in *P. aeruginosa* SPAdes assemblies (Dataset PA2). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



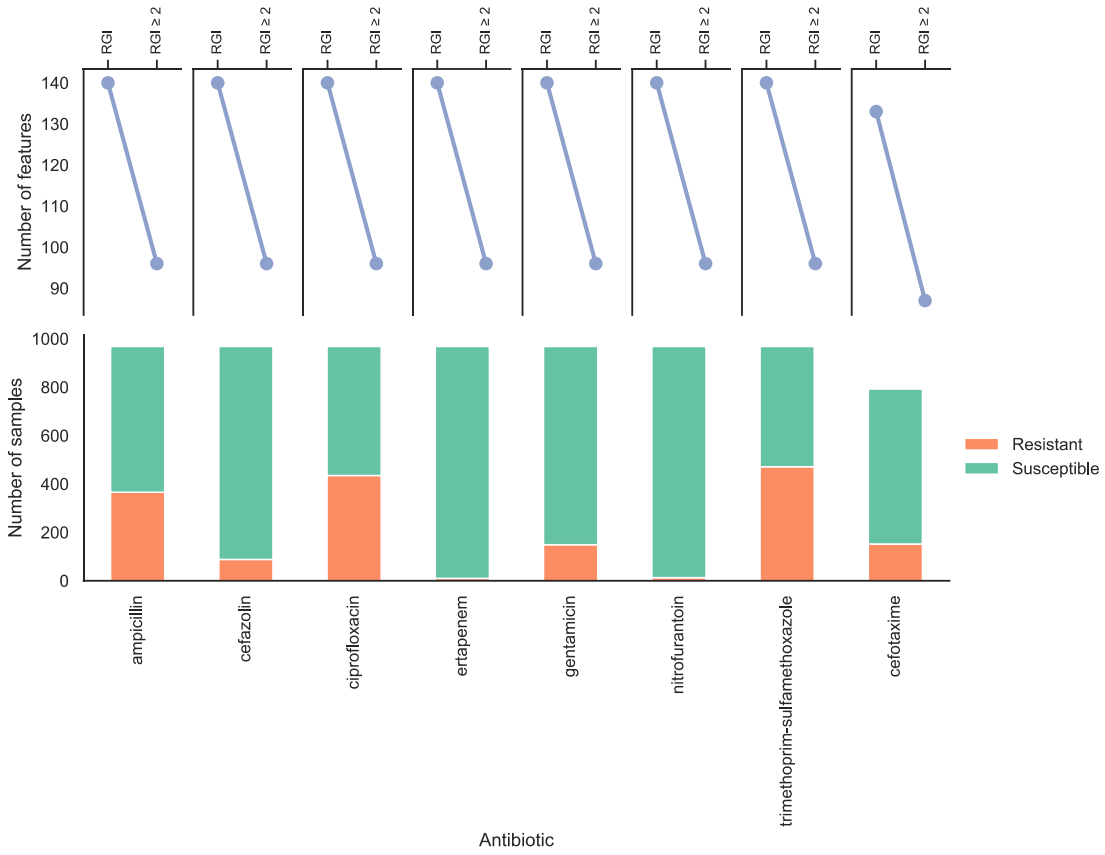
Supplementary Figure 3-5. Resistance determinant prediction and distribution AMR phenotypes in *N. gonorrhoeae* SPAdes assemblies (Dataset NG1). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



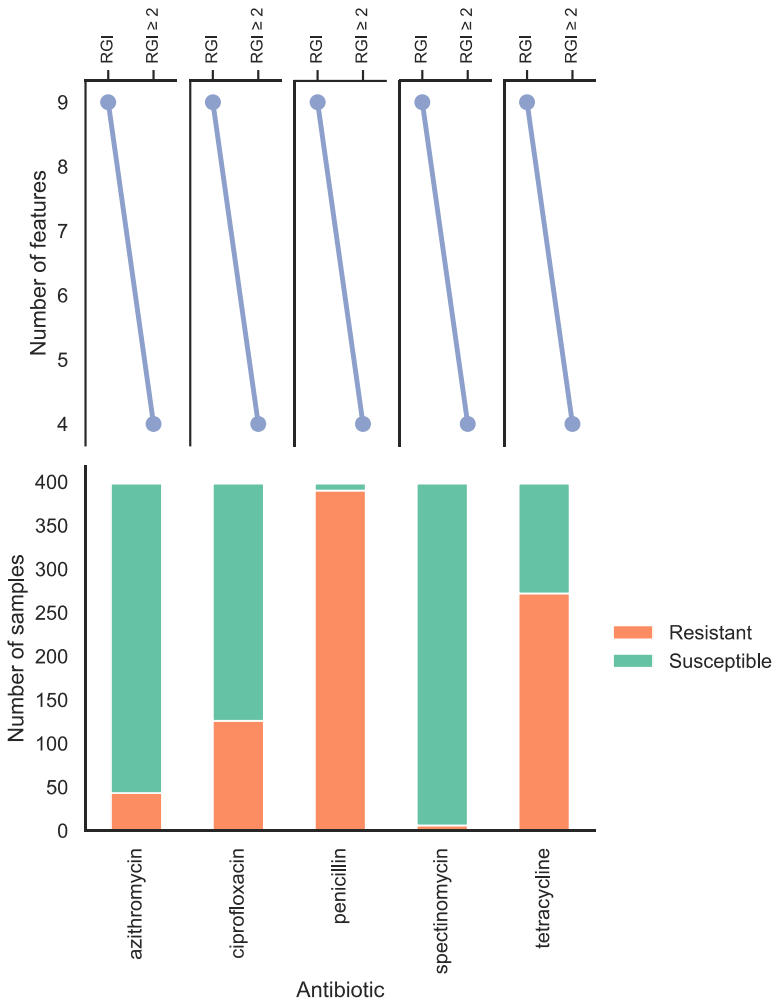
Supplementary Figure 3-6. Resistance determinant prediction and distribution AMR phenotypes in *N. gonorrhoeae* SPAdes assemblies (Dataset NG2). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



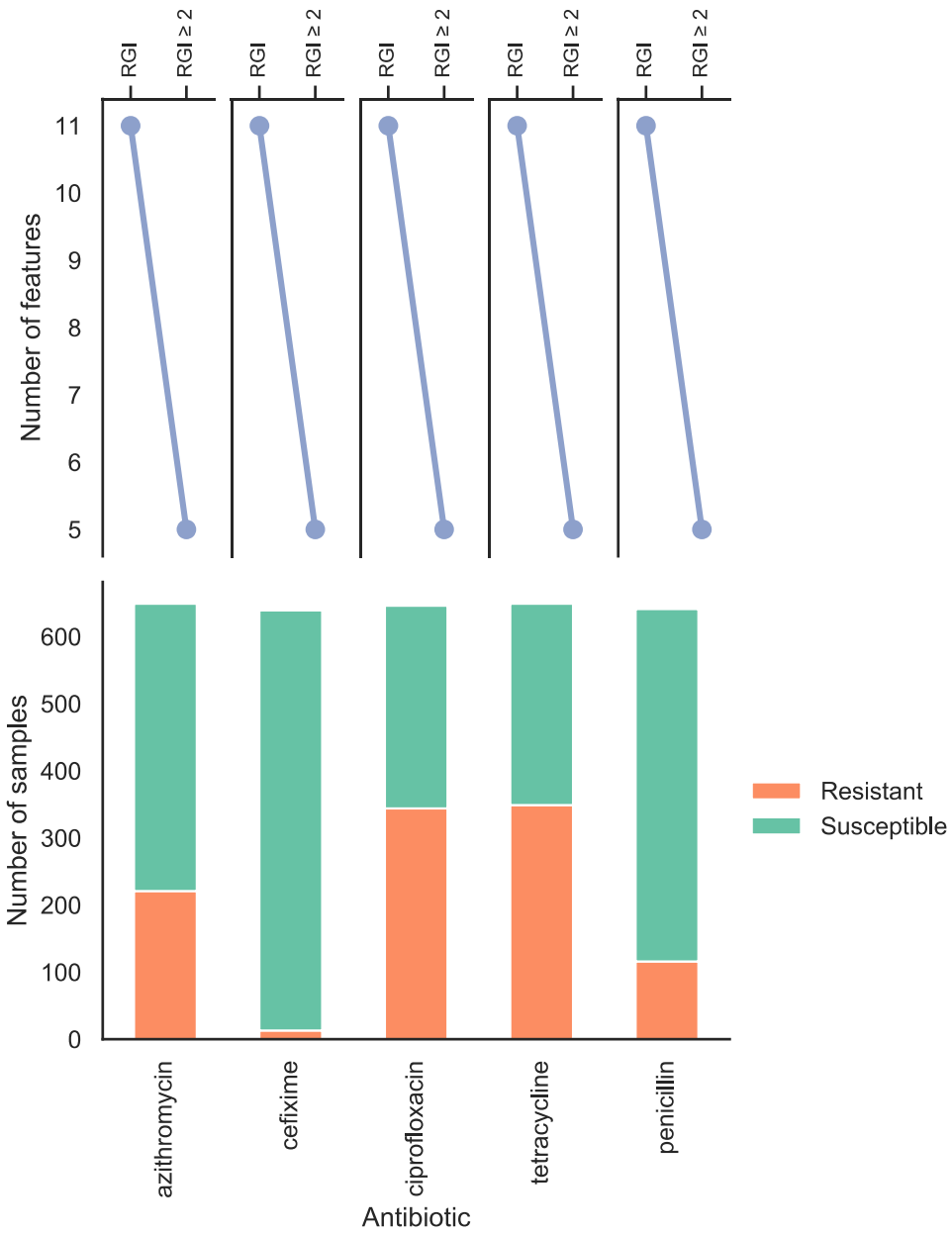
Supplementary Figure 3-7. Resistance determinant prediction and distribution AMR phenotypes in *E. coli* HyAsP assemblies (Dataset EC1). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



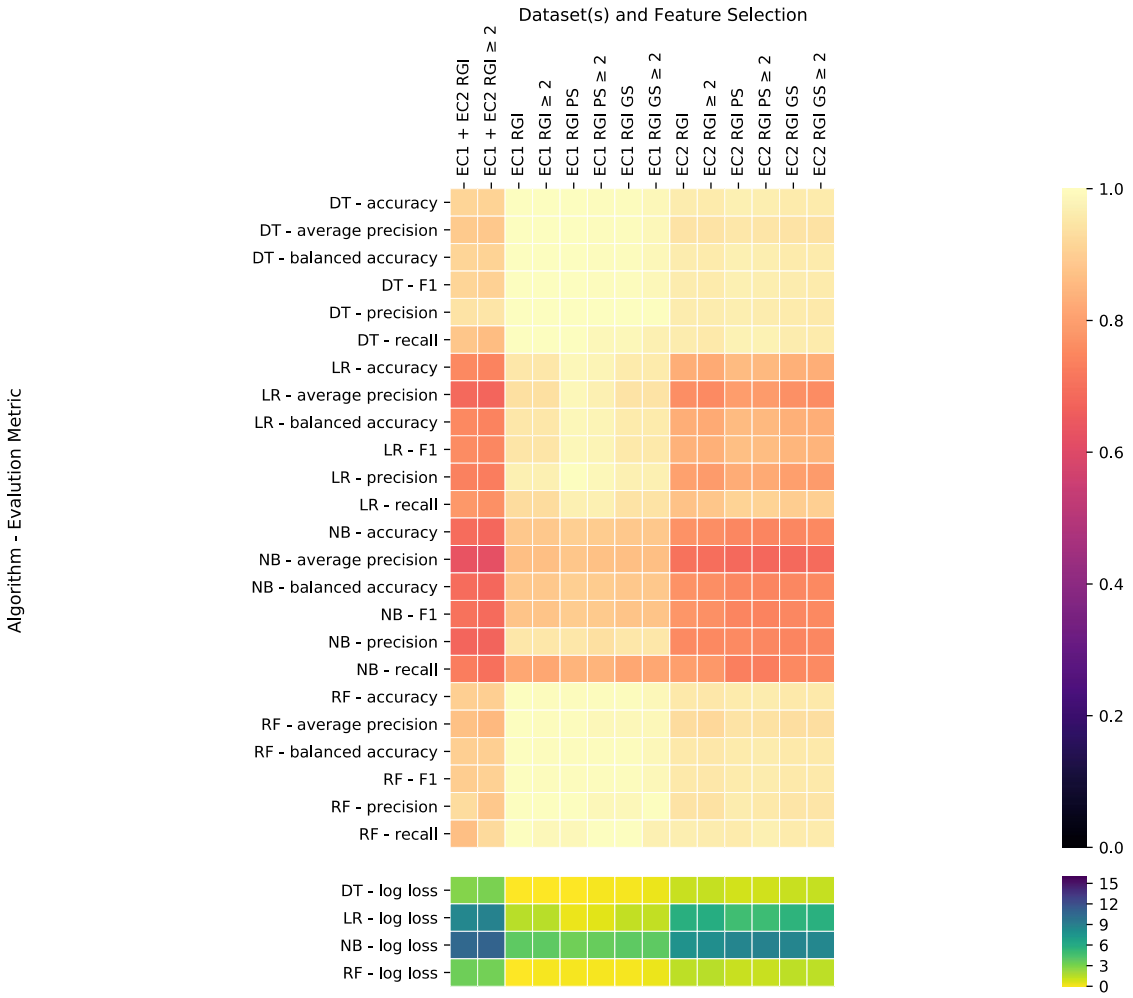
Supplementary Figure 3-8. Resistance determinant prediction and distribution AMR phenotypes in *E. coli* HyAsP assemblies (Dataset EC2). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



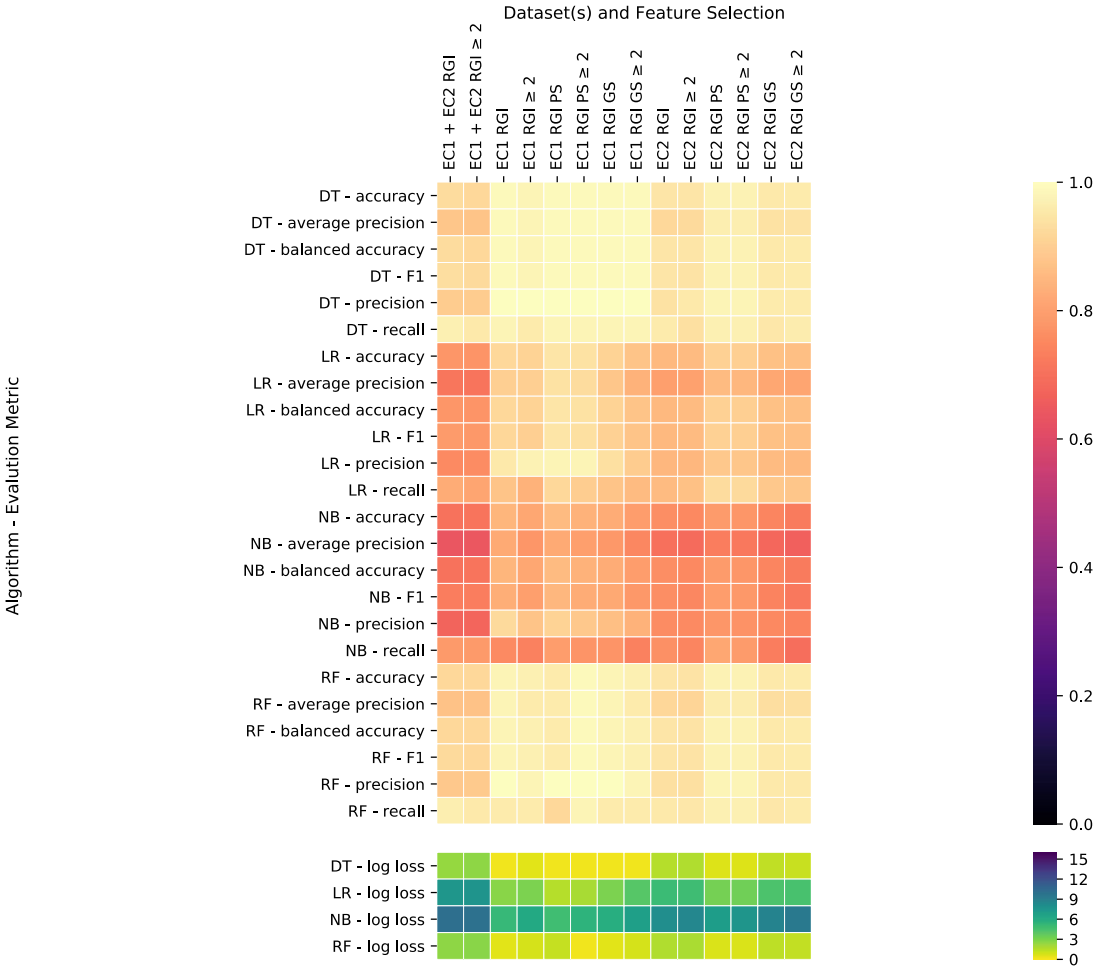
Supplementary Figure 3-9. Resistance determinant prediction and distribution AMR phenotypes in *N. gonorrhoeae* HyAsP assemblies (Dataset NG1). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



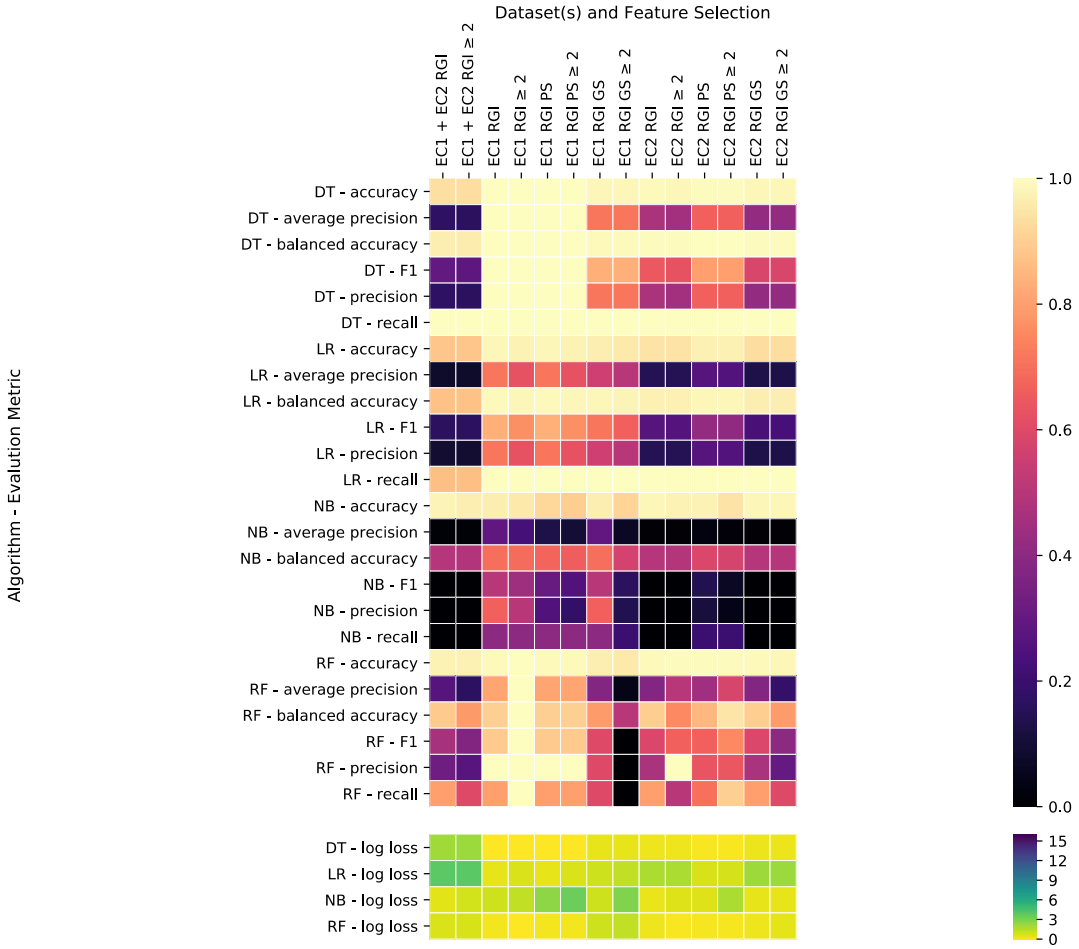
Supplementary Figure 3-10. Resistance determinant prediction and distribution AMR phenotypes in *N. gonorrhoeae* HyAsP assemblies (Dataset NG2). The top plots show the number of resistance determinants before (RGI) and after filtering for resistance determinants found in more than one sample (RGI ≥ 2).



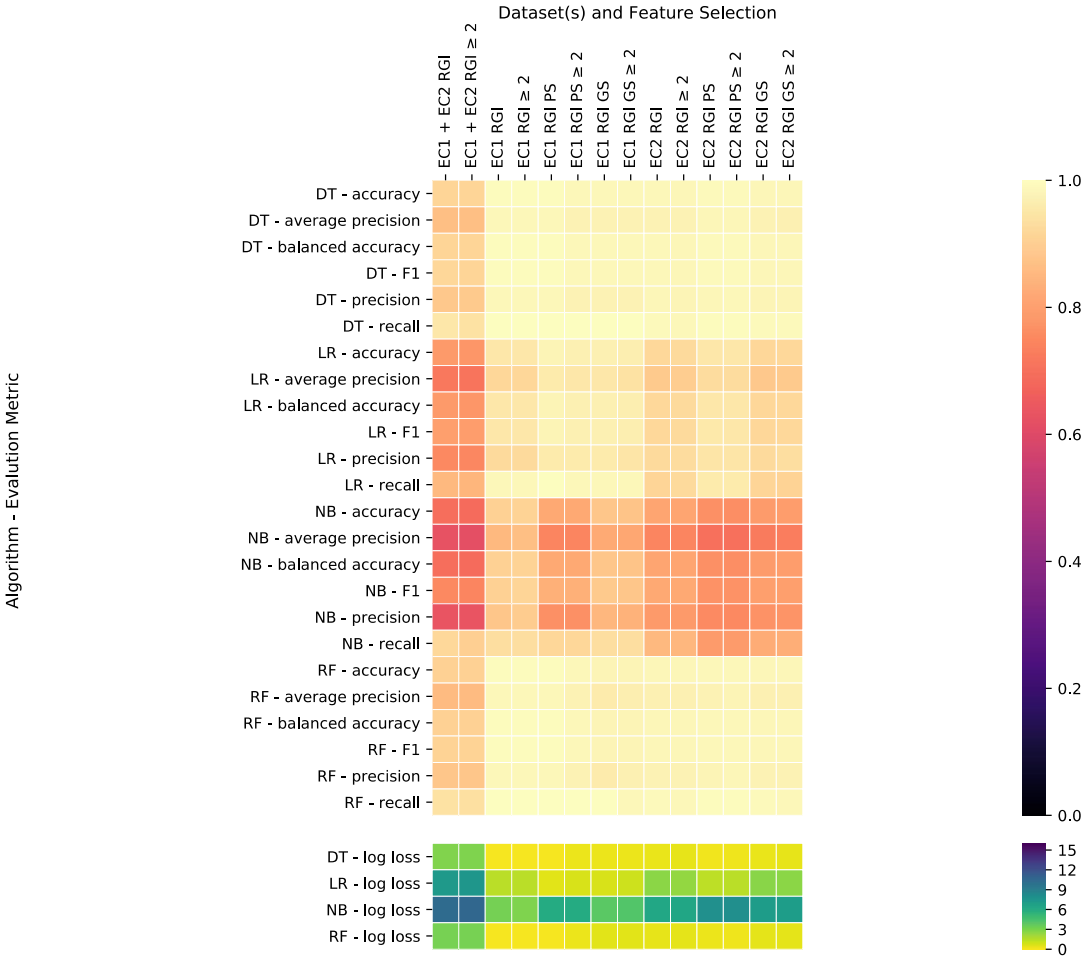
Supplementary Figure 3-11. *E. coli* ampicillin resistance prediction models using datasets EC1 and EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



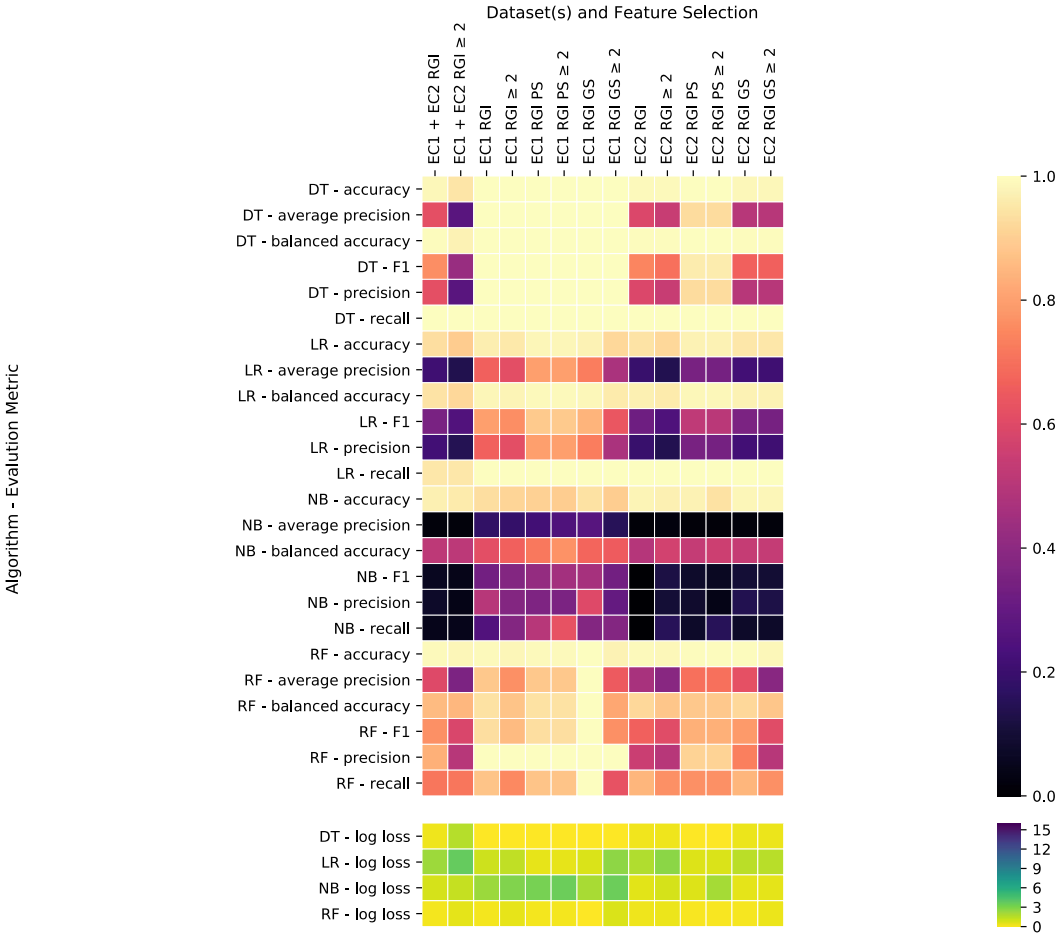
Supplementary Figure 3-12. *E. coli* cefazolin resistance prediction models using datasets EC1 and EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



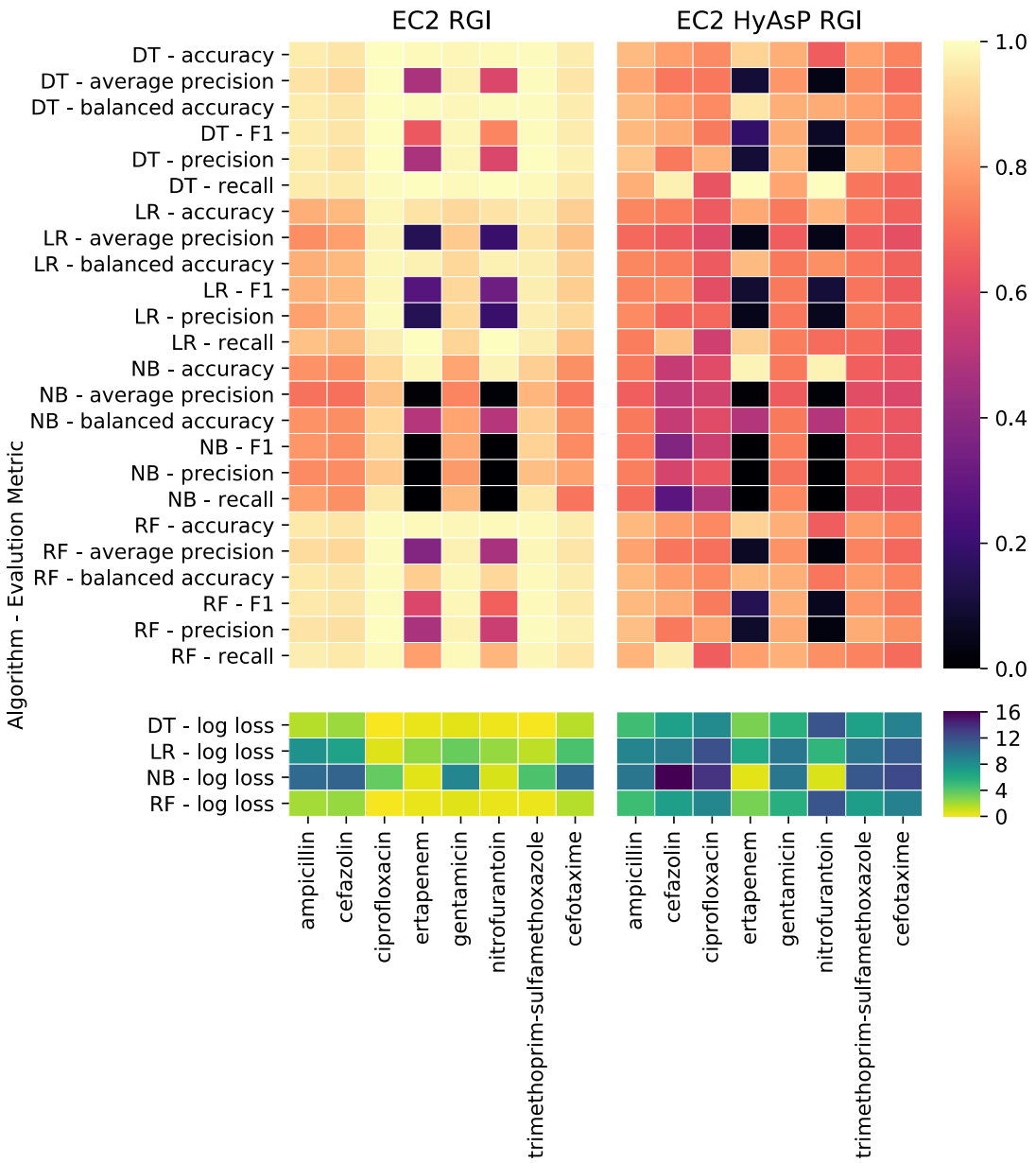
Supplementary Figure 3-13. *E. coli* ertapenem resistance prediction models using datasets EC1 and EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



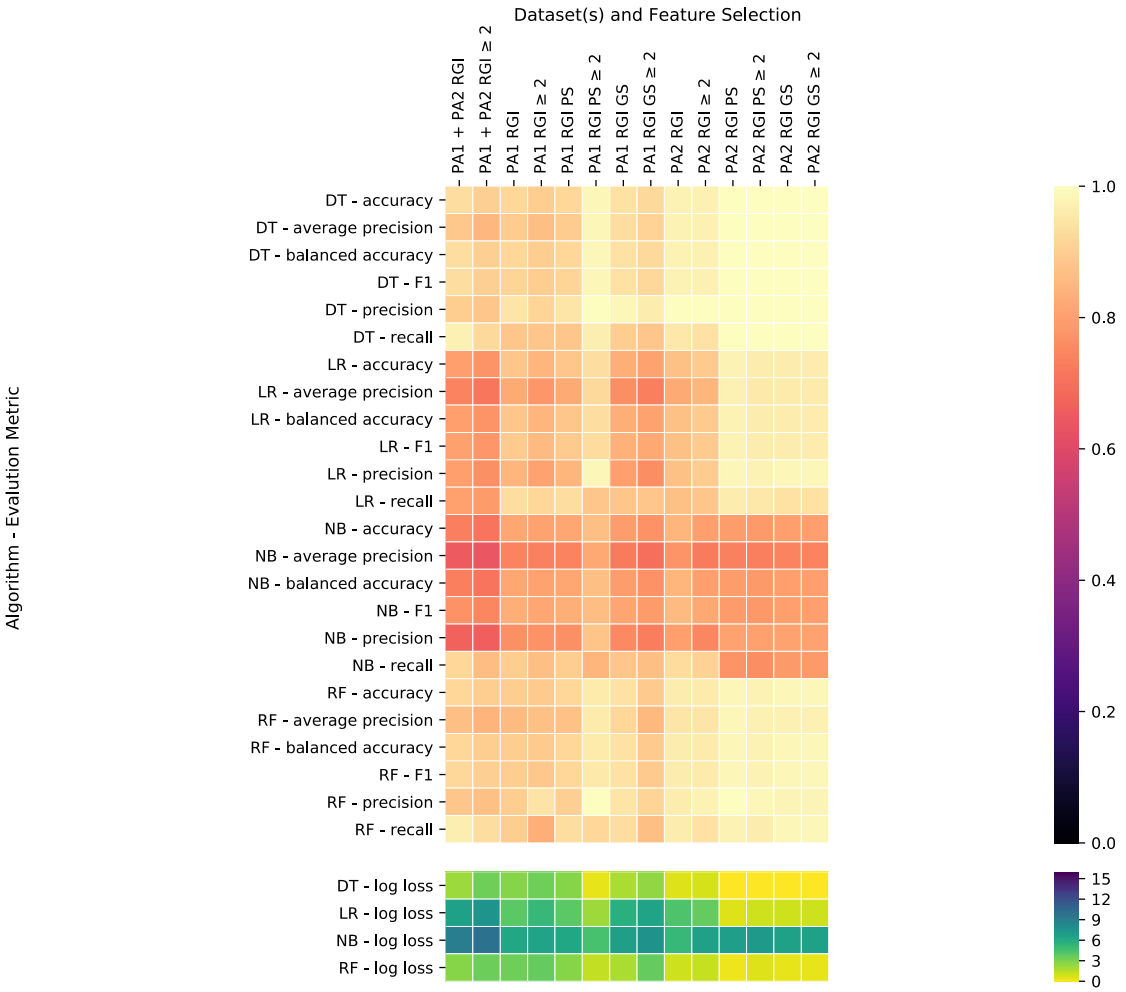
Supplementary Figure 3-14. *E. coli* gentamicin resistance prediction models using datasets EC1 and EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



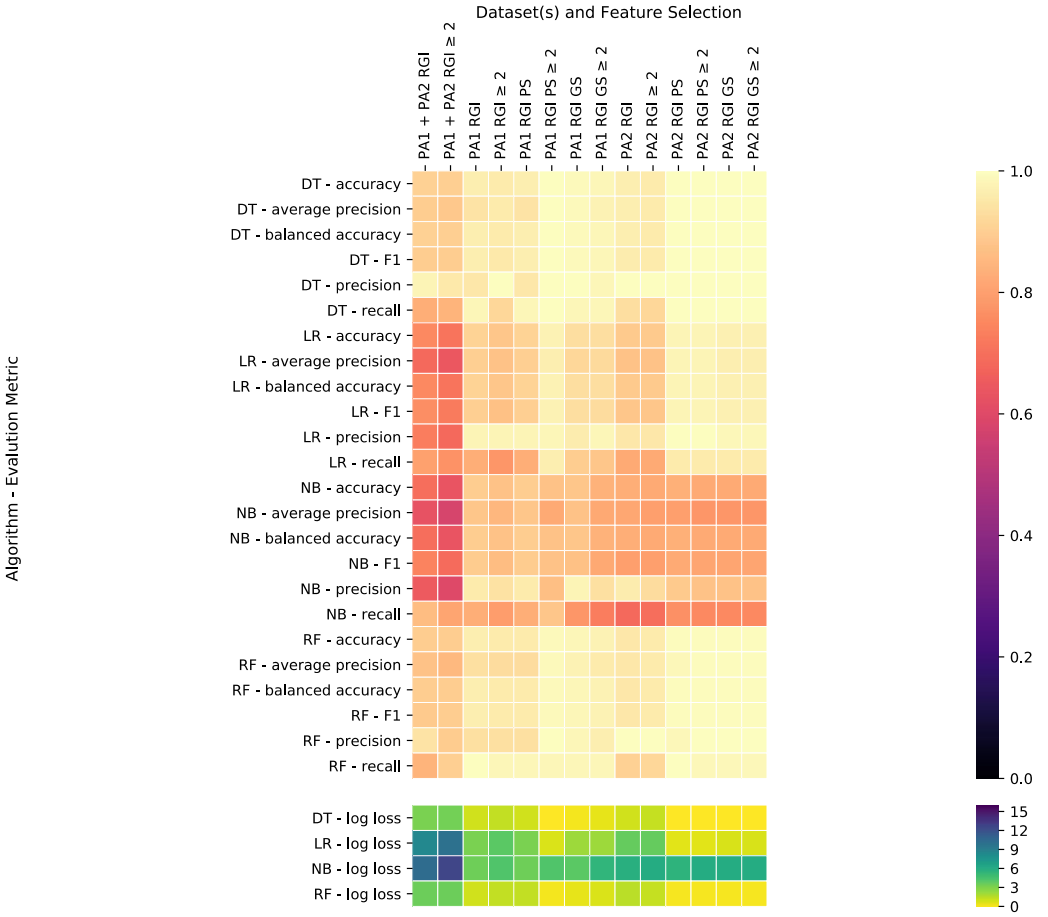
Supplementary Figure 3-15. *E. coli* nitrofurantoin resistance prediction models using datasets EC1 and EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



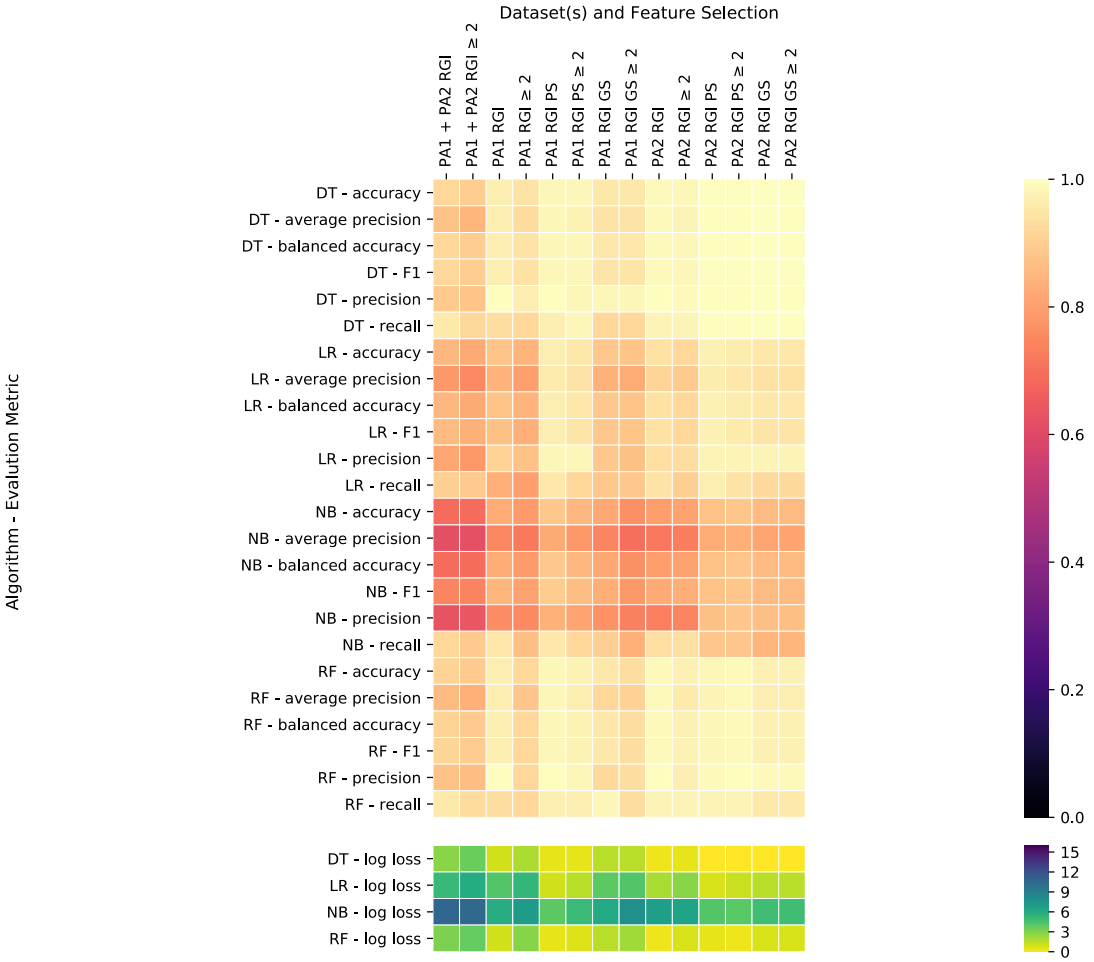
Supplementary Figure 3-16. *E. coli* AMR prediction models using SPAdes (chromosome + plasmid) or HyAsP (plasmid) assemblies in dataset EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



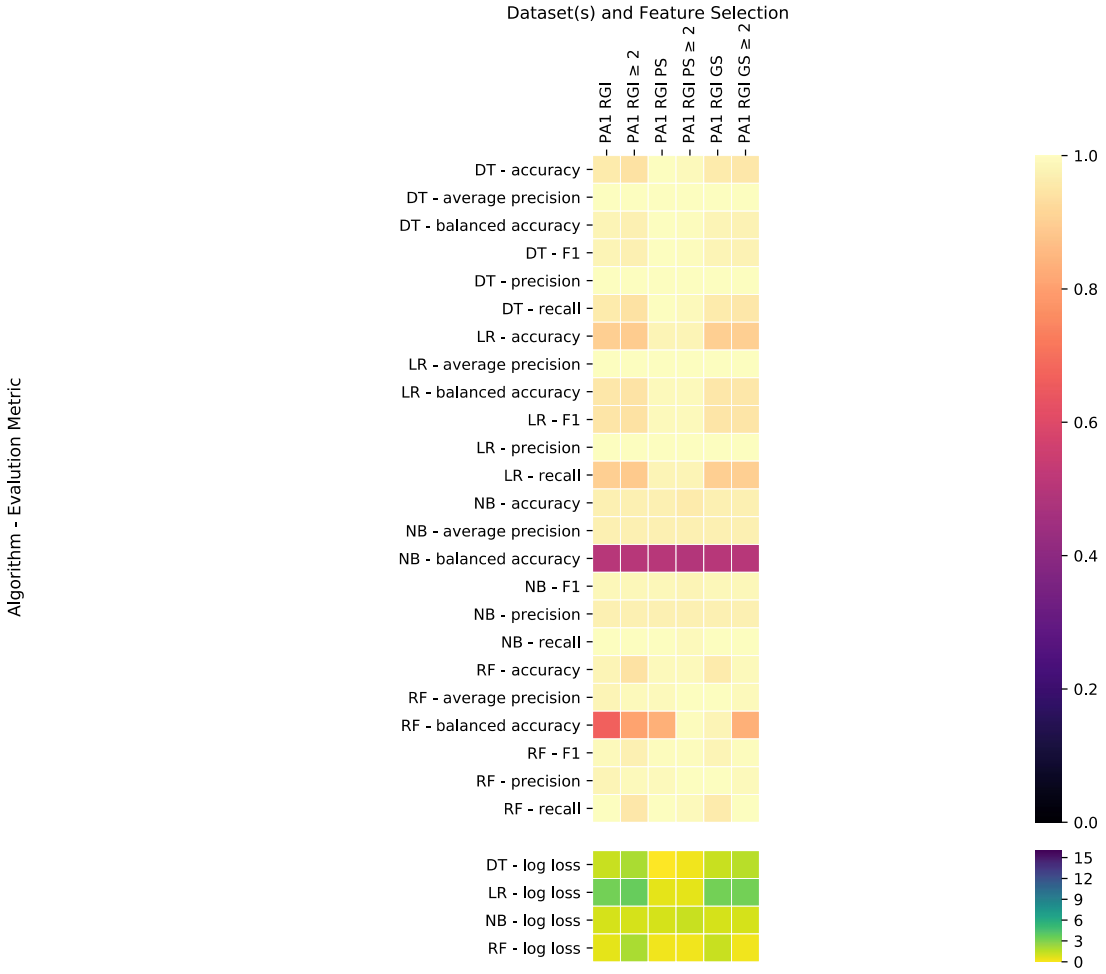
Supplementary Figure 3-17. *P. aeruginosa* ceftazidime resistance prediction models using datasets PA1 and PA2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



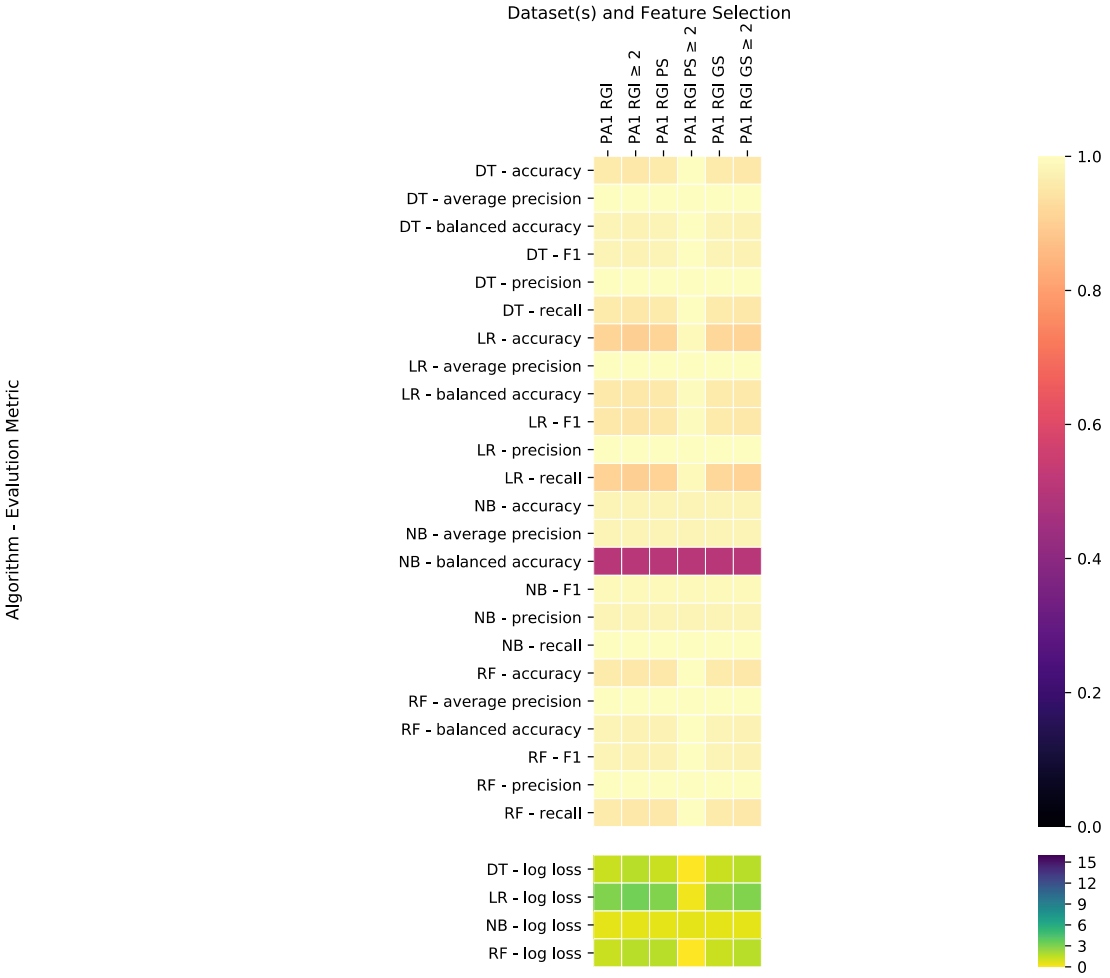
Supplementary Figure 3-18. *P. aeruginosa* ciprofloxacin resistance prediction models using datasets PA1 and PA2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



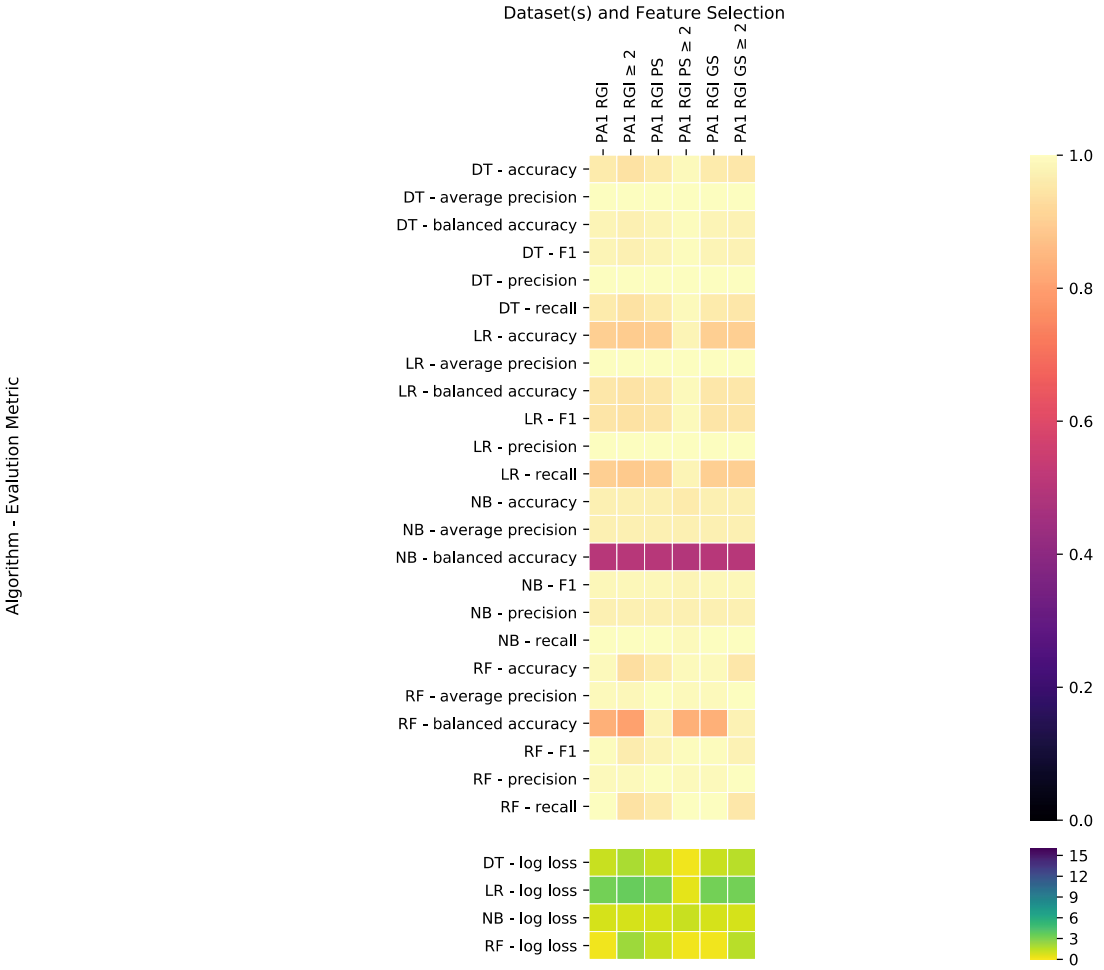
Supplementary Figure 3-19. *P. aeruginosa* piperacillin-tazobactam resistance prediction models using datasets PA1 and PA2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



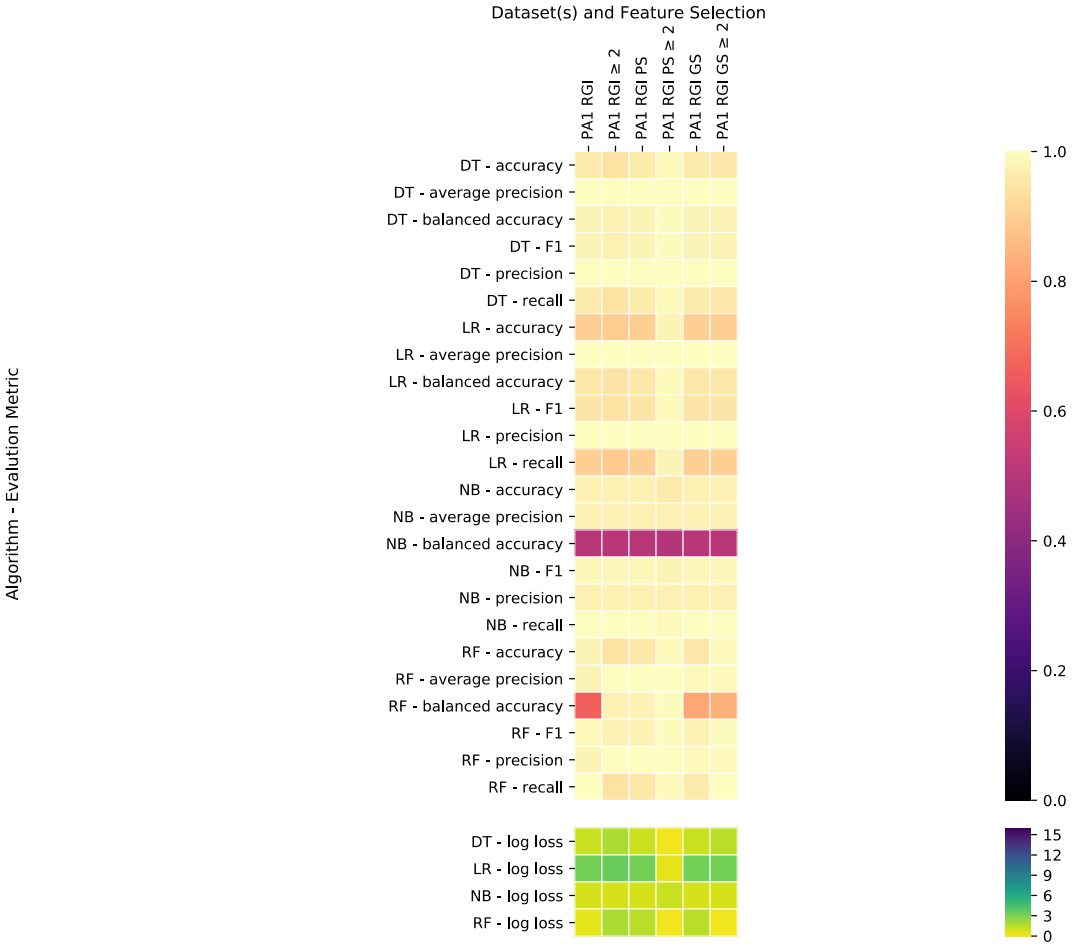
Supplementary Figure 3-20. AMR prediction models for *P. aeruginosa* amoxicillin-clavulanic acid resistance prediction models using dataset PA1. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



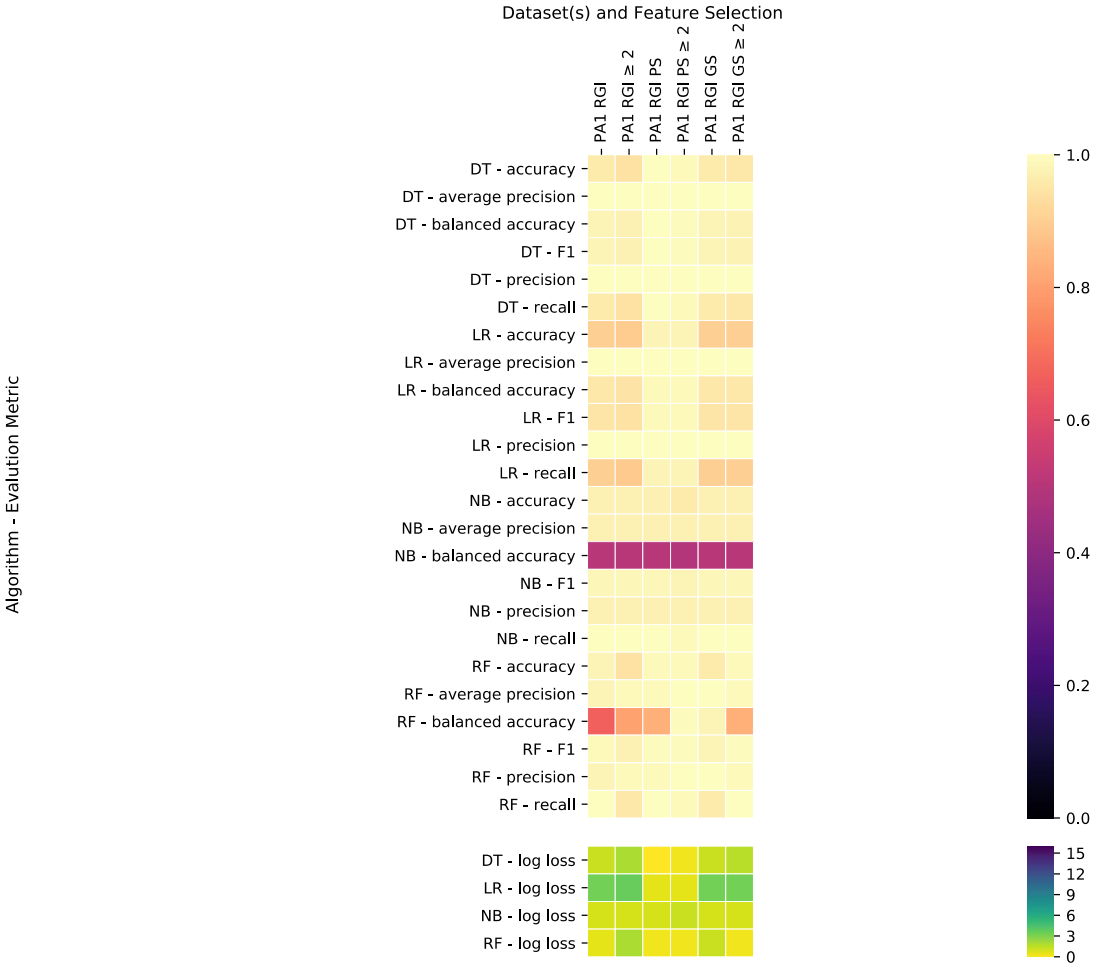
Supplementary Figure 3-21. *P. aeruginosa* cefixime resistance prediction models using dataset PA1. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



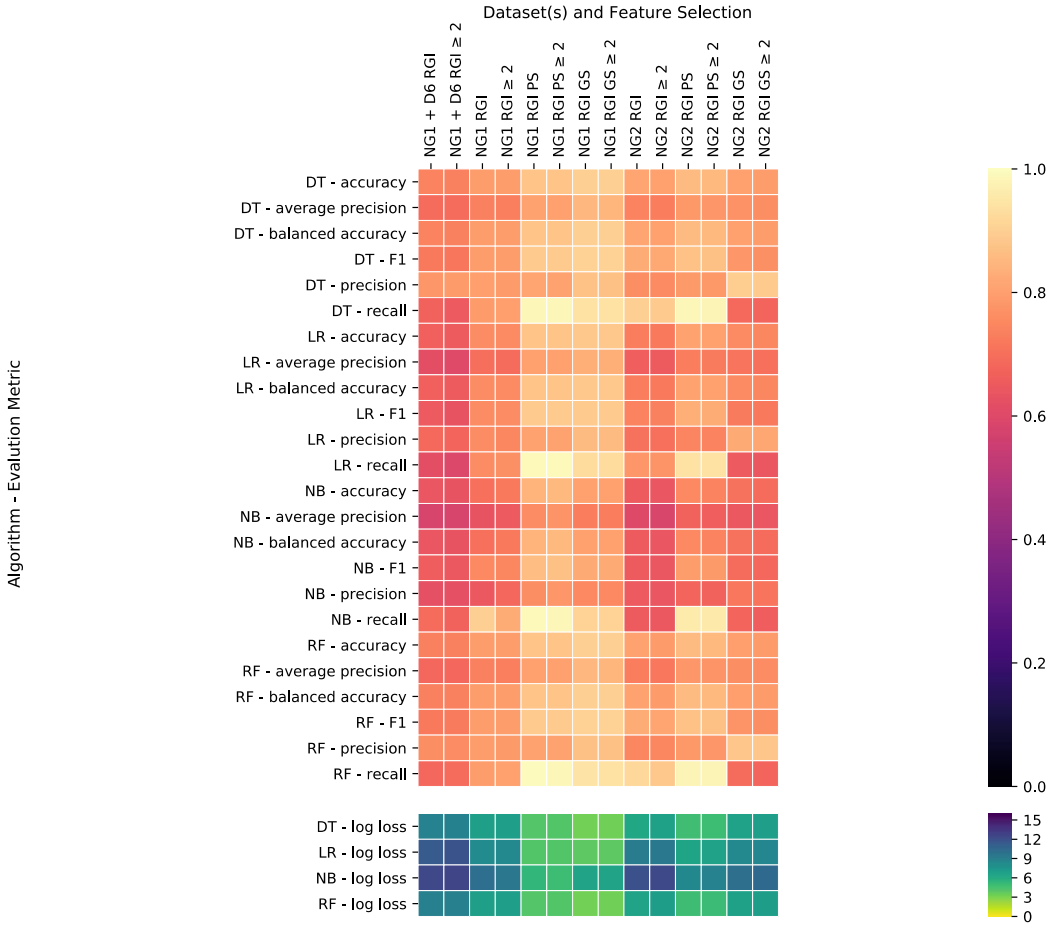
Supplementary Figure 3-22. *P. aeruginosa* cefoxitin resistance prediction models using dataset PA1. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



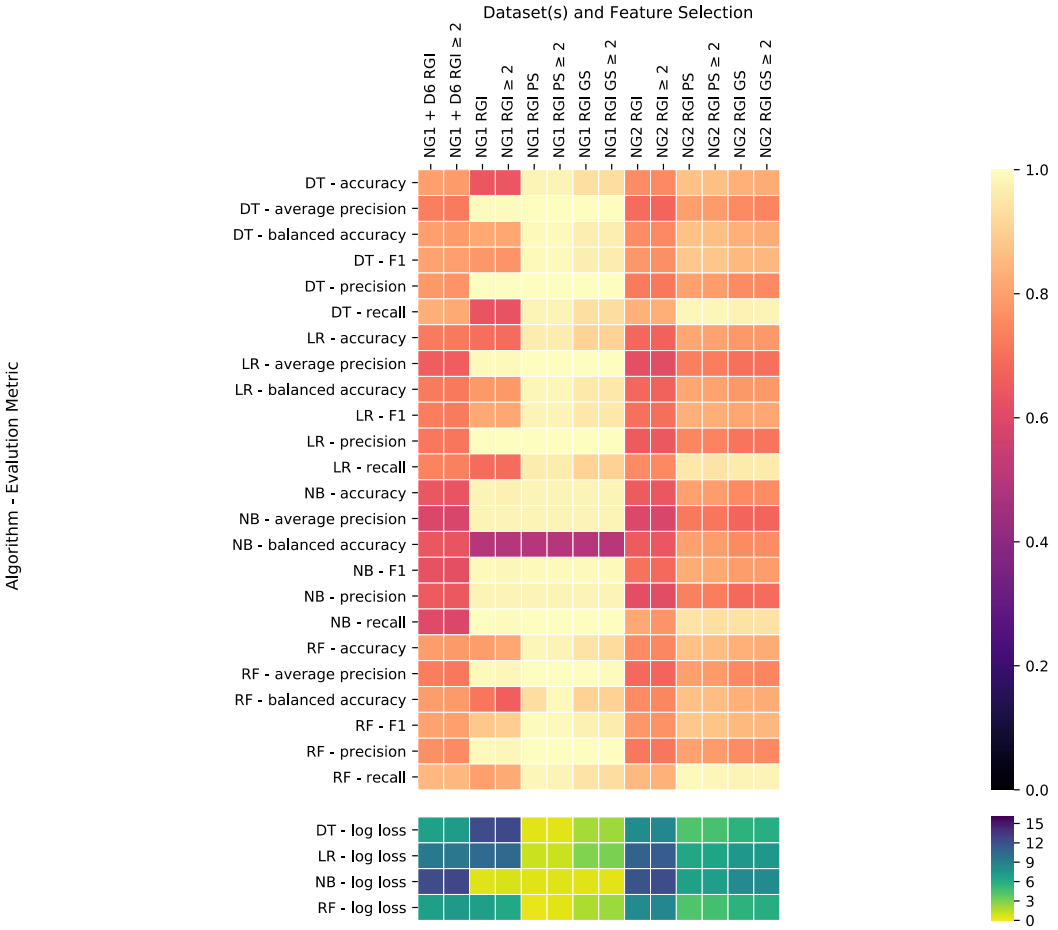
Supplementary Figure 3-23. *P. aeruginosa* ceftriaxone resistance prediction models using dataset PA1. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



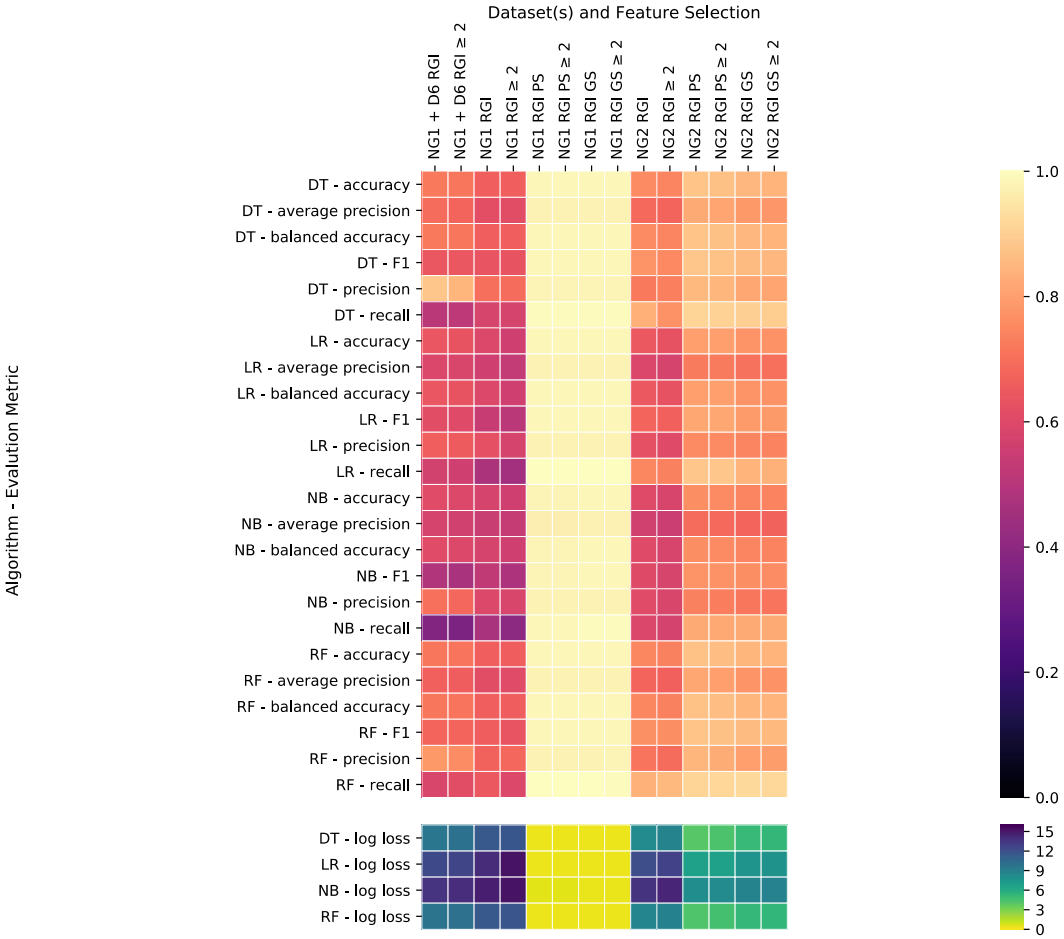
Supplementary Figure 3-24. *P. aeruginosa* trimethoprim-sulfamethoxazole resistance prediction models using dataset PA1. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



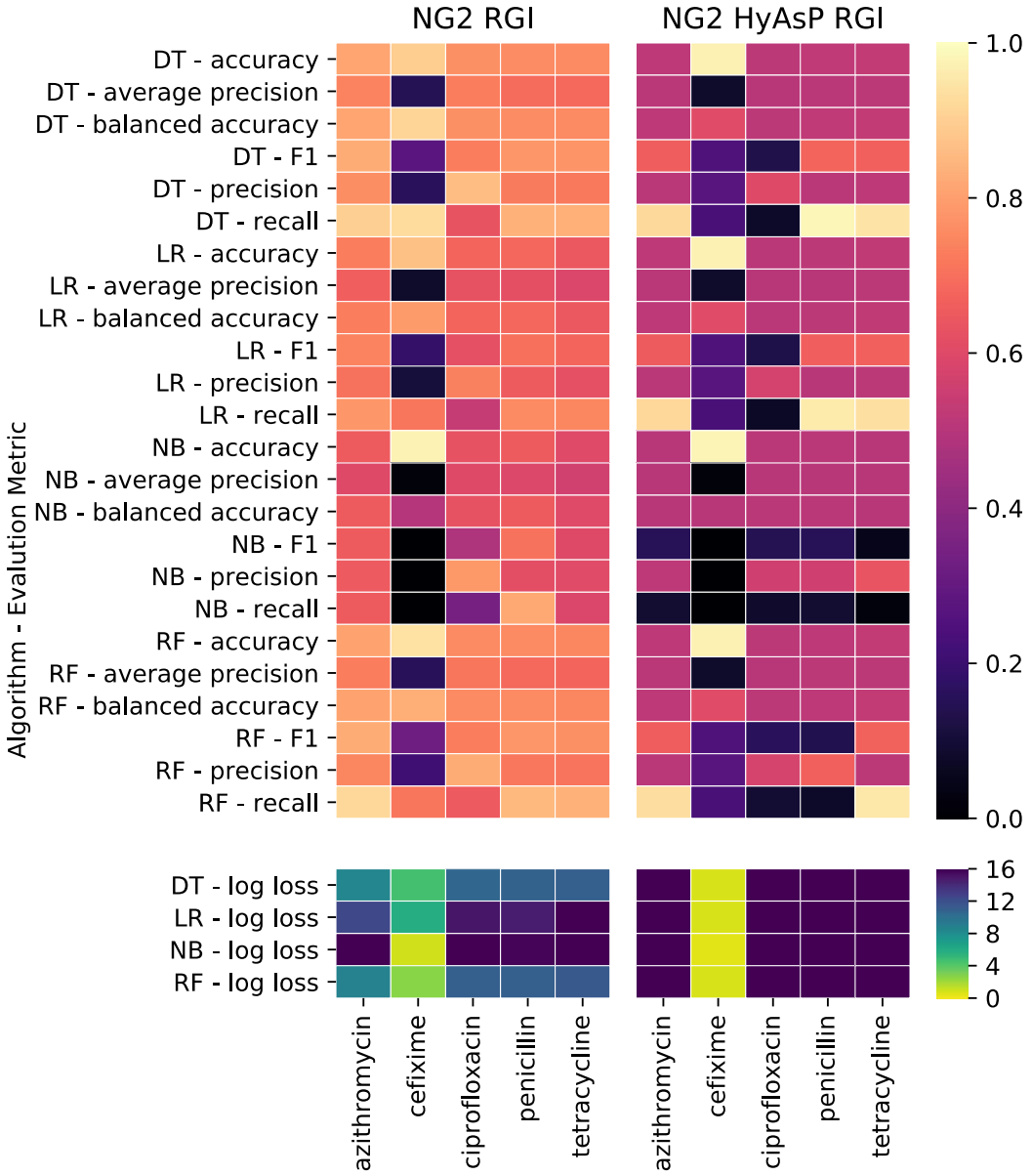
Supplementary Figure 3-25. *N. gonorrhoeae* azithromycin resistance prediction models using datasets NG1 and NG2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



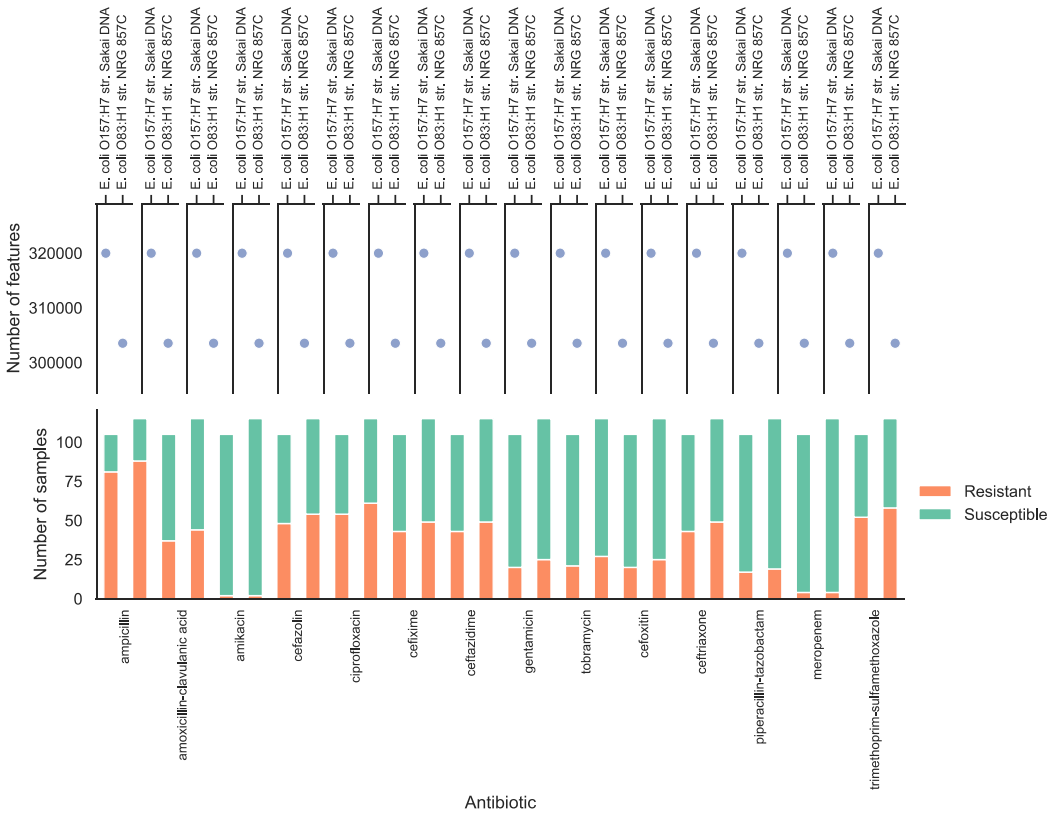
Supplementary Figure 3-26. *N. gonorrhoeae* penicillin resistance prediction models using datasets NG1 and NG2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



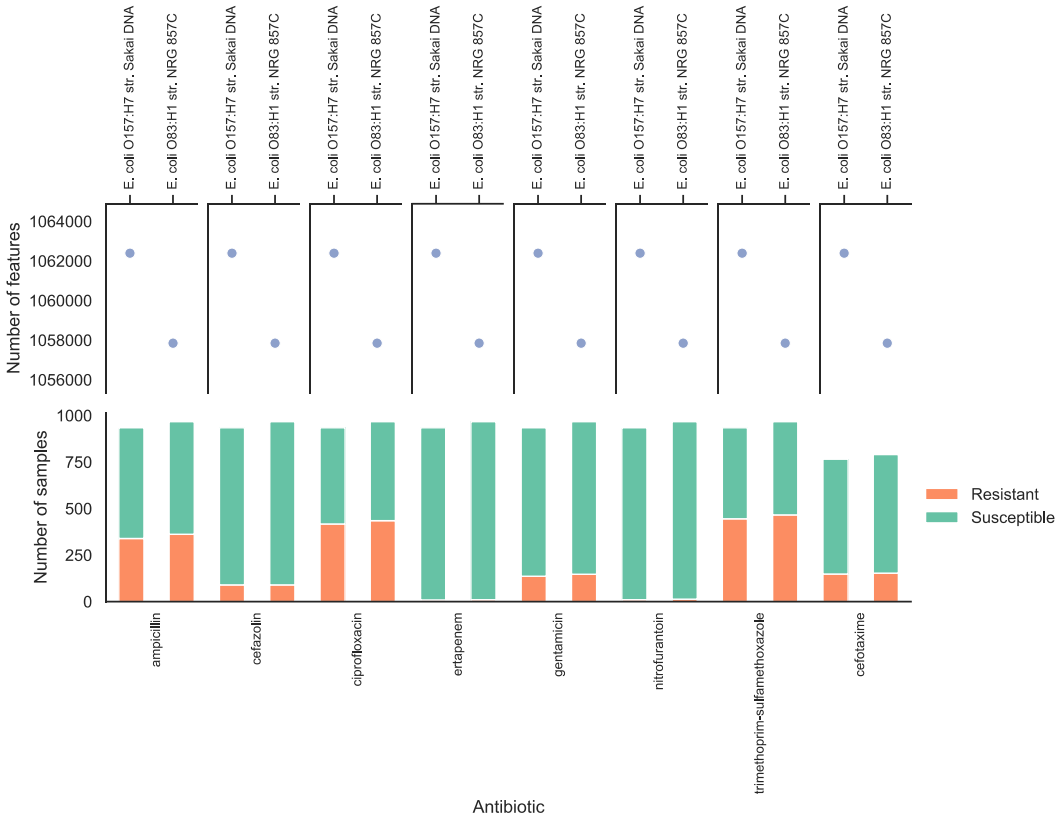
Supplementary Figure 3-27. *N. gonorrhoeae* tetracycline resistance prediction models using datasets NG1 and NG2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



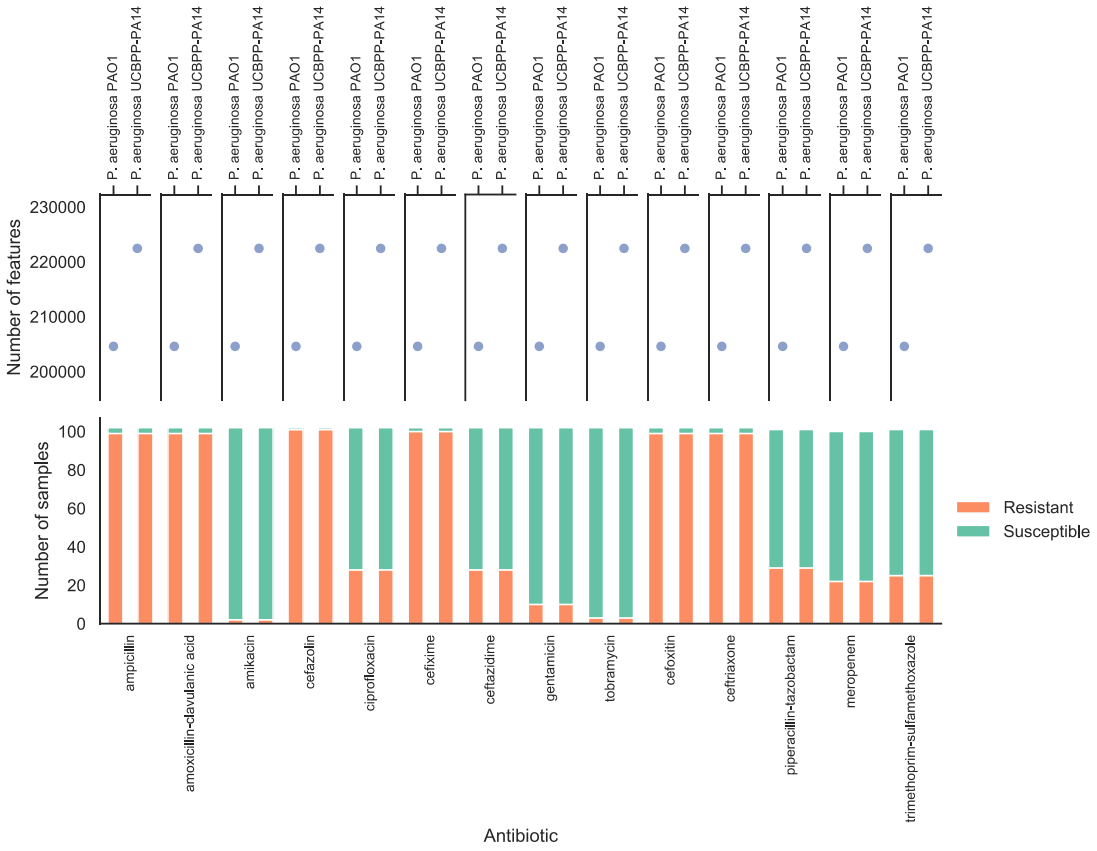
Supplementary Figure 3-28. *N. gonorrhoeae* resistance prediction models using SPAdes (chromosome + plasmid) or HyAsP (plasmid) assemblies in dataset NG2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants in ≥ 2 samples (i.e., PS), as in Figure 3-1.



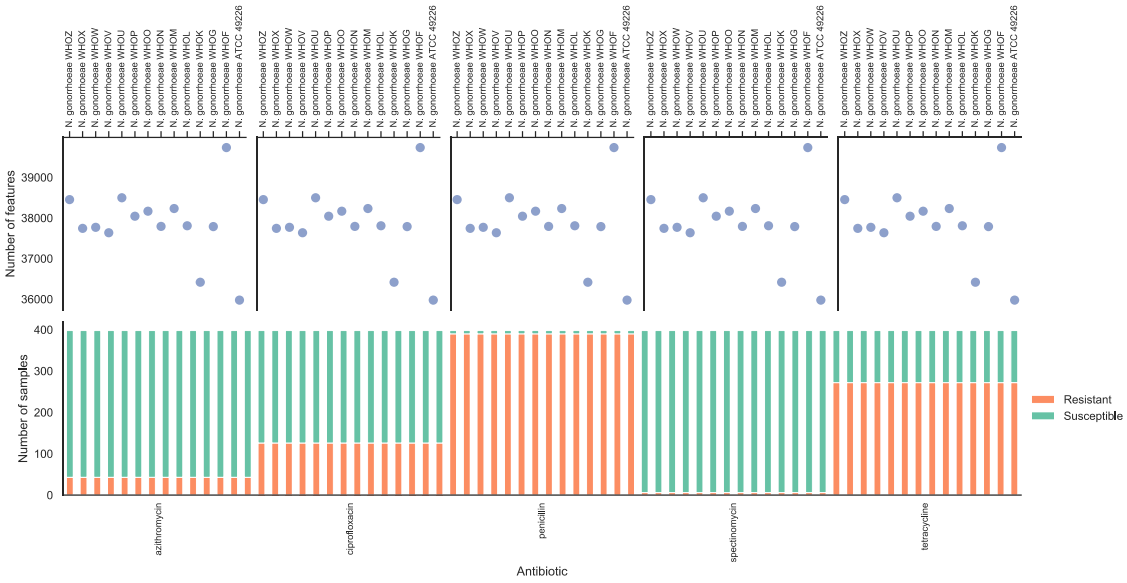
Supplementary Figure 3-29. Mutation generation and AMR phenotype distribution in *E. coli* dataset EC1. Two *E. coli* reference sequences were used.



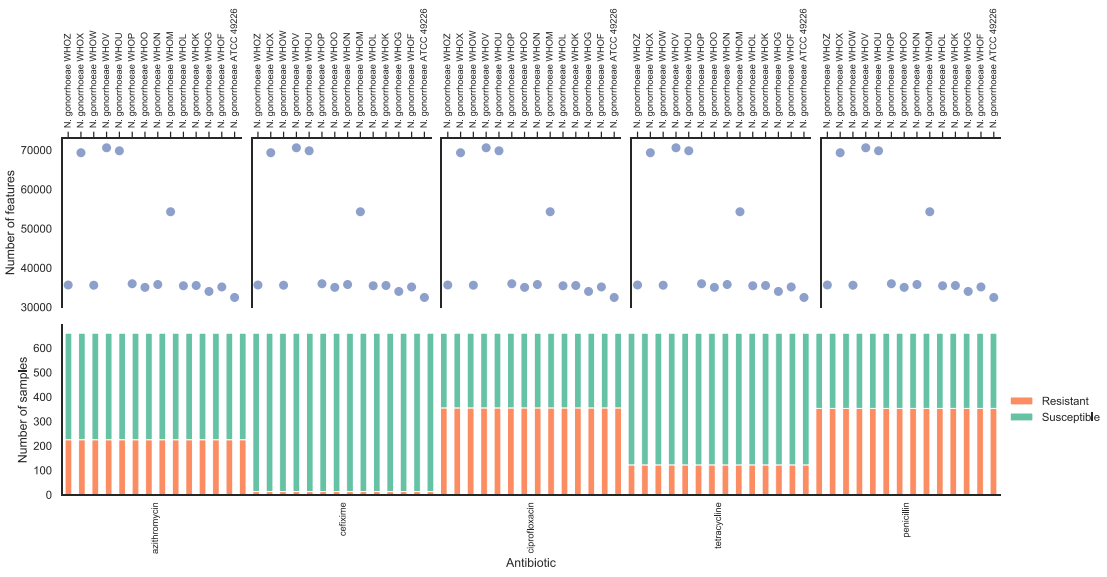
Supplementary Figure 3-30. Mutation generation and AMR phenotype distribution in *E. coli* dataset EC2. Two *E. coli* reference sequences were used.



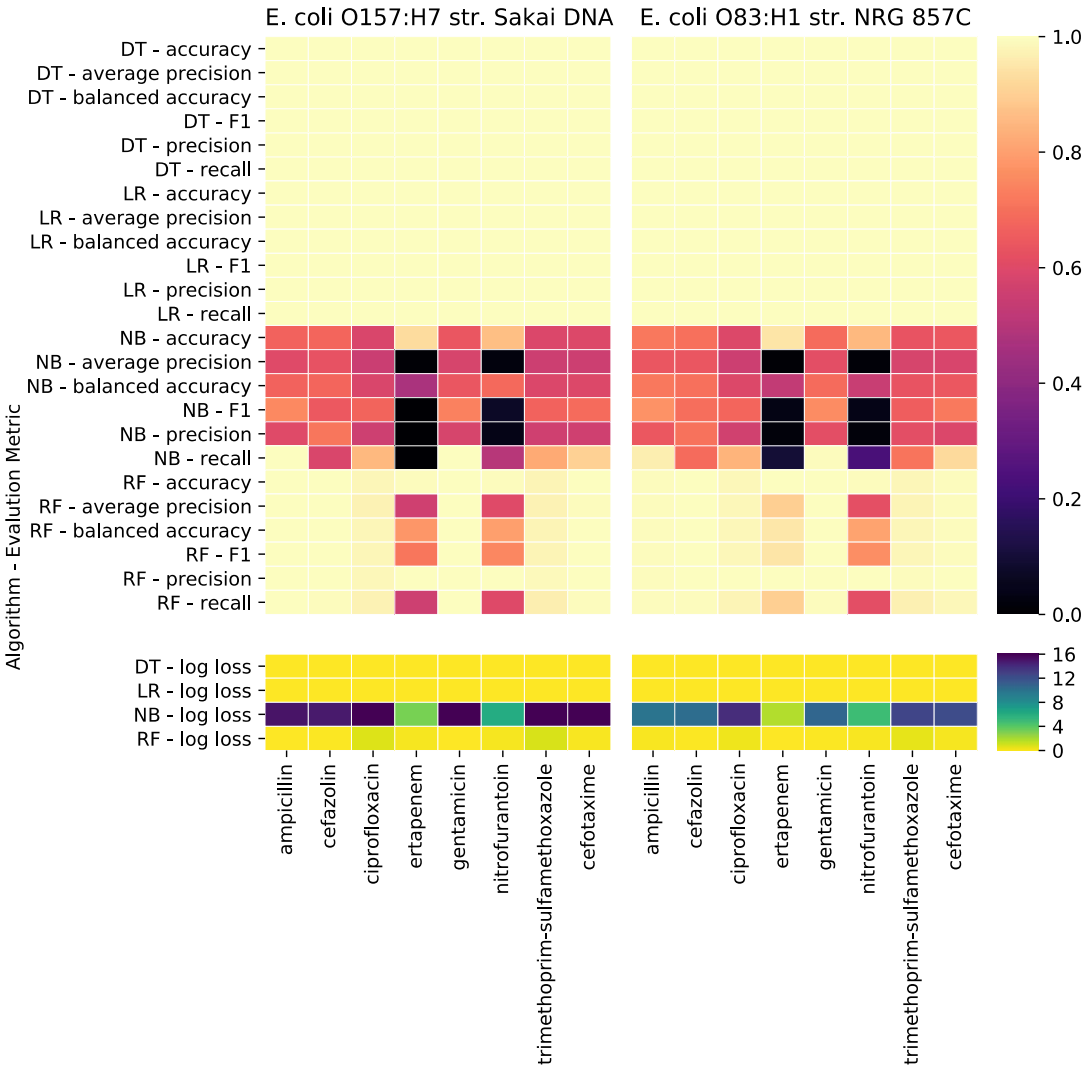
Supplementary Figure 3-31. Mutation generation and AMR phenotype distribution in *P. aeruginosa* dataset PA1. Two *P. aeruginosa* reference sequences were used.



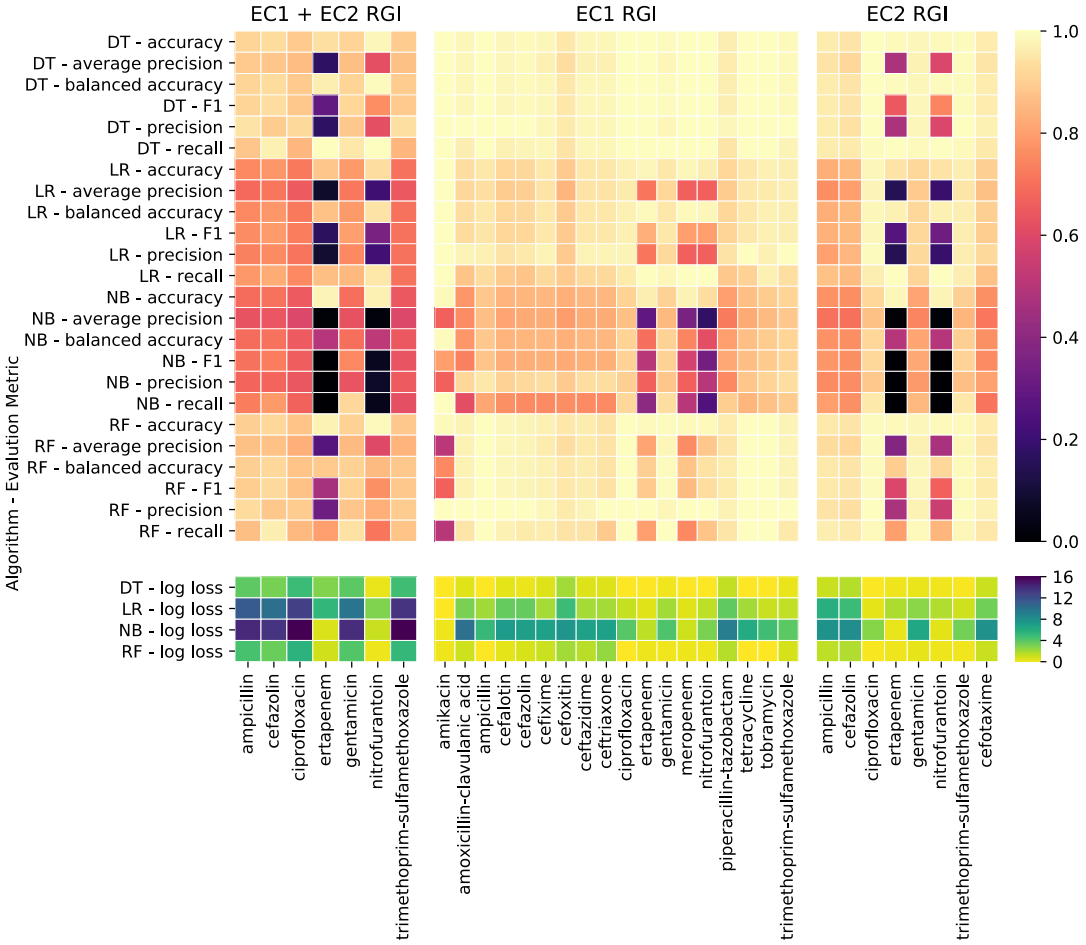
Supplementary Figure 3-32. Mutation generation and AMR phenotype distribution in *N. gonorrhoeae* dataset NG1. Fifteen *N. gonorrhoeae* reference sequences were used.



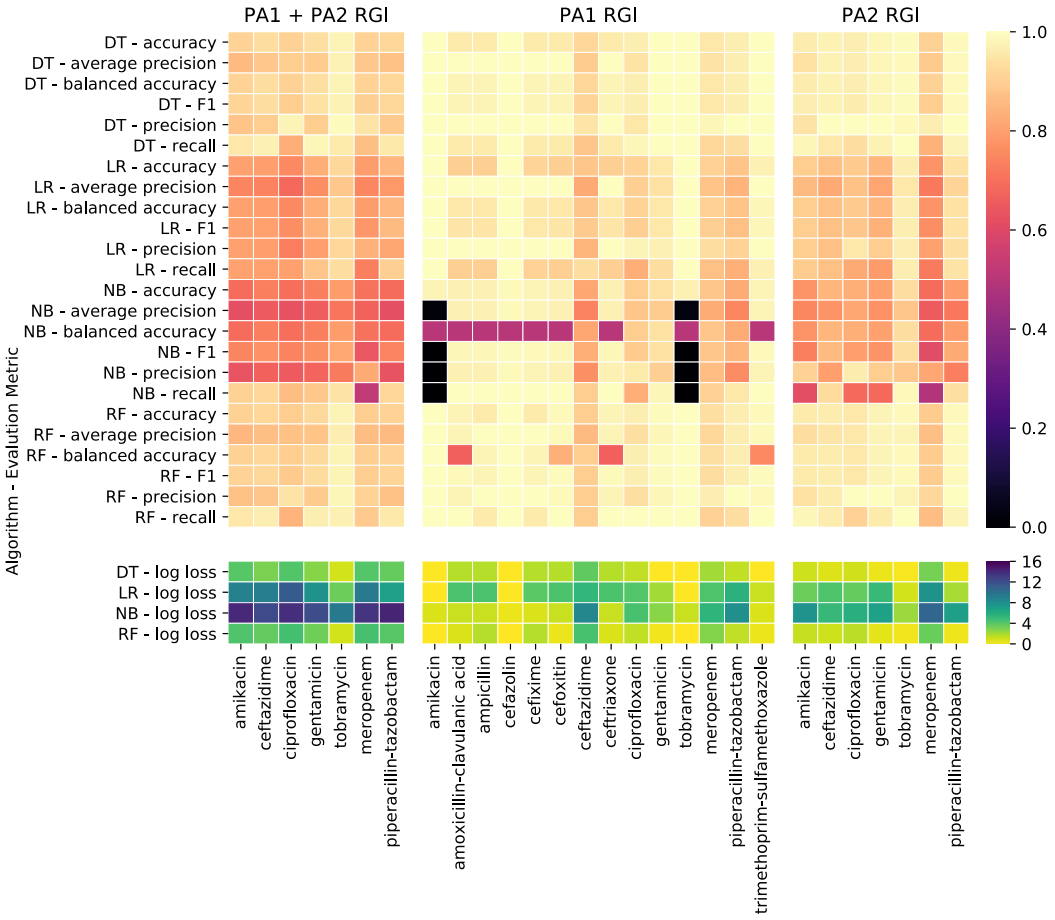
Supplementary Figure 3-33. Mutation generation and AMR phenotype distribution in *N. gonorrhoeae* dataset NG2. Fifteen *N. gonorrhoeae* reference sequences were used.



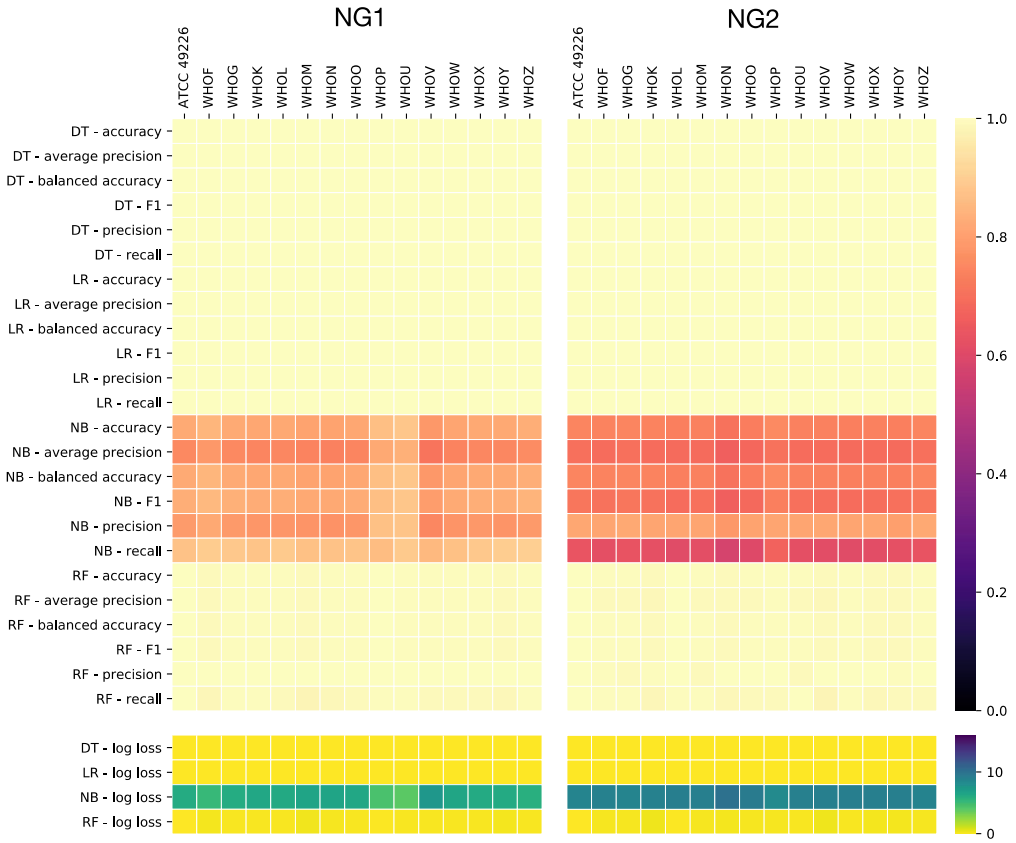
Supplementary Figure 3-34. *E. coli* AMR prediction models using mutations in dataset EC2. Two *E. coli* references were used. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.



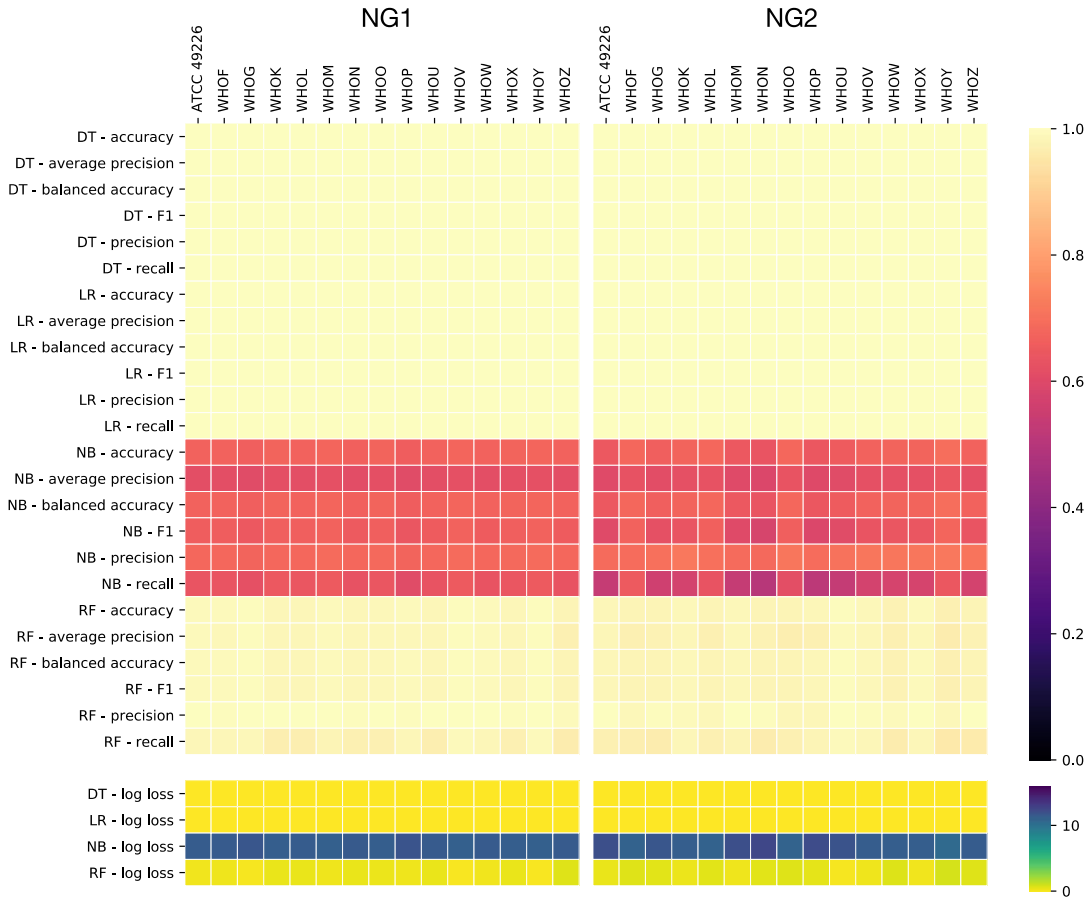
Supplementary Figure 3-35. *E. coli* AMR prediction models using known resistance determinants and no filtering with datasets EC1 and EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.



Supplementary Figure 3-36. *P. aeruginosa* AMR prediction models using known resistance determinants and no filtering with datasets PA1 and PA2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

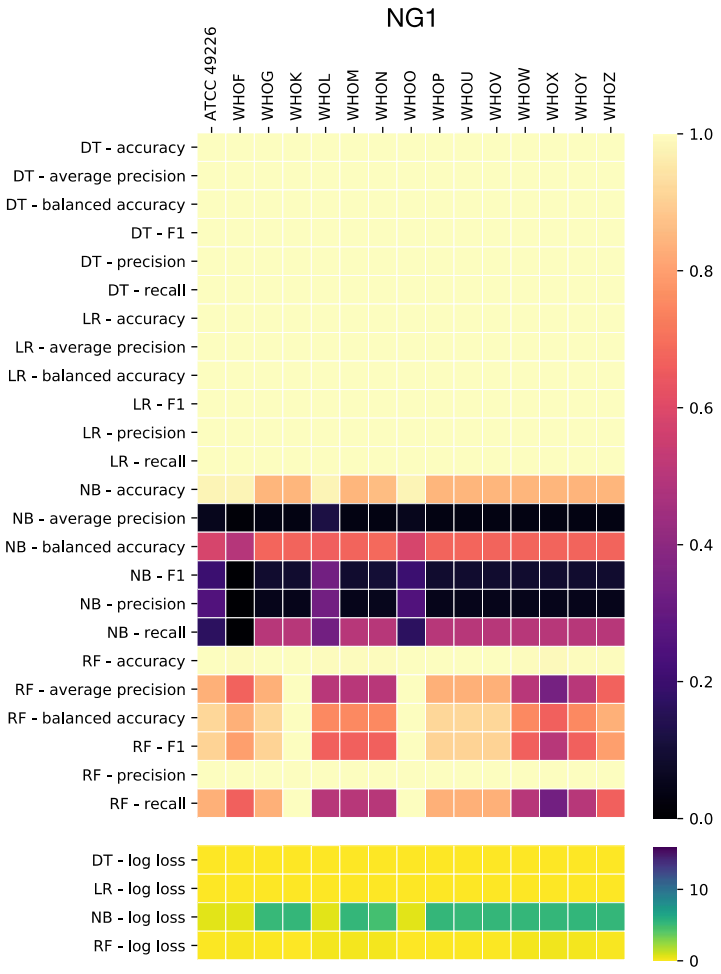


Supplementary Figure 3-37. *N. gonorrhoeae* azithromycin resistance prediction models using mutations in datasets NG1 and NG2. Fifteen *N. gonorrhoeae* reference sequences were used. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

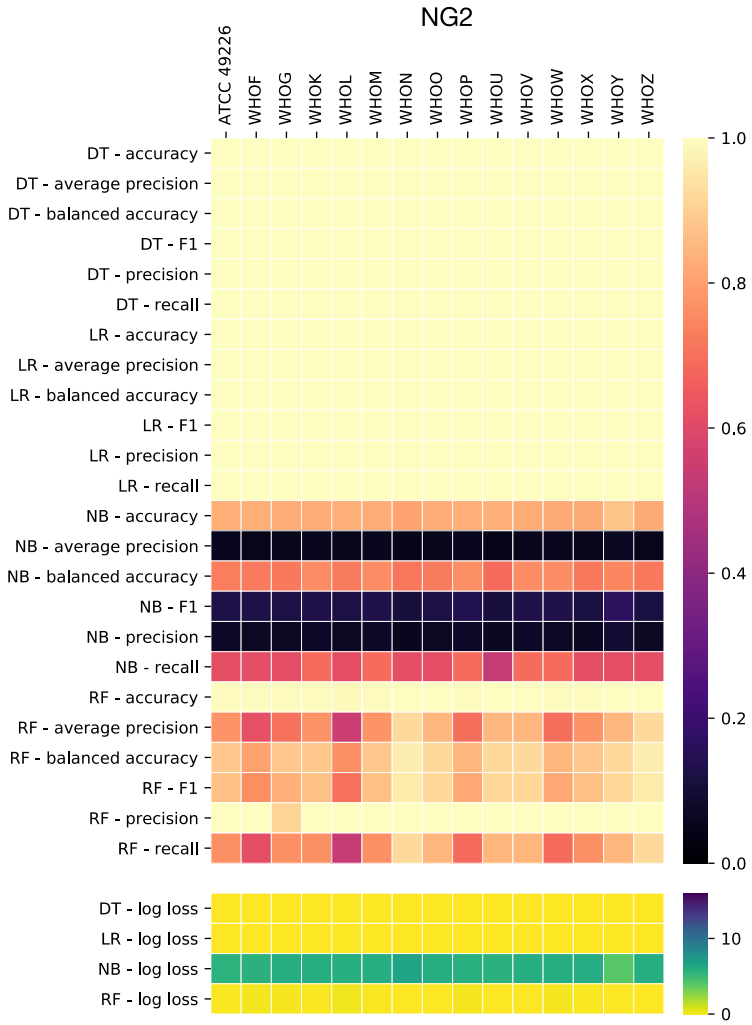


Supplementary Figure 3-38. *N. gonorrhoeae* ciprofloxacin resistance prediction models using mutations in datasets NG1 and NG2. Fifteen *N. gonorrhoeae* reference sequences were used. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

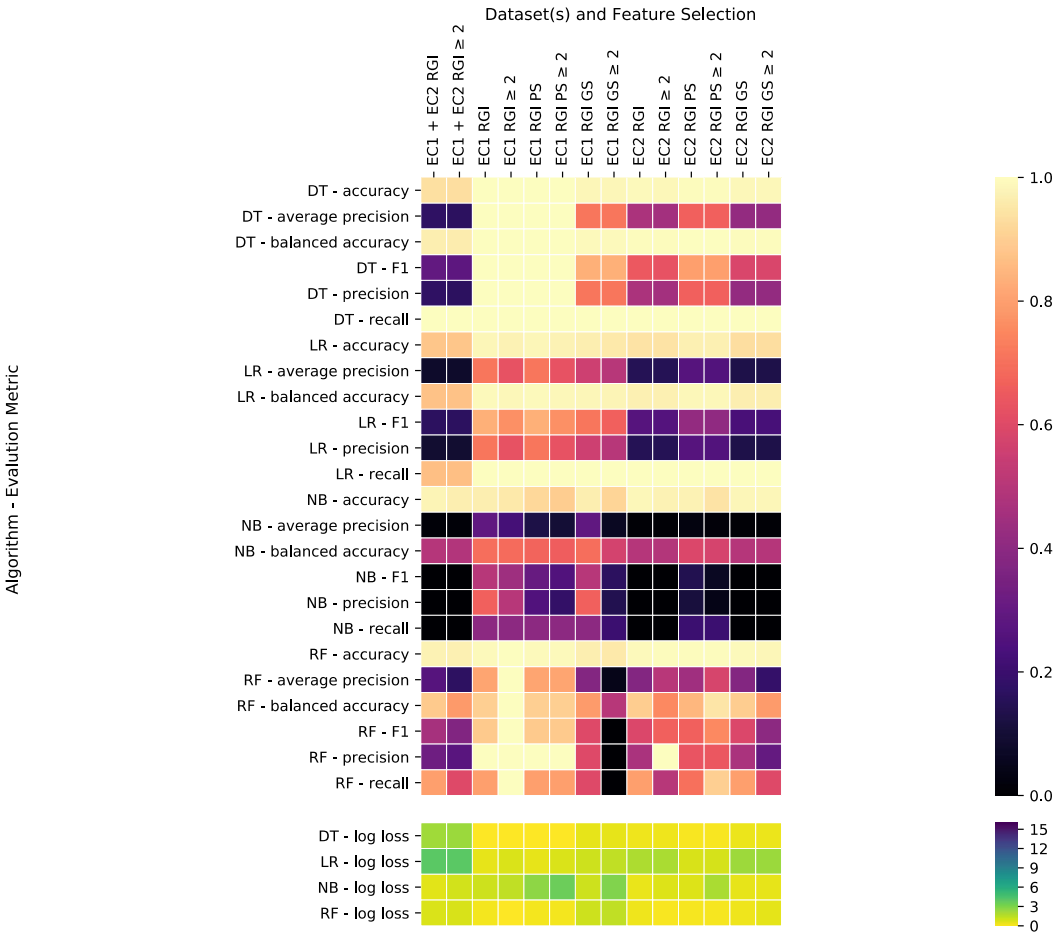
Supplementary Figure 3-39. *N. gonorrhoeae* tetracycline resistance prediction models using mutations in datasets NG1 and NG2. Fifteen *N. gonorrhoeae* reference sequences were used. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.



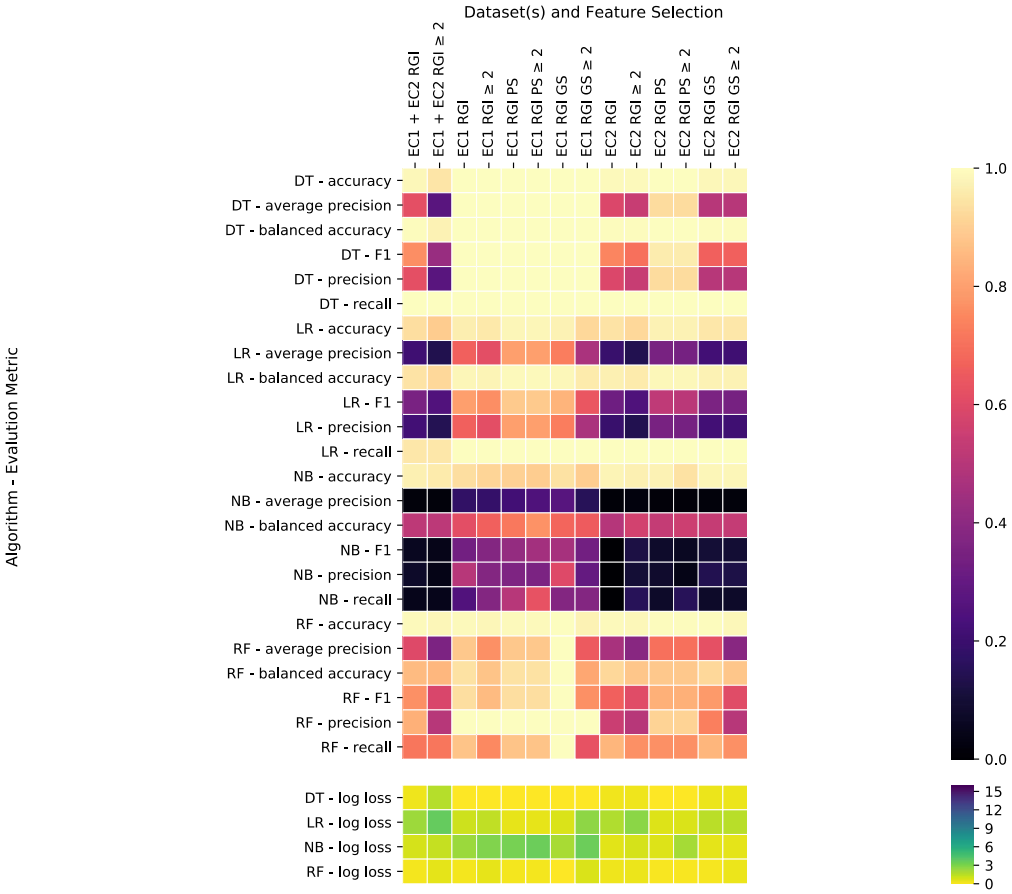
Supplementary Figure 3-40. *N. gonorrhoeae* spectinomycin resistant prediction models for *N. gonorrhoeae* (NG1). Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.



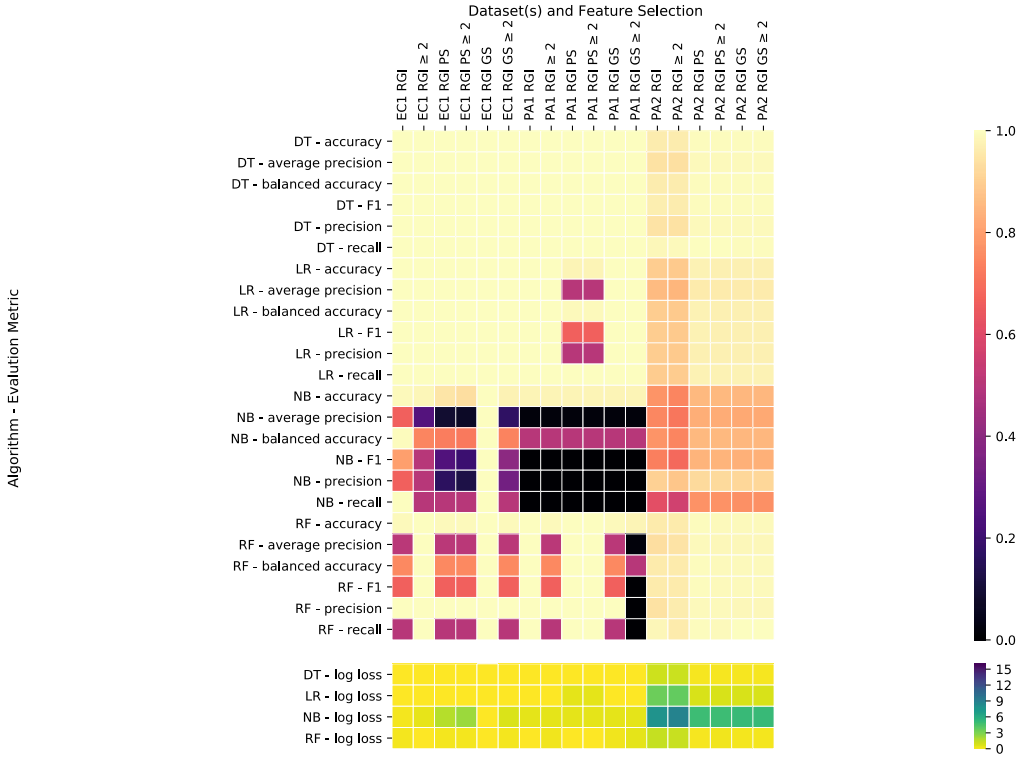
Supplementary Figure 3-41. *N. gonorrhoeae* cefixime resistance prediction models using mutations in dataset NG1. Fifteen *N. gonorrhoeae* reference sequences were used. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.



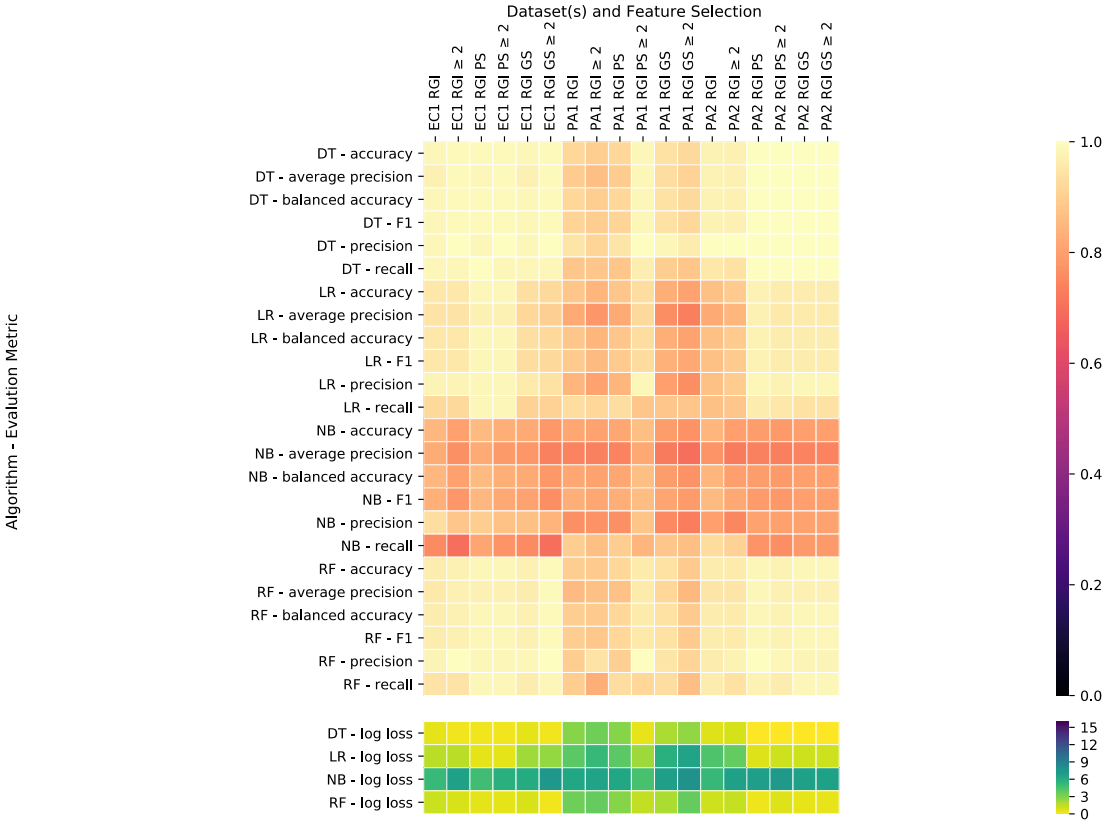
Supplementary Figure 3-42. *E. coli* ertapenem resistance prediction models using known resistance determinants in datasets EC1 and EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.



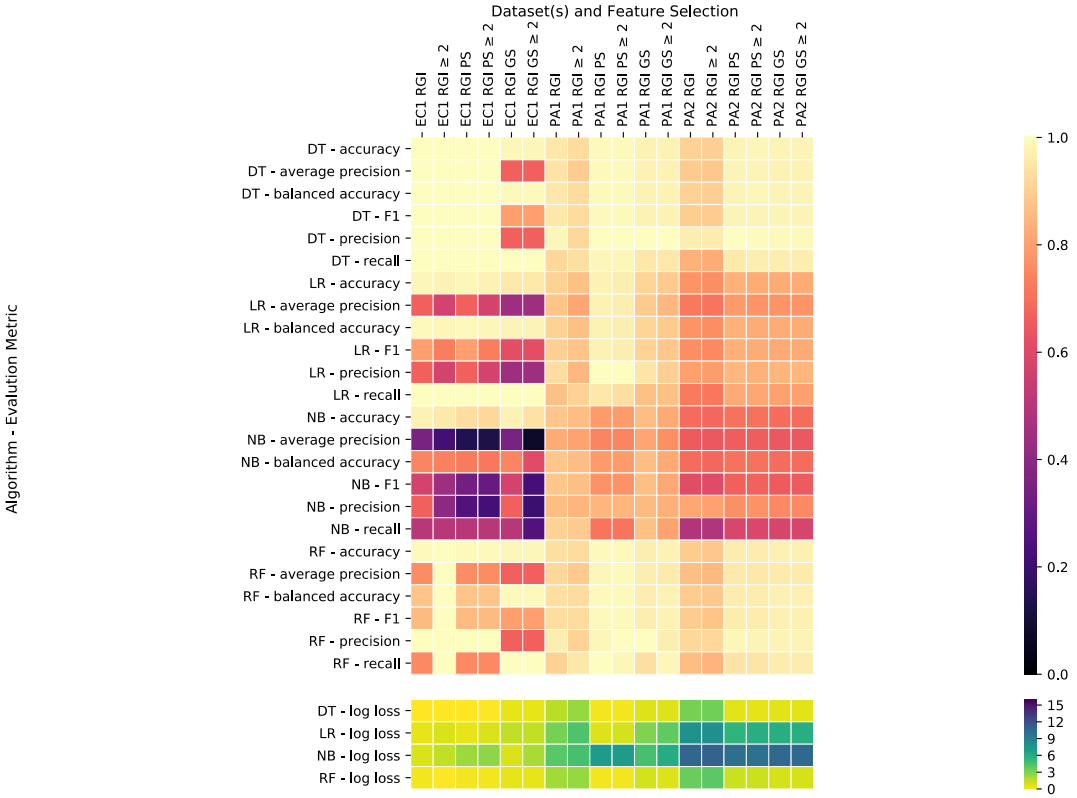
Supplementary Figure 3-43. *E. coli* nitrofurantoin resistance prediction models using known resistance determinants in datasets EC1 and EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.



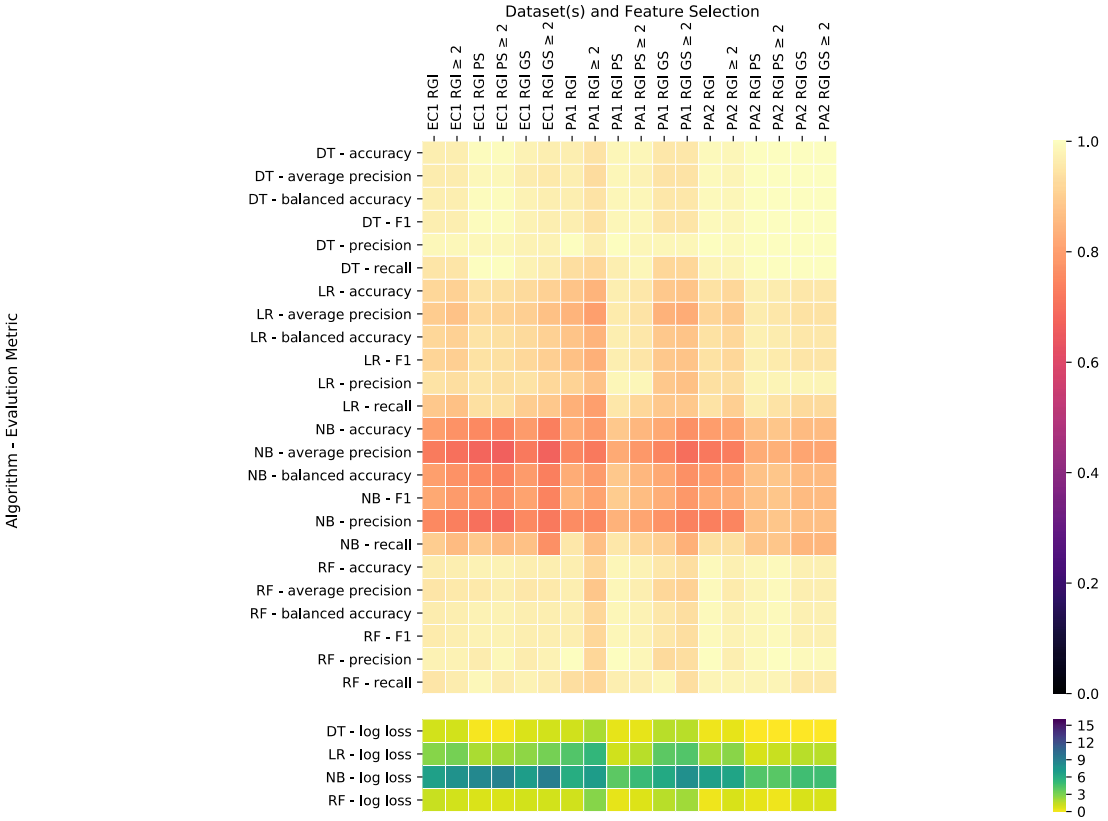
Supplementary Figure 3-44. *E. coli* and *P. aeruginosa* amikacin resistance prediction models using known resistance determinants in datasets EC1, PA1, and PA2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.



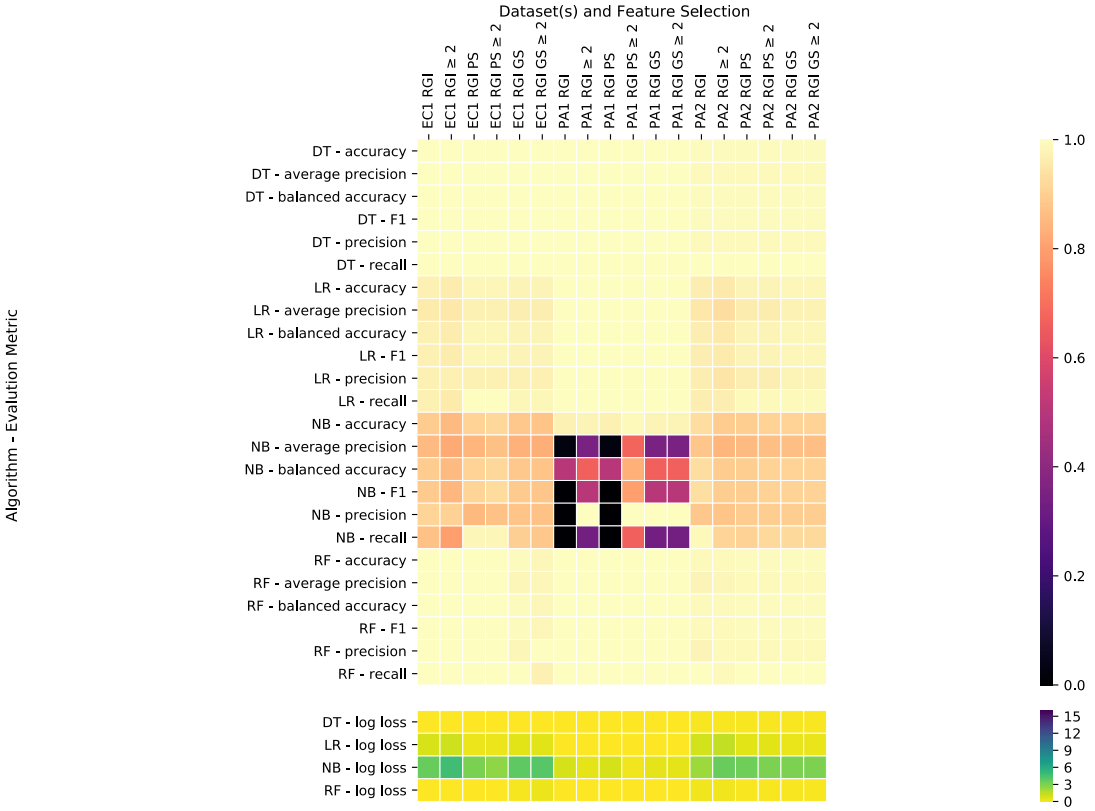
Supplementary Figure 3-45. *E. coli* and *P. aeruginosa* ceftazidime resistance prediction models using known resistance determinants in datasets EC1, PA1, and PA2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.



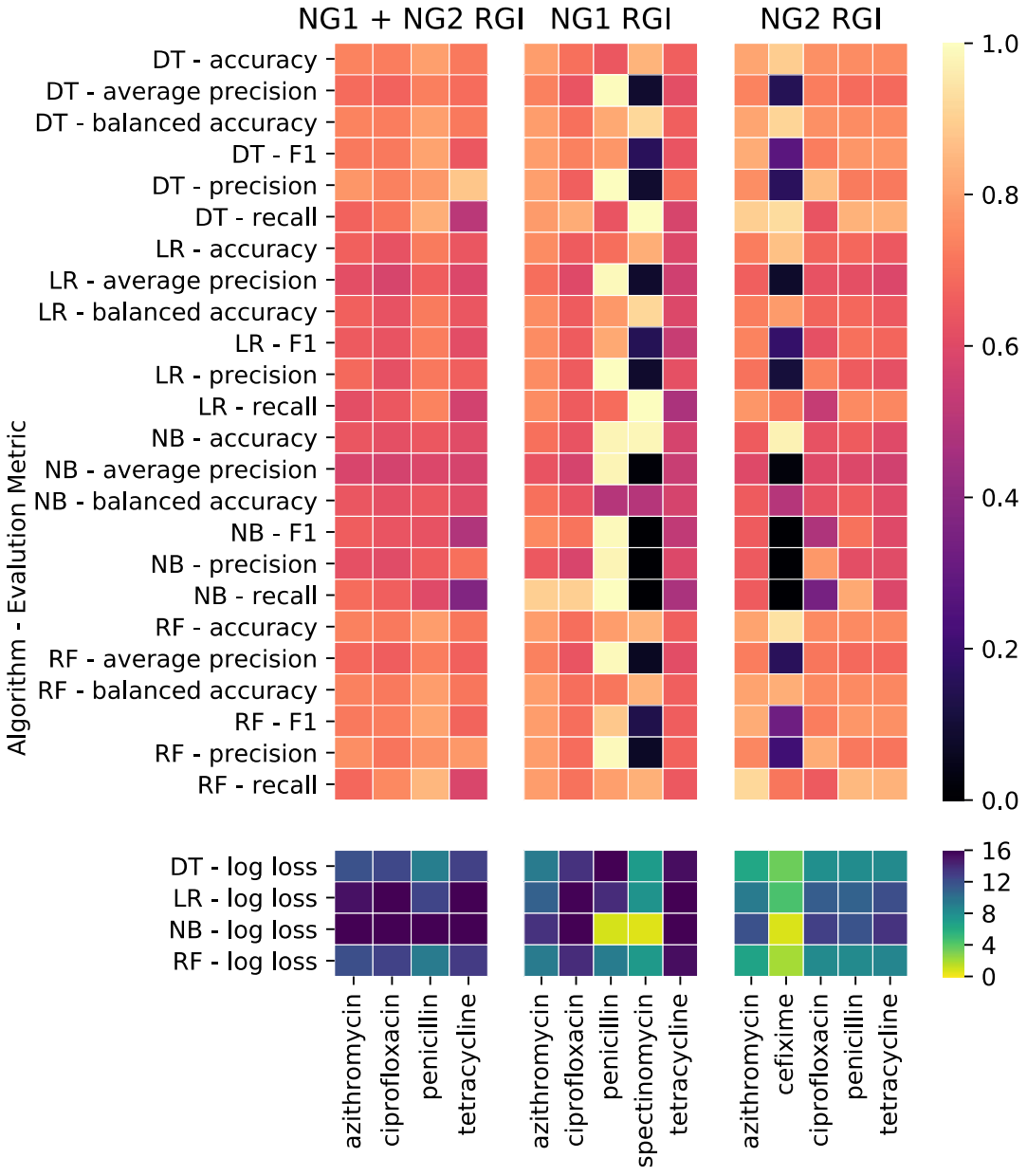
Supplementary Figure 3-46. *E. coli* and *P. aeruginosa* meropenem resistance prediction models using known resistance determinants in datasets EC1, PA1, and PA2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.



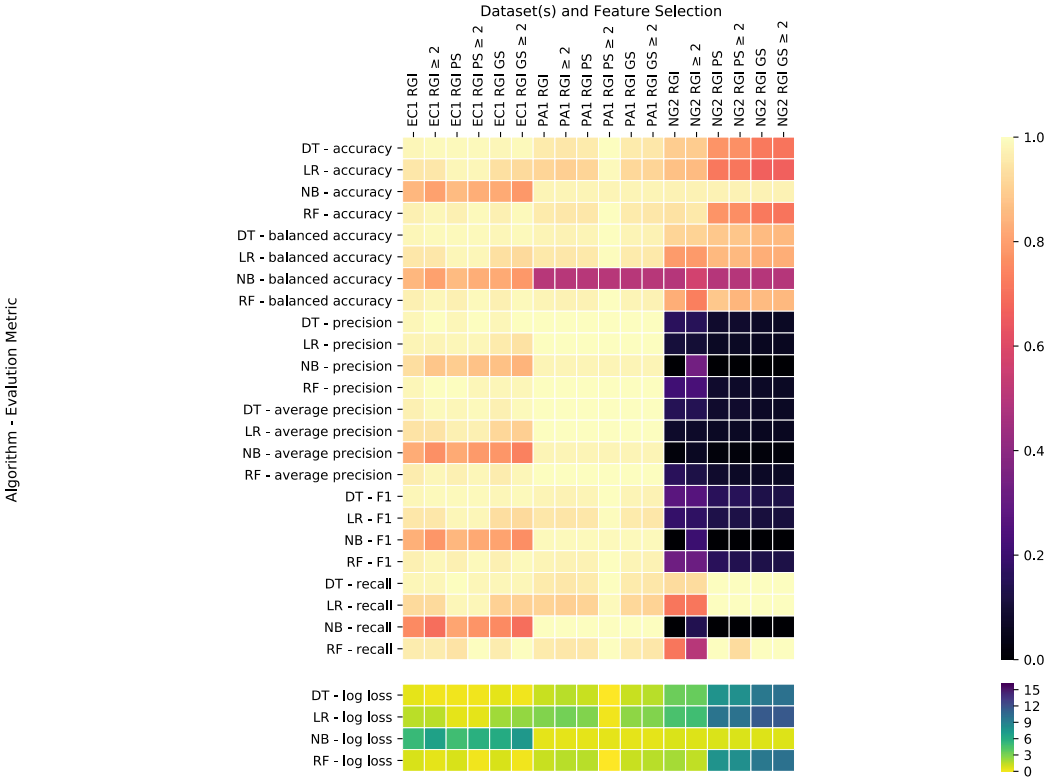
Supplementary Figure 3-47. *E. coli* and *P. aeruginosa* piperacillin-tazobactam resistance prediction models using known resistance determinants in datasets EC1, PA1, and PA2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.



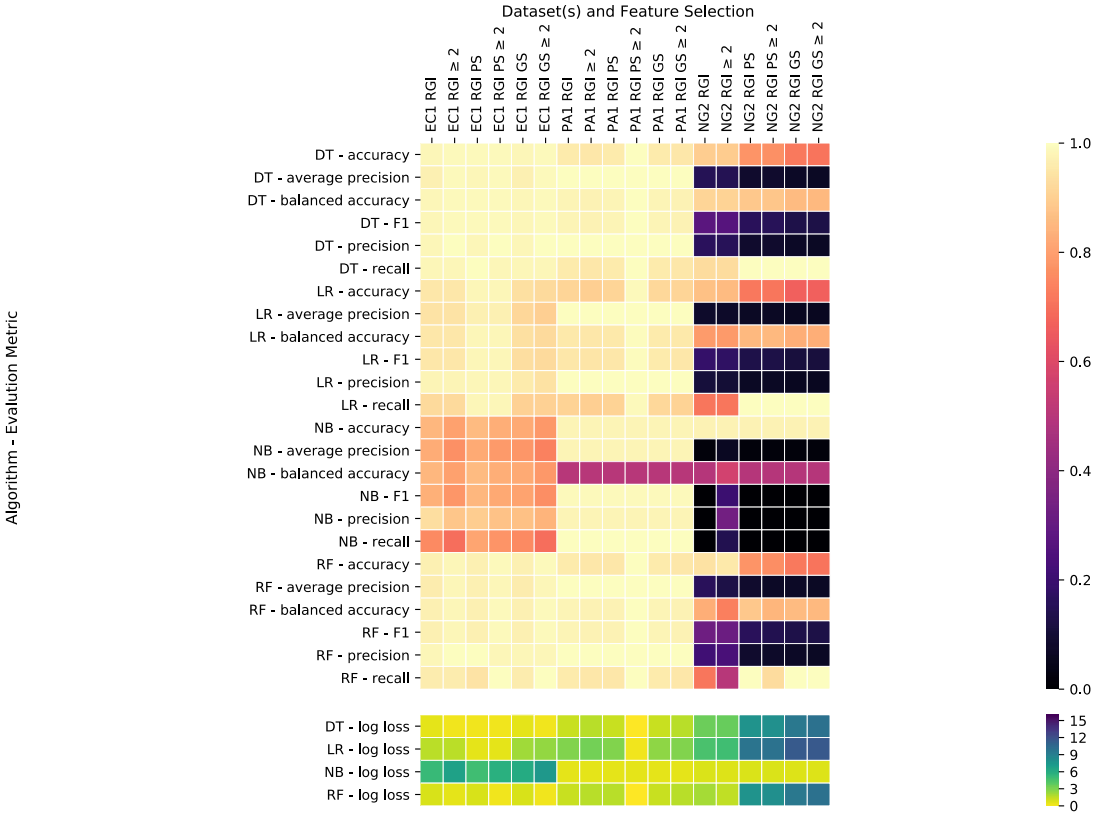
Supplementary Figure 3-48. *E. coli* and *P. aeruginosa* tobramycin resistance prediction models using known resistance determinants in datasets EC1, PA1, and PA2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.



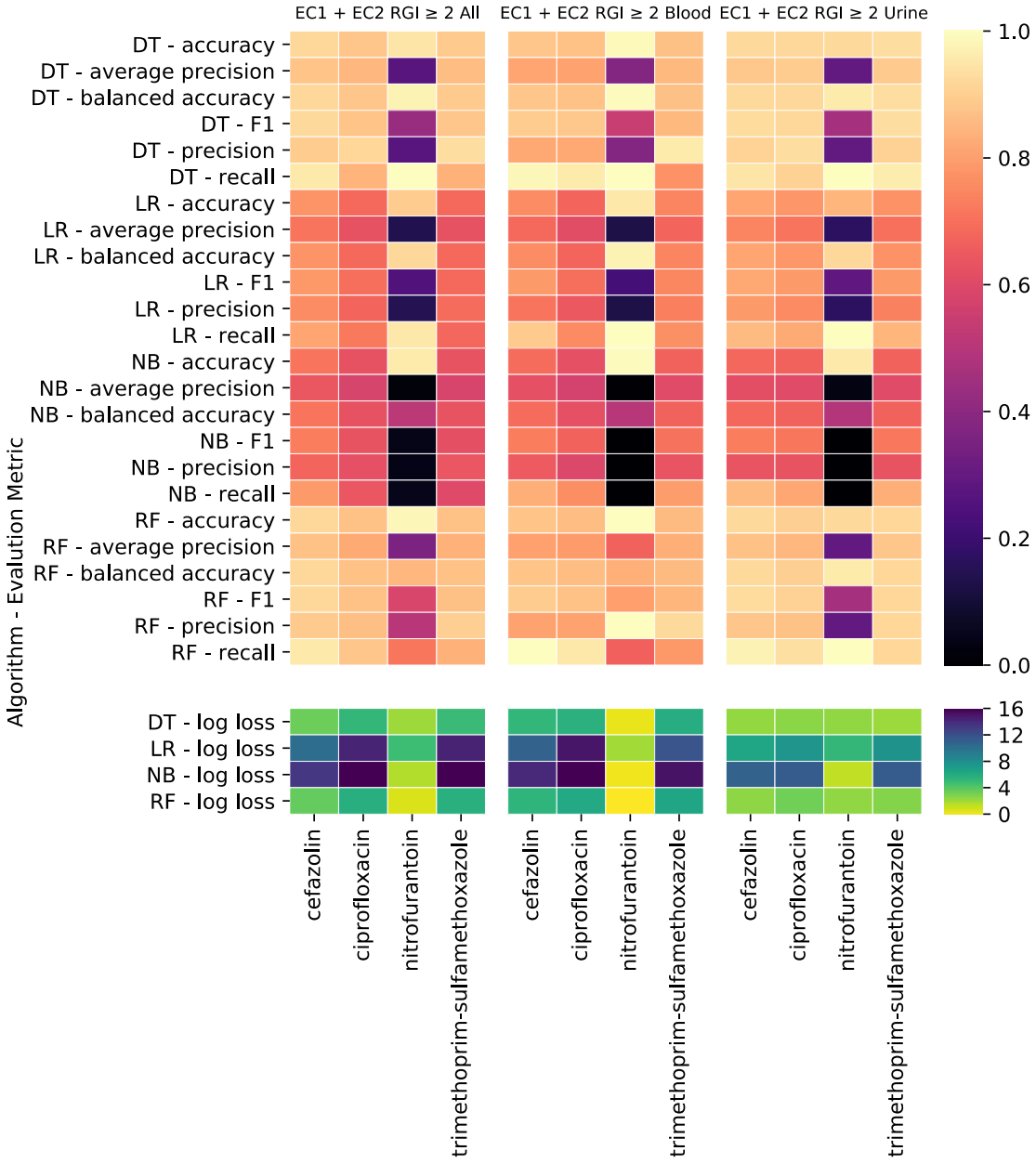
Supplementary Figure 3-49. *N. gonorrhoeae* AMR prediction models using known resistance determinants and no filtering with datasets NG1 and NG2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.



Supplementary Figure 3-50. *E. coli, P. aeruginosa* and *N. gonorrhoeae* cefixime resistance prediction models using known resistance determinants in datasets EC1, PA1, and NG2 grouped by evaluation metric. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.



Supplementary Figure 3-51. *E. coli, P. aeruginosa* and *N. gonorrhoeae* cefixime resistance prediction models using known resistance determinants in datasets EC1, PA1, and NG2 grouped by algorithm. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. Representation of features (i.e., PS), physicochemical filtering (i.e., GS), inclusion of resistance determinants if they are found in ≥ 2 samples (i.e., PS), as in Figure 3-1.



Supplementary Figure 3-52. *E. coli* AMR prediction models stratified by site of infection using known resistance genes found in more than one sample in datasets EC1 and EC2. Each square represents an AMR prediction model created using an algorithm, features, and assessed using an evaluation metric where its colour represents the performance. On the y-axis are the algorithms (e.g., logistic regression (LR), decision tree (DT), random forest (RF) and naïve Bayes (NB)) and evaluation metrics used to assess model performance. For more detail on performance interpretation, see Figure 3-1.

Supplementary Tables

Supplementary Table 3-1. Evaluation metric formulas. TP, TN, FP, FN indicate true positive, true negative, false positive, false negative.

Evaluation Metric	Formulas
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R) (or Sensitivity)	$\frac{TP}{TP + FN}$
Specificity (S)	$\frac{TN}{TN + FP}$
F1	$\frac{2P \cdot R}{P + R}$
Balanced accuracy	$\frac{R + S}{2}$

CHAPTER FOUR: Genomic feature selection drives antibiotic minimum inhibitory concentration prediction performance

CHAPTER FOUR PREFACE

Author contributions: KKT and AGM conceived the project and designed experiments. KKT performed the analysis. KKT wrote this chapter.

Acknowledgements

This research was funded by the Canadian Institutes of Health Research (PJT-156214 to A.G.M.), a David Braley Chair in Computational Biology to A.G.M., and an Ontario Graduate Scholarship, a McMaster University MacDATA Institute Graduate Fellowship, and a Michael G. DeGroote Institute for Infectious Disease Research Michael Kamin Hart Memorial Scholarship to K.K.T. Computer resources were supplied by the McMaster Service Lab and Repository computing cluster, funded in part by grants from the Canadian Foundation for Innovation (34531 to A.G.M.) and donation of hardware from Cisco Systems Canada, Inc.

ABSTRACT

The World Health Organization and Centers for Disease Control and Prevention have described antimicrobial resistance (AMR) as a public and global health crisis. One challenge is the turnaround time for AMR diagnosis, which may be improved by using next-generation sequencing technology in clinical microbiology laboratories. The optimal workflow includes by-passing culture-based phenotypic methods through genome sequencing and bioinformatics analyses, with the result an accurate AMR phenotype prediction. While a few studies have predicted antibiotic minimum inhibitory concentrations (MICs), we show the effect of using different genetic features, algorithms, and evaluation metrics on *Escherichia coli, Neisseria gonorrhoeae* and *Pseudomonas aeruginosa* MIC prediction models. We used two different evaluation metrics to assess the MIC prediction models. First, we showed that using chromosome and plasmid-borne resistance determinants generated better prediction models than only plasmid-borne resistance determinants. Then, we showed that feature filtering and representation had little effect on MIC prediction model performance. We further demonstrated use of known resistance determinants generates improved AMR *E. coli* MIC prediction models, whereas using mutations generated using a reference genome sequence improved AMR *N. gonorrhoeae* MIC prediction models. However, with *P. aeruginosa* it was challenging to generate highly accurate prediction models with resistance determinants or mutations. The accuracy of the MIC prediction models within a two-fold dilution factor of the laboratory determined MIC were 86% for *E. coli* (9 antibiotics), 98% for *N. gonorrhoeae* (8 antibiotics), and 39% for *P. aeruginosa* (7 antibiotics). We illustrate that interpreting

MIC prediction models is useful for understanding the mechanisms driving resistance when using hundreds of known resistance determinants but not thousands of mutations relative to a reference. Lastly, we show that MIC prediction models only describe correlative and not causative relationships between genetic features and an antibiotic MIC. Our work demonstrates the parameters that need to be considered when genomic-based MIC prediction models are used in clinical microbiology laboratories to inform clinical treatment and public health initiatives.

INTRODUCTION

Antimicrobial resistance (AMR) is a growing public health threat, but new technologies are being used to address the lack of new interventions to prevent the spread and treatment of drug resistant infections (Baker, Payne, Rappuoli, & De Gregorio, 2018). Currently, most clinical microbiology labs use gold standard culture-dependent phenotypic methods to determine antibiotic susceptibility profiles to inform antibiotic treatment. However, culture-dependent methods rely on *in vitro* growth of viable organisms and face inability to routinely test novel antibiotics, recipe modifications to support fastidious pathogens, reproducibility challenges for certain antibiotics, and do not account for factors that can influence the outcomes of infection, e.g., biofilm formation (Anonymous, 2019; Burnham, Leeds, Nordmann, O'Grady, & Patel, 2017). As such, there is growing interest for clinical laboratories to test the potential of next-generation sequencing technologies to predict AMR and bypass the time required for conventional culture-dependent AMR phenotypic methods (Ransom, Potter, Dantas, & Burnham, 2020).

The current workflow for predicting AMR phenotypes from genotypes is to use next-generation sequencing technology (e.g., Illumina, PacBio, Oxford Nanopore) to generate sequence reads which can then be processed in several ways to identify genetic features. After evaluating and excluding low quality reads, mutations can be identified using breseq (Deatherage & Barrick, 2014), *k*-mers (short nucleotide sequences) can be generated using Ray (Boisvert, Laviolette, & Corbeil, 2010), or genomes can be assembled using SPAdes and HyAsP (Bankevich *et al.*, 2012; Müller & Chauve, 2019).

Typically, the latter is performed and known resistance determinants are annotated from the genome assemblies using databases such as the Comprehensive Antibiotic Resistance Database (Alcock *et al.*, 2020), ARG-ANNOT (Gupta *et al.*, 2014), ResFinder (P. T. L. C. Clausen, Zankari, Aarestrup, & Lund, 2016), homology models such as ResFams (Gibson *et al.*, 2015), or manual use of alignment software such and BLAST (Madden, 2013). These genetic features (e.g., mutations, *k*-mers, known resistance determinants) are then used to predict AMR phenotypes using either a rules-based or machine learning modelling method. Rules-based algorithms typically assume that the presence of a known resistance determinant leads to a resistant phenotype, whereas a susceptible phenotype is inferred from its absence (Bradley *et al.*, 2015; Pesesky *et al.*, 2016; Shelburne *et al.*, 2017; Tsang *et al.*, 2021). Yet, machine learning approaches have increasingly illustrated the inaccuracy of rules-based algorithms (Moradigaravand *et al.*, 2018; Pesesky *et al.*, 2016; Tsang *et al.*, 2021), in part due to the fact the presence of an AMR gene does not guarantee its expression nor generation of a clinically actionable MIC.

Machine learning methods infer patterns between genetic features and AMR phenotypes in a training dataset to build a model that is then evaluated using a test dataset. The end goal is production of a computational method to accurately predict AMR phenotypes from genome sequences of newly acquired samples based upon underlying patterns in the genomic data. A number of different machine learning algorithms can be used, including logistic regression, random forest, naïve Bayes, decision trees, set covering machine, XGBoost, AdaBoost, and neural networks (Avershina *et al.*, 2021; Davis *et al.*, 2016; Drouin *et al.*, 2016; Hicks *et al.*, 2019; Kim *et al.*, 2019; Nguyen *et al.*,

2018; Nguyen *et al.*, 2019; Shi *et al.*, 2019; Tsang *et al.*, 2021; Yang *et al.*, 2018). These machine learning algorithms differ in their ability to interpret the model (i.e., not only predict phenotype but identify genetic drivers) and in model performance. In addition to the variety of algorithms used, previously published AMR machine learning models have either predicted resistant (R) / susceptible (S) categories to match existing clinical guidelines or minimum inhibitory concentration values (MICs) for direct interpretation, with varying success. While predicting R/S may be sufficient for informing antibiotic treatment for an individual patient, predicting MICs can allow for more informative local and global surveillance, particularly for pathogens that have antibiotic MICs nearing clinical breakpoints. In addition, R/S categories are defined by international guidelines, such as the Clinical & Laboratory Standards Institute (CLSI, 2018) and the European Committee on Antimicrobial Susceptibility Testing (EUCAST, 2015), that have differing breakpoints and laboratory methods for different bacteria and/or antibiotics and are thus not generalizable (Cusack, Ashley, Ling, Roberts, *et al.*, 2019; Hombach *et al.*, 2013; Kassim *et al.*, 2016; Rodríguez-Baño *et al.*, 2012; Wolfensberger *et al.*, 2013). Previously published MIC prediction models have been built for *Klebsiella pneumoniae* (Nguyen *et al.*, 2018; Nguyen *et al.*, 2020), *Neisseria gonorrhoeae* (Eyre *et al.*, 2019; Eyre *et al.*, 2017; Hicks *et al.*, 2019), *Escherichia coli* (Pataki *et al.*, 2020), nontyphoidal *Salmonella* (Nguyen *et al.*, 2020; Nguyen *et al.*, 2019), and *Streptococcus pneumoniae (Y. Li et al., 2016; B. J. Metcalf et al., 2016)*. Most of these MIC prediction models perform with greater than 80% accuracy yet are difficult to compare because each uses different genetic features, algorithms, and evaluation metrics. To our knowledge, there is no study that

compares the effect of using different genetic features, algorithms, and evaluation metrics on MIC prediction model performance.

In this work, we develop MIC prediction models for *E. coli*, *N. gonorrhoeae*, and *P. aeruginosa* using different genetic features and filtering methods (known resistance determinants and mutations) and algorithms (linear regression, lasso LARS CV, and ridge regression), which we assess using two evaluation metrics (coefficient of determination and mean squared error). We demonstrate that using known resistance determinants improves *E. coli* MIC prediction model performance, whereas using mutations improves *N. gonorrhoeae* MIC prediction model performance. However, neither known resistance determinants nor mutations perform well for *P. aeruginosa* MIC prediction models.

METHODS

Bacterial isolates

We used datasets of *E. coli* (EC2) (MacFadden *et al.*, 2019), *P. aeruginosa* (PA2) (Davis *et al.*, 2020) from the PATRIC database, and two previously published *N. gonorrhoeae* collections (NG1 (Lee *et al.*, 2018) and NG2 (Eyre *et al.*, 2017)) that were also used in Chapter 3. Unpublished phenotypic testing data from dataset EC2 are available from https://github.com/karatsang/MICprediction. All datasets included minimum inhibitory concentrations for a number of different antibiotics using CLSI guidelines (CLSI, 2018). For the PA2 dataset, we only included antibiotic phenotypes generated using the CLSI guidelines with 50 or more genomes. For more information about the phenotypic measurements of each dataset, refer to their respective primary publication source, as outlined in Chapter 3. Descriptions of these genomes are presented in Table 3-1. In terms of balance, all datasets had more than three minimum inhibitory concentration values based on more than 10% of all genomes, except for EC2 (ertapenem, meropenem, nitrofurantoin) and NG1 (spectinomycin) (Supplementary Figure 4-1 to 4-4).

Genetic feature generation

For each isolate, raw short read sequences were first trimmed using Trimmomatic and then either used to identify mutations using breseq (v 0.35.3) (Deatherage & Barrick, 2014), assembled into chromosomal and plasmid DNA using SPAdes, or assembled into plasmid DNA alone using HyAsP (v1.0.0) (Müller & Chauve, 2019). Since the PA2 dataset only provided genome assemblies (FASTAs), we were unable to include HyAsP plasmids as a feature set, but we simulated reads from the FASTA sequences using ART (v 2.3.7) (Huang, Li, Myers, & Marth, 2012) with the options `art_illumina -ss HS25 -

sam --paired --len 150 --fcov 10 --mflen 200 --sdev 10` for mutation identification by breseq. Mutations were identified using breseq with default parameters and the following reference sequences: *E. coli* O83:H1 str. NRG 857C (ASM18334v1), *E. coli O157:H7 str. Sakai DNA* (NC_002695.2), *P. aeruginosa* PAO1 (NC_002516.2), *P. aeruginosa* UCBPP-PA14 (NC_008463.1), *N. gonorrhoeae* ATCC 49226 (NZ_CP045728.1), WHOF, WHOG, WHOK, WHOL, WHOM, WHON, WHOO, WHOP, WHOU, WHOV, WHOW, WHOX, WHOZ (Unemo *et al.*, 2016). Gdtools (Deatherage & Barrick, 2014) was used to annotate the breseq results, while known resistance determinants were predicted in the chromosomal and plasmid DNA assemblies using the Resistance Gene Identifier (RGI, v 5.1.0) and Comprehensive Antibiotic Resistance Database (CARD, v 3.0.8) (Alcock *et al.*, 2020). RGI categorizes resistance determinants as 'Perfect' or 'Strict' if the predicted amino acid sequence is 100% identical to the reference sequence in CARD or if the predicted amino acid sequence passes a curated bitscore cutoff, respectively. Since RGI is dependent on CARD, RGI is unable to identify new resistance determinants, while breseq is CARD-independent, meaning it can identify unknown mutations.

Genetic feature filtering

We removed any mutations from breseq that were only observed in one isolate in a given dataset to reduce the potential misrepresentation of data, as it is difficult to differentiate between sequencing error, transcription error, and a *bona fide* mutation if it appears in a single isolate. In contrast, if a mutation is identified in multiple isolates, it is

less likely to be a sequencing/transcription error. To remove potential spurious 'Strict' resistance determinant predictions by RGI, we applied a Grantham Score (Grantham, 1974) filter to categorize amino acid substitutions (relative to CARD reference) into classes of physicochemical dissimilarity: conservative (0-50), moderately conservative (51-100), moderately radical (101-150) or radical (≥151). We removed any RGI hits that had a Grantham Score greater than 151. As with mutations, we included a filter that only allowed resistance determinants found in two or more samples (≥2),

Minimum inhibitory concentration (MIC) prediction modelling

We used three different algorithms for predicting MICs (linear regression, lasso least-angle regression (LARS) cross validation (CV), ridge regression CV). Linear regression is the simplest algorithm of the three and it attempts to predict AMR phenotype using the best straight line fit to a set of genetic features (Lai *et al.*, 1978). The latter two algorithms are penalized versions of linear regression, with Lasso LARS CV setting less contributive genetic features to be zero (Efron, Hastie, Johnstone, & Tibshirani, 2004) and ridge regression CV assigning genetic features with minor contribution to be close to zero (Hoerl & Kennard, 1970). Lasso LARS CV is useful when only a few genetic features are driving resistance, yet it is not useful for interpretation of drivers of resistance in complicated data sets as it will arbitrarily select one (and remove the other) if there are two highly collinear genetic features. Ridge regression is appropriate when all genetic features need to be incorporated into the model, as it will not reduce any genetic features and can be reflective of additive or synergistic

AMR gene interactions. For each approach, hyperparameters were tuned using a threefold stratified shuffle split cross-validation scheme and models evaluated using the coefficient of determination (R^2) on the test set and mean squared error (MSE) on the test set and overall dataset. R^2 compares the fit of the chosen model to that of a straight line (e.g., the null hypothesis). If R^2 is negative, it means the model has worse predictions than a baseline model that always predicts the average of the data. The closer R^2 is to 1, the better the model explains the MIC prediction. A limitation of R^2 is that it does not explain prediction error, which is why MSE was also used. MSE measures the average of the squares of the errors and an ideal MSE is 0 as it indicates a better prediction model. Model selection was performed for each dataset in its entirety, and not each antibiotic-pathogen combination, because we strived for the most parsimonious method of model prediction. In other words, for each dataset we selected one feature set and one algorithm for final model performance evaluation and evaluated the performance of each feature set across all antibiotics using R^2 and MSE, selecting the best algorithm based on majority rule (e.g., the feature set and model combination that worked best for the most antibiotics). If there were two feature sets that had the same R^2 and MSE values, then the most parsimonious feature set (i.e., less filtering / computational effort) was selected.

To predict MICs, the $\log_2$ values of the MICs were used for generation of all prediction models. We bound the accuracy to within a two-fold dilution factor of the laboratory determined MIC, which is consistent with current U.S. Food and Drug Administration standards for diagnostic tools, Canadian National Microbiology Lab quality assurance for antibiotic susceptibility testing of *N. gonorrhoeae,* and conventional

clinical microbiology practices (Reller, Weinstein, Jorgensen, & Ferraro, 2009; Sawatzky *et al.*, 2015; U.S. Department of Health and Human Services, 2009).

Machine learning and dataset partitioning were performed using scikit-learn (Pedregosa *et al.*, 2011) (v0.20.0), with data otherwise manipulated using numpy (Oliphant, 2006) (v1.17.2) and pandas (McKinney, 2010) (v0.25.1). Heatmaps were generated using seaborn (v0.11.0). The code and conda environments (using python v3.7.2), and intermediate data files required to generate this analysis are available: https://github.com/karatsang/MICprediction.

RESULTS

Genetic feature generation and filtering

We used the *Escherichia coli* (EC2), *Pseudomonas aeruginosa* (PA2), and *Neisseria gonorrhoeae* (NG1 and NG2) datasets from Chapter 3 (Table 3-1) and as in Chapter 3 we used SPAdes (Bankevich *et al.*, 2012) to assemble chromosomes and plasmids (Table 3-2), and HyAsP (Müller & Chauve, 2019) to predict plasmid assembles alone (Table 3-3). The only exception was the *Pseudomonas aeruginosa* PA2 dataset, for which raw sequencing reads were not available and we used the available assemblies. We simulated reads using the assemblies, which had resulted in an average read coverage of 5.5-fold. Using breseq and the simulated reads, we generated a range of 390,000-430,000 mutations, depending on the reference sequence used (Supplementary Table 4-1). As in Chapter 3, we represented RGI as a single feature or the presence of the resistance gene combined with its RGI criteria called the ‘PS’ representation (Perfect amino acid

sequence match to CARD reference sequence or Strict variant of the CARD reference sequence). We also excluded resistance determinants that were only found in one genome and created a Grantham score ('GS') filter to remove any predicted resistance genes that were considered to have radical amino acid changes, and thus likely representative of RGI false positives. Generally, the PS representation increased the number of resistance determinants by 10-30%, while excluding resistance determinants only found in one genome and the GS filter decreased the number of resistance determinants by 10-50%. Refer to this section in Chapter 3 for a more detailed description of genetic feature generation and filtering.

Evaluation metrics to assess resistance determinant-based MIC prediction models

One of the first decisions made was to determine which genetic features to use for MIC prediction model creation. We tested a number of different methods to represent and filter known resistance determinants from chromosomes and plasmids or plasmids alone. We decided to use the coefficient of determination (R^2) and mean squared error (MSE) as evaluation metrics. For most antibiotics in the *E. coli* and *N. gonorrhoeae* datasets, using only plasmid-borne resistance determinants resulted in poorer performing MIC prediction models, regardless of any filter (Figure 4-1 to 4-3). Using R^2 better illustrates the discrepancy in MIC model performance between using resistance determinants in plasmids and chromosomes versus plasmids alone. For example, the performance of *E. coli* cefazolin, meropenem, and ertapenem MIC prediction models did not notably change using MSE but decreased when using R^2 (Figure 4-1). There was very little difference in

performance among filters applied to use of known resistance determinants from chromosomes and plasmids in *E. coli* (Figure 4-1), yet for *N. gonorrhoeae* using the Perfect and Strict (PS) representation decreased R^2 and MSEs in both datasets (Figure 4-1,4-2). The Grantham Score filter only impacted (and decreased) performance for the NG2 dataset (Figure 4-2). For *N. gonorrhoeae*, spectinomycin resistance prediction model performance conflicted among evaluation metrics: R^2 scored poorly but MSEs scores were improved. In *P. aeruginosa* (PA2), filtering had no effect on lasso LARS CV and linear regression MIC prediction models, however the PS representation performed worse than no filtering or use of the GS filter in the ridge regression CV prediction models (Figure 4-4). Overall, we observed use of resistance determinants from both plasmids and chromosomes with no filters was both the simplest approach and generated MIC prediction models with the best R^2 and MSE values across all pathogens.

Alongside determining the genetic features to power MIC prediction, different algorithms were also tested and assessed using both evaluation metrics. For *E. coli* (EC2) and both *N. gonorrhoeae* datasets (NG1 and NG2), we observed that use of lasso LARS CV and ridge regression CV both generated similarly performing models that were improved compared to those produced by linear regression (Figure 4-1 to 4-3). In contrast, with *P. aeruginosa* (PA2) use of lasso LARS CV outperformed both linear regression and ridge regression CV (Figure 4-4), although performance was generally low across all algorithms for this pathogen.

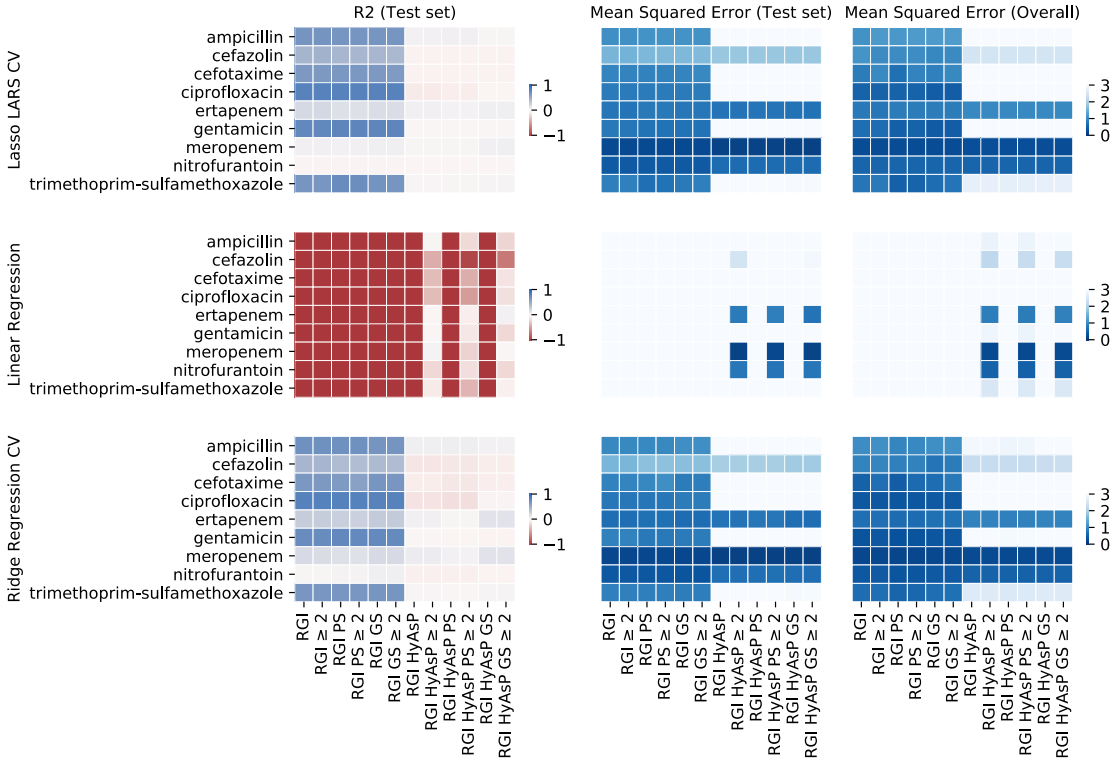


Figure 4-1. *E. coli* MIC prediction models using resistance determinants assessed with evaluation metrics (EC2). Each box represents an antibiotic prediction model generated using specific features (x-axis) and algorithm (y-axis). Lasso LARS Cross Validation (CV), linear regression, and ridge regression CV were used to generate MIC prediction models. The Resistance Gene Identifier (RGI) was used to predict known resistance genes in chromosomes and plasmids or just plasmids (HyAsP). We use a filter that included resistance determinants found in two or more samples (≥2). Perfect and Strict (PS) representation was used to capture the redundancy of multiple resistance determinants and the Grantham Score (GS) filter removed any spurious RGI predictions. Coefficient of determination (R^2) measures is the proportion of the variance in the dependent variable that is predictable from the independent variable. The closer R^2 is to 1 (i.e., darker blue), the better the prediction model. Mean squared error is measures the average of the squares of the errors. The closer the mean squared error is to 0 (i.e., darker blue), the better the prediction model.

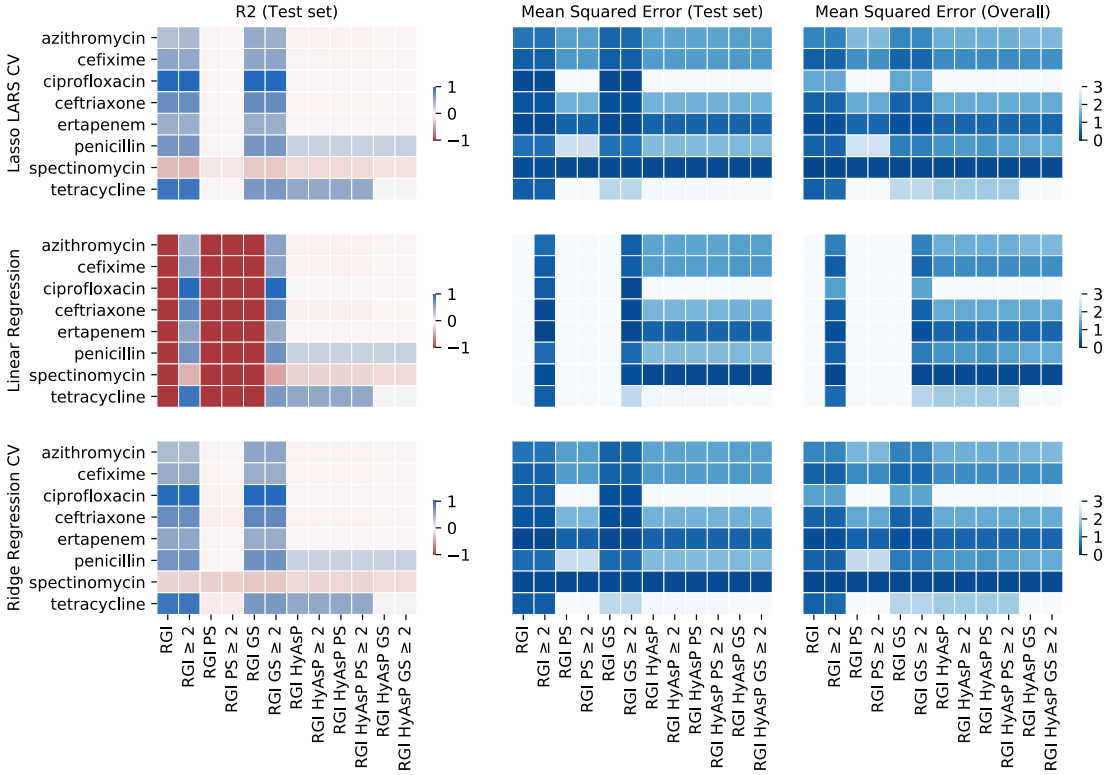


Figure 4-2. *N. gonorrhoeae* MIC prediction models using resistance determinants assessed with evaluation metrics (NG1). Each box represents an antibiotic prediction model generated using specific features (x-axis) and algorithm (y-axis). Lasso LARS Cross Validation (CV), linear regression, and ridge regression CV were used to generate MIC prediction models. Resistance Gene Identifier (RGI) was used to predict known resistance genes in chromosomes and plasmids or just plasmids (HyAsP). Only including features found in two or more samples (≥2), representation of features (i.e., PS), physicochemical filtering (i.e., GS) as in Figure 4-1. Coefficient of determination (R^2) and mean squared error were used as evaluation metrics with details in Figure 4-1.

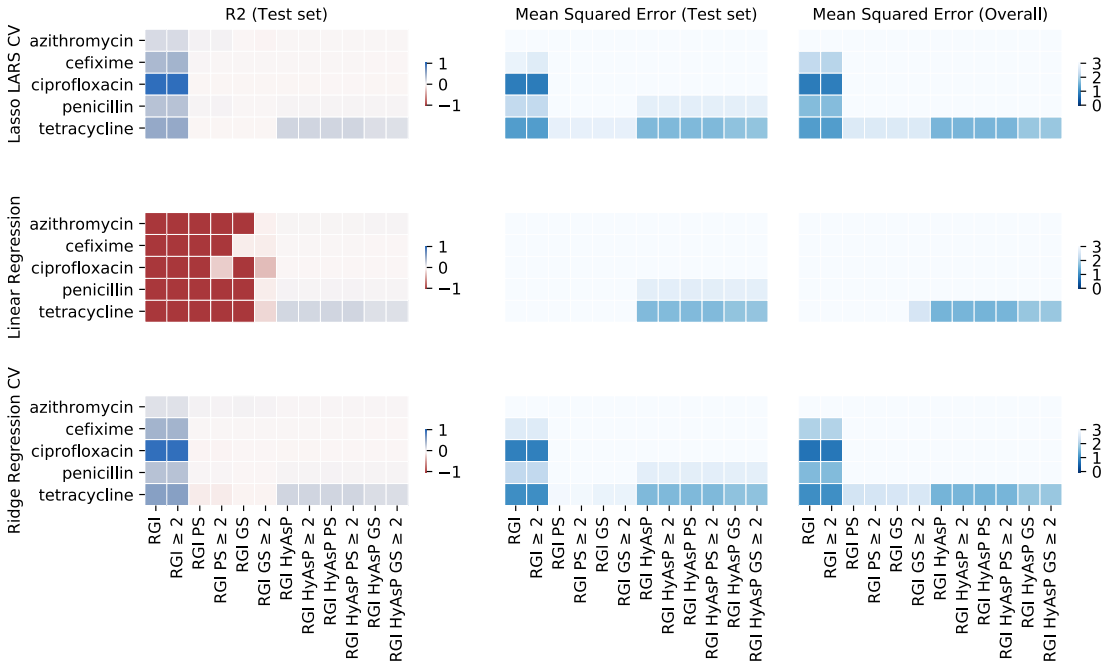


Figure 4-3. *N. gonorrhoeae* MIC prediction models using resistance determinants assessed with evaluation metrics (NG2). Each box represents an antibiotic prediction model generated using specific features (x-axis) and algorithm (y-axis). Lasso LARS Cross Validation (CV), linear regression, and ridge regression CV were used to generate MIC prediction models. Resistance Gene Identifier (RGI) was used to predict known resistance genes in chromosomes and plasmids or just plasmids (HyAsP). Only including features found in two or more samples (≥2), representation of features (i.e., PS), physicochemical filtering (i.e., GS) as in Figure 4-1. Coefficient of determination (R^2) and mean squared error were used as evaluation metrics with details in Figure 4-1.

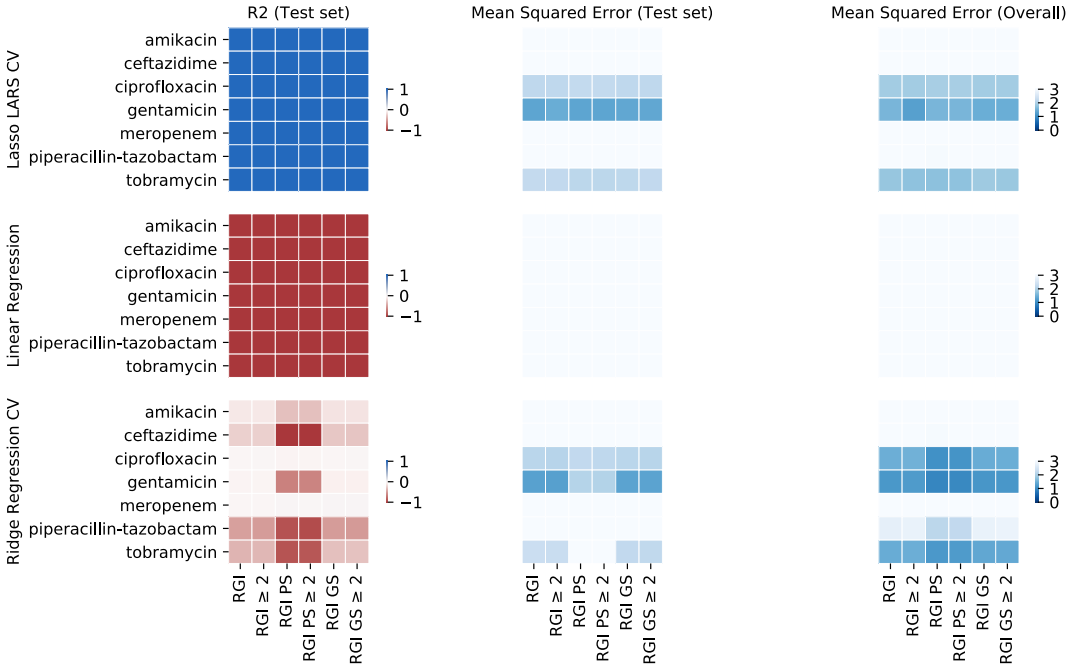


Figure 4-4. *P. aeruginosa* MIC prediction models using resistance determinants assessed with evaluation metrics (PA2). Each box represents an antibiotic prediction model generated using specific features (x-axis) and algorithm (y-axis). Lasso LARS Cross Validation (CV), linear regression, and ridge regression CV were used to generate MIC prediction models. Resistance Gene Identifier (RGI) was used to predict known resistance genes in chromosomes and plasmids. Only including features found in two or more samples (≥2), representation of features (i.e., PS), physicochemical filtering (i.e., GS) as in Figure 4-1. Coefficient of determination (R^2) and mean squared error were used as evaluation metrics with details in Figure 4-1.

Using mutations instead of known resistance determinants can improve MIC prediction models

We used two different reference strains to detect mutations against each sample in the *E. coli* EC2 dataset. Using R^2 and MSE, the mutations generated using *E. coli* O157:H7 str. Sakai produce similar but improved MIC prediction models than RGI determinants (Figure 4-5). For ampicillin MIC prediction, the R^2 value is better than MSE and these values are consistent across all algorithms (Figure 4-5). The best performing algorithm using mutations in the EC2 dataset was lasso LARS CV (Figure 4-5), compared to either lasso LARS CV or ridge regression CV with RGI determinants (Figure 4-1). In the *N. gonorrhoeae* datasets (NG1 and NG2), we used 5 different reference strains to generate mutations. In dataset NG1, there was no notable difference among combinations of the 5 different reference strains and both evaluation metrics, particularly in the MIC prediction models created using linear regression and ridge regression CV (Figure 4-6). In contrast, with dataset NG2 using MSE on the test set showed use of reference strain *N. gonorrhoeae* ATCC 49226 was superior, particularly for azithromycin MIC prediction (Figure 4-7). Yet, using R^2 there was little difference between using the 5 reference strains on MIC prediction model performance, suggesting MSE may be more discriminating for NG2 (Figure 4-7). Overall, for both NG1 and NG2 use of linear regression or ridge regression CV improved prediction models compared to using lasso LARS CV, with this difference being most pronounced for MSEs on the overall dataset (Figure 4-6, 4-7). In addition, the *N. gonorrhoeae* MIC prediction models generated using

mutations overall had improved R^2 and MSE values (Figure 4-6, 4-7) than those generated using RGI determinants (Figure 4-2, 4-3),

Overall, prediction of MICs for *P. aeruginosa* was considerably worse than for the other pathogens. For the *P. aeruginosa* PA2 dataset, we simulated reads using genome assemblies and then generated mutation feature sets using two different reference strains and breseq. Using *P. aeruginosa* PAO1 as a reference strain to generate mutations created improved prediction models compared to *P. aeruginosa* PA14 when applying R^2, a finding particularly highlighted by gentamicin resistance prediction (Figure 4-8). The best predictive algorithm was lasso LARS CV, based on R^2 values for amikacin, meropenem, piperacillin-tazobactam, and tobramycin MICs (Figure 4-8). Unlike the other two pathogens, the values of R^2 with lasso LARS CV were improved using known resistance genes generated by RGI compared to using mutations (Figure 4-4, 4-8).

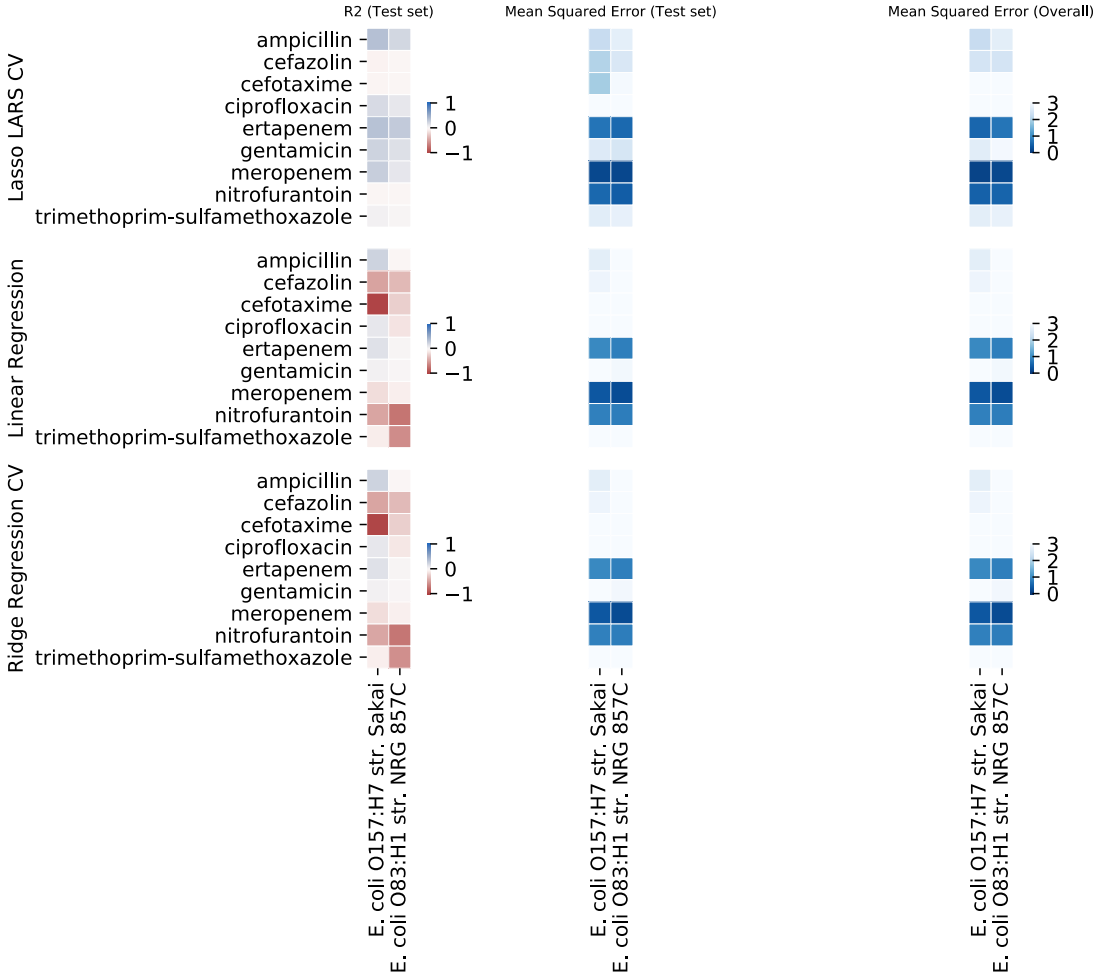


Figure 4-5. *E. coli* MIC prediction models using mutations assessed with evaluation metrics (EC2). Each box represents an antibiotic prediction model generated using a reference genome (x-axis) and algorithm (y-axis). Lasso LARS Cross Validation (CV), linear regression, and ridge regression CV were used to generate MIC prediction models. Breseq was used to identify mutations based on a reference genome and we only included mutations found in two or more samples (≥2). Coefficient of determination (R^2) and mean squared error were used as evaluation metrics with details in Figure 4-1.

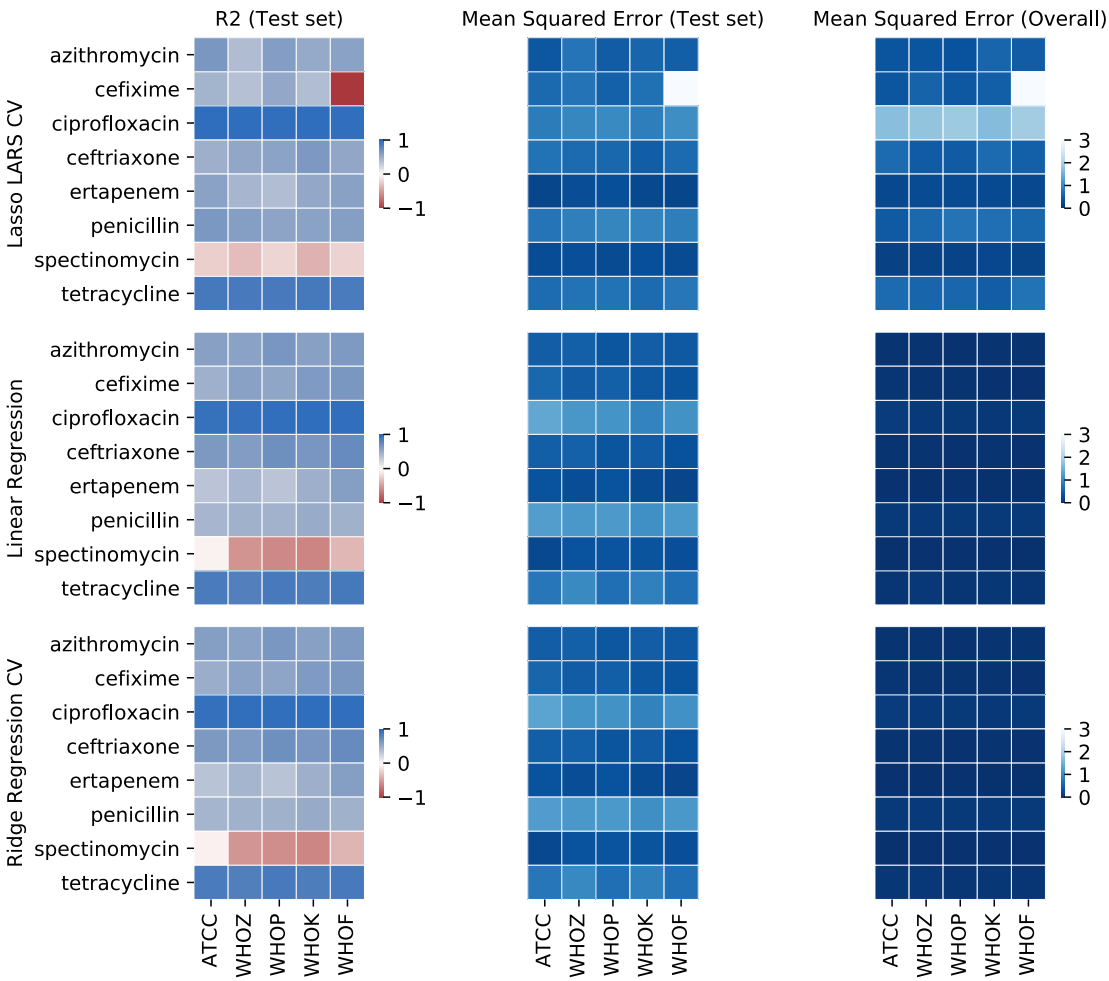


Figure 4-6. *N. gonorrhoeae* MIC prediction models using mutations assessed with evaluation metrics (NG1). Each box represents an antibiotic prediction model generated using a reference genome (x-axis) and algorithm (y-axis). Lasso LARS Cross Validation (CV), linear regression, and ridge regression CV were used to generate MIC prediction models. Breseq was used to identify mutations based on a reference genome and we only included mutations found in two or more samples (≥2). Coefficient of determination (R^2) and mean squared error were used as evaluation metrics with details in Figure 4-1.

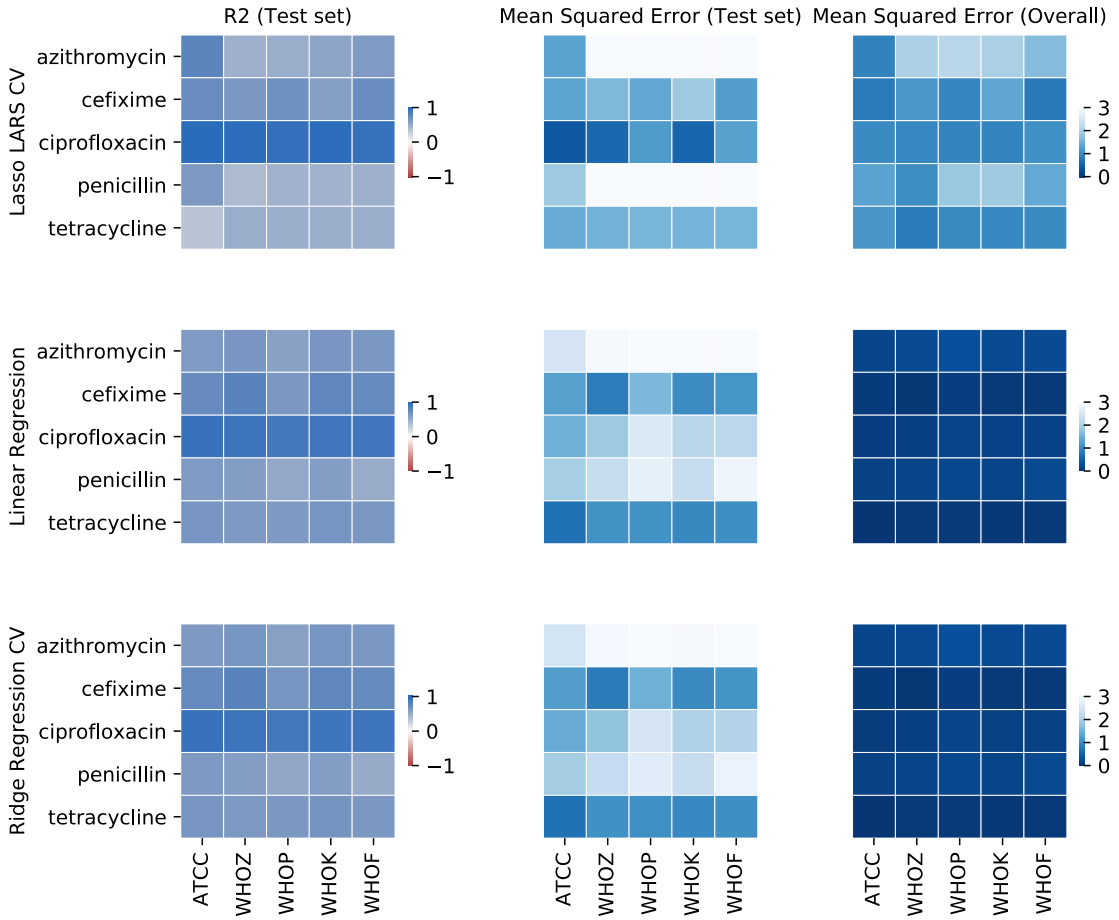


Figure 4-7. *N. gonorrhoeae* MIC prediction models using mutations assessed with evaluation metrics (NG2). Each box represents an antibiotic prediction model generated using a reference genome (x-axis) and algorithm (y-axis). Lasso LARS Cross Validation (CV), linear regression, and ridge regression CV were used to generate MIC prediction models. Breseq was used to identify mutations based on a reference genome and we only included mutations found in two or more samples (≥2). Coefficient of determination (R^2) and mean squared error were used as evaluation metrics with details in Figure 4-1.

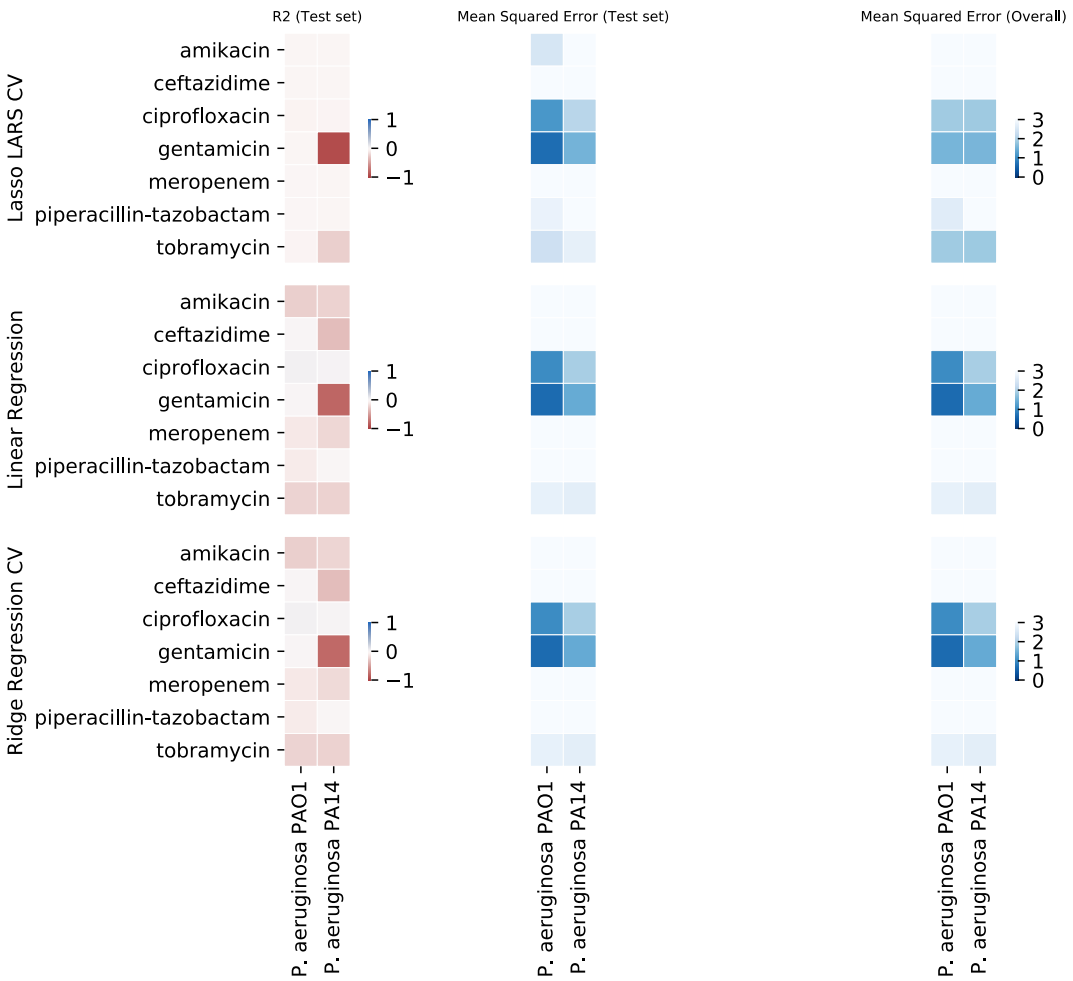


Figure 4-8. *P. aeruginosa* MIC prediction models using mutations assessed with evaluation metrics (PA2). Each box represents an antibiotic prediction model generated using a reference genome (x-axis) and algorithm (y-axis). Lasso LARS Cross Validation (CV), linear regression, and ridge regression CV were used to generate MIC prediction models. Breseq was used to identify mutations based on a reference genome and we only included mutations found in two or more samples (≥2). Coefficient of determination (R^2) and mean squared error were used as evaluation metrics with details in Figure 4-1.

Picking a single algorithm for model selection among antibiotics

Using the R^2 and MSE evaluation metrics, we identified the best filter (for known resistance determinants) or reference sequence (for mutations) and algorithm to use for each dataset, regardless of antibiotic, as a step towards evaluation of ML performance for each antibiotic for each pathogen (Table 4-1). In making these decisions, if genetic features and filtering methods performed similarly in terms of R^2 and MSEs, then the simplest method was chosen. In cases using known resistance determinants, inclusion of RGI features from both chromosomes and plasmids (without the Perfect and Strict representation or Grantham score filter, but only including resistance determinants in ≥ genomes) was best for all pathogens and datasets. If algorithms performed similarly for a specific data set, then the more interpretable algorithm was selected to maximize scientific value, i.e., the algorithm allowing the clearest identification of genetic drivers of resistance and their relative importance to the prediction model (best to worst: ridge regression CV, linear regression, lasso LARS CV). Overall, ridge regression CV and lasso LARS CV were the best for *N. gonorrhoeae* and *P. aeruginosa* MIC prediction models regardless of genetic features, respectively, whereas in *E. coli* ridge regression CV was best suited for known resistance determinants and lasso LARS generated the best MIC prediction models when using mutations (Table 4-1). Linear regression was not suitable for any MIC prediction models.

Table 4-1. The genetic features and algorithms selected to evaluate final MIC prediction models, based on either known resistance determinants or mutations. Known resistance determinants were identified in the chromosome and plasmid (without the Perfect and Strict representation or Grantham score filter, but only including resistance determinants present in ≥2 genomes). The mutation feature set was filtered to only include mutations in ≥2 genomes.

Species	Dataset	Genetic features	Algorithm
E. coli	EC2	Known resistance determinants	Ridge Regression CV
		Mutations (*E. coli* O157:H7 str. Sakai)	Lasso LARS CV
N. gonorrhoeae	NG1	Known resistance determinants	Ridge Regression CV
		Mutations (*N. gonorrhoeae* ATCC 49226)	Ridge Regression CV
	NG2	Known resistance determinants	Ridge Regression CV
		Mutations (*N. gonorrhoeae* ATCC 49226)	Ridge Regression CV
P. aeruginosa	PA2	Known resistance determinants	Lasso LARS CV
		Mutations (*P. aeruginosa* PAO1)	Lasso LARS CV

MIC prediction models are specific to pathogen, antibiotic, and genetic features

After identifying the algorithms and filters best suited for building MIC prediction models for every dataset for either known resistance determinants or mutations (Table 4-1), we generated and tested the models, and determined rates of accurate prediction (i.e., predicted MIC value is within a two-fold dilution factor of the laboratory determined MIC), over prediction (i.e., predicted MIC value is greater than a two-fold dilution factor of the laboratory determined MIC), and under prediction (i.e., predicted MIC value is less than a two-fold dilution factor of the laboratory determined MIC). *E. coli* ampicillin, cefazolin, cefotaxime, ciprofloxacin, gentamicin, and trimethoprim-sulfamethoxazole MIC prediction models performed worse using mutations compared to using known resistance genes, with the largest decrease in accurate predictions for ciprofloxacin

(reduced 92%) and trimethoprim-sulfamethoxazole (reduced 78%) (Figure 4-9). Only the ertapenem MIC prediction model improved using mutations (15% increase in accurate predictions), while meropenem and nitrofurantoin prediction models performed similarly using known resistance determinants or mutations (Figure 4-9). For the *N. gonorrhoeae* datasets, using mutations provided superior MIC prediction models for all antibiotics, with the greatest improvements over known resistance genes for azithromycin in NG1 (increased 58%) and NG2 (increased 55%) (Figure 4-10,4-11). With *P. aeruginosa*, using mutations had very minimal (less than 1%) improvement over known resistance genes for amikacin, ceftazidime, ciprofloxacin, gentamicin, and meropenem MIC prediction and a 10% improvement for piperacillin-tazobactam, while tobramycin MIC prediction had less than 1% improvement based on known resistance determinants compared to mutations (Figure 4-12). Overall, we observed that prediction models were specific to the species, the antibiotic, and genetic features. Selecting between the best performing features (i.e., known resistance determinants vs. mutations), *E. coli* MIC prediction models using known resistance determinants had an average accuracy of 86%, *N. gonorrhoeae* MIC prediction models using mutations had an average accuracy of 98% (NG1) and 97% (NG2), and *P. aeruginosa* MIC prediction models using mutations had an average accuracy of 41% (Supplementary Table 4-2).

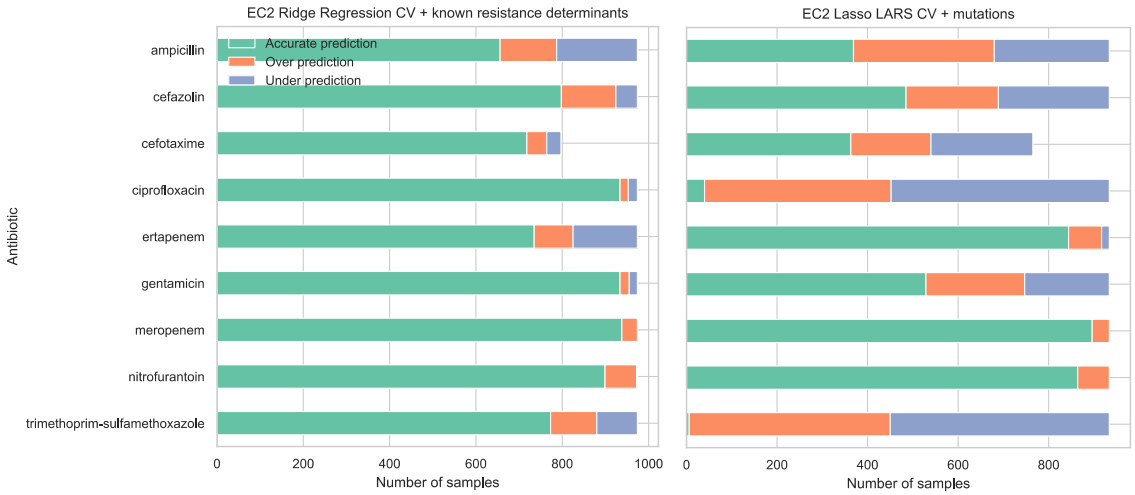


Figure 4-9. Best performing *E. coli* MIC prediction models (EC2). Mutations were generated using *E. coli* O157:H7 str. Sakai. Accurate prediction means the predicted MIC value is within ±1- two-fold dilution factor of the laboratory determined MIC. Over prediction means the predicted MIC value is >1 two-fold dilution factor of the laboratory determined MIC. Under prediction means the predicted MIC value is <1 two-fold dilution factor of the laboratory determined MIC. Not every sample was able to generate mutations using breseq.

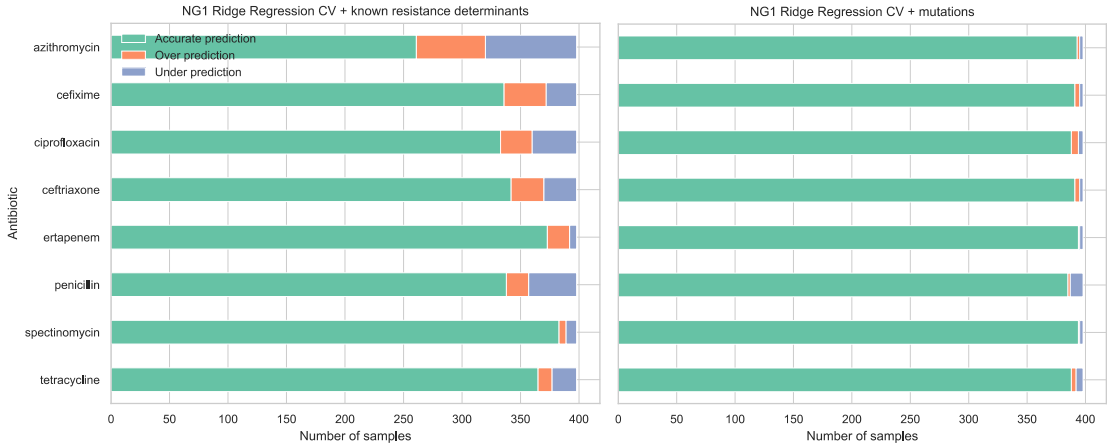


Figure 4-10. Best performing *N. gonorrhoeae* MIC prediction models (NG1). Mutations were generated using *N. gonorrhoeae* ATCC 49226. Accurate prediction means the predicted MIC value is within ±1- two-fold dilution factor of the laboratory determined MIC. Over prediction means the predicted MIC value is >1 two-fold dilution factor of the laboratory determined MIC. Under prediction means the predicted MIC value is <1 two-fold dilution factor of the laboratory determined MIC. Not every sample was able to generate mutations using breseq.

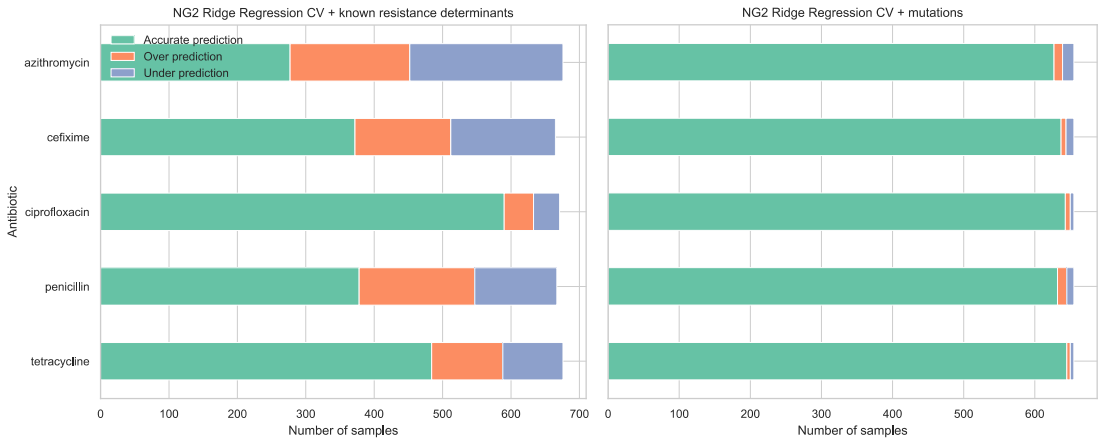


Figure 4-11. Best performing *N. gonorrhoeae* MIC prediction models (NG2). Mutations were generated using *N. gonorrhoeae* ATCC 49226. Accurate prediction means the predicted MIC value is within ±1- two-fold dilution factor of the laboratory determined MIC. Over prediction means the predicted MIC value is >1 two-fold dilution factor of the laboratory determined MIC. Under prediction means the predicted MIC value is <1 two-fold dilution factor of the laboratory determined MIC. Not every sample was able to generate mutations using breseq.

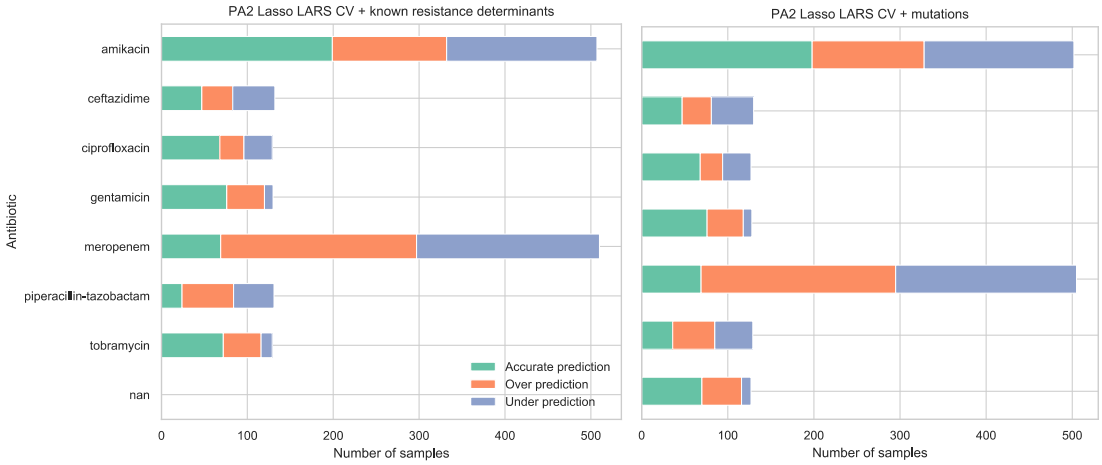


Figure 4-12. Best performing *P. aeruginosa* MIC prediction models (PA2). Mutations were generated using *P. aeruginosa* PAO1 and simulated reads. Accurate prediction means the predicted MIC value is within ±1- two-fold dilution factor of the laboratory determined MIC. Over prediction means the predicted MIC value is >1 two-fold dilution factor of the laboratory determined MIC. Under prediction means the predicted MIC value is <1 two-fold dilution factor of the laboratory determined MIC. Not every sample was able to generate mutations using breseq.

Interpretation of models is only meaningful using known resistance determinants

Our goal was not only to predict MICs, but to interpret the MIC prediction models to better understand the mechanisms driving resistance. For each final model using resistance determinants, we investigated the resistance determinants with the top three coefficients and determined whether there was supporting experimental evidence in CARD or the literature (i.e., whether the resistance determinant is known to confer resistance to a particular antibiotic). We only investigated final models using resistance determinants because our method of interpretation is not suitable for the mutation feature set: mutation feature sets ranged from 31,000 to >1,000,000 mutations and thus only exploring 3 mutations is not informative. New approaches are needed to evaluate these

data holistically. In addition, a brief exploration of these mutations revealed that many are within unannotated “hypothetical proteins”, which makes it challenging to infer mechanism or function without deeper analysis or experimental approaches.

For the *E. coli* MIC prediction models based on RGI, many of the highest coefficients were assigned to resistance determinants that are known to confer resistance to that particular antibiotic (Table 4-2). It is noteworthy that the substrate specificity of many β-lactamases (e.g., CTX-M-15, CMY-2, CTX-M-27) were validated in Chapter 2 (Supplementary Table 2-1). While the resistance determinants with high coefficients for nitrofurantoin, ertapenem, and meropenem resistance may not be causal, they still generated models that performed well (Figure 4-9). In contrast, it is evident that understanding the mechanisms driving resistance for *N. gonorrhoeae* will be difficult for most antibiotics based on model coefficients (Tables 4-3, 4-4), with the exception that the resistance determinants with high coefficients for penicillin and tetracycline resistance prediction are known to confer resistance to these antibiotics (Table 4-3, 4-4). Lastly and matching the general trend of poor phenotype prediction for *P. aeruginosa*, most of the top three resistance determinants with the highest coefficients in *P. aeruginosa* do not confer resistance to the expected antibiotic (Table 4-5).

Table 4-2. Highest coefficients from each *E. coli* (EC2) MIC prediction model. Only the top three resistance determinants with the highest coefficients are included in table. Resistance determinants and whether they confer resistance to an antibiotic was determined using the Comprehensive Antibiotic Resistance Database and/or previously published literature. Asterisk indicates that the substrate specifies were experimentally validated using the Antibiotic Resistance Platform in Chapter 2 (Supplementary Table 2-1).

Antibiotic	Resistance determinant	Coefficient	Does resistance determinant confer resistance to the antibiotic?
ampicillin	TEM-1	1.8	Yes*
	CTX-M-27	1.2	Yes*
	CMY-2	1.1	Yes*
cefazolin	CTX-M-15	2.4	Yes*
	CMY-2	1.9	Yes*
	CTX-M-27	1.4	Yes*
cefotaxime	CTX-M-14	3.7	Likely
	CTX-M-15	3.6	Likely
	CTX-M-27	3.4	Likely
ciprofloxacin	*Escherichia coli* parC	3.0	Yes
	Escherichia coli gyrA	0.7	Yes
	linG	0.3	No
ertapenem	CTX-M-15	0.7	No*
	CMY-2	0.7	Yes*
	CMY-42	0.4	No
gentamicin	AAC(3)-IId	3.9	Yes
	AAC(3)-IIe	3.6	Yes
	AAC(3)-VIa	2.8	Yes
meropenem	tetM	0.3	No
	CTX-M-15	0.2	No*
	dfrA24	0.2	No
nitrofurantoin	dfrA20	0.4	No
	AAC(3)-IIc	0.4	No
	adeL	0.4	No
trimethoprim-sulfamethoxazole	Sul1	1.0	Yes
	dfrA1	1.0	Yes
	drfA14	0.99	Yes

Table 4-3. Highest coefficients from each *N. gonorrhoeae* (NG1) MIC prediction model. Only the top three resistance determinants with the highest coefficients are included in table. Resistance determinants and whether they confer resistance to an antibiotic was determined using the Comprehensive Antibiotic Resistance Database, NG-STAR, and/or previously published literature.

Antibiotic	Resistance determinant	Coefficient	Does resistance determinant confer resistance to the antibiotic?
azithromycin	*Neisseria gonorrhoeae* porin PIB (por)	0.6	Unknown
	Neisseria gonorrhoeae PBP2	0.5	Unlikely
	Neisseria gonorrhoea PBP1	0.4	Unlikely
cefixime	*Neisseria gonorrhoeae* porin PIB (por)	0.8	Unknown
	FEZ-1	0.7	Likely
	Neisseria gonorrhoea PBP1	0.5	Likely
ciprofloxacin	*Neisseria gonorrhoeae* gyrA	1.0	Yes
	Neisseria gonorrhoeae porin PIB (por)	1.0	Unlikely
	Neisseria gonorrhoea PBP1	0.99	Unlikely
ceftriaxone	*Neisseria gonorrhoeae* porin PIB (por)	1.1	Yes
	Neisseria gonorrhoea PBP1	0.6	Unlikely
	Neisseria gonorrhoeae gyrA	0.6	No
ertapenem	*Neisseria gonorrhoeae PBP2*	0.9	No
	FloR	0.8	No
	Neisseria meningitidis PBP2	0.7	No
penicillin	TEM-1	4.7	Yes
	TEM-135	4.3	Yes
	TEM-163	3.5	Yes
spectinomycin	FEZ-1	0.2	No
	Neisseria gonorrhoeae folP	0.1	No
	Neisseria gonorrhoeae PBP2	0.1	No
tetracycline	tetM	3.8	Yes

	rpsJ	2.5	Yes
	Neisseria gonorrhoeae porin PIB (por)	0.2	Yes

Table 4-4. Highest coefficients from each *N. gonorrhoeae* (NG2) MIC prediction model. Only the top three resistance determinants with the highest coefficients are included in table. Resistance determinants and whether they confer resistance to an antibiotic was determined using the Comprehensive Antibiotic Resistance Database, NG-STAR, and/or previously published literature.

Antibiotic	Resistance determinant	Coefficient	Does resistance determinant confer resistance to the antibiotic?
azithromycin	*Neisseria gonorrhoeae* FolP	1.5	No
	RpsJ	1.2	No
	Acinetobacter baumannii AmvA	1.1	No
cefixime	*Neisseria gonorrhoeae parC*	1.3	No
	TriC	1.2	No
	Neisseria gonorrhoeae porin PIB (por)	1.0	Unknown
ciprofloxacin	*Neisseria gonorrhoeae* gyrA	7.8	Yes
	Neisseria gonorrhoeae parC	1.8	Yes
	Neisseria gonorrhoeae porin PIB (por)	1.0	Unlikely
penicillin	TEM-1	2.4	Yes
	Neisseria gonorrhoeae porin PIB (por)	1.3	Yes
	TEM-135	1.2	Yes
tetracycline	tetM	1.9	Yes
	Neisseria gonorrhoeae PBP2	0.9	No
	Neisseria gonorrhoeae gyrA	0.7	No

Table 4-5. Highest coefficients from each *P. aeruginosa* (PA2) MIC prediction model. Only the top three resistance determinants with the highest coefficients are included in table. Resistance determinants and whether they confer resistance to an antibiotic was determined using the Comprehensive Antibiotic Resistance Database, NG-STAR, and/or previously published literature.

Antibiotic	Resistance determinant	Coefficient	Does resistance determinant confer resistance to the antibiotic?
amikacin	carA	0.0	No
	mexL	0.0	Yes
	OXA-301	0.0	No
ceftazidime	OXA-488	0.2	Likely
	cmx	0.0	No
	cmlb	0.0	No
ciprofloxacin	OXA-2	0.1	No
	ANT(3``)-IIa	0.1	No
	OXA-488	0.1	No
gentamicin	cmx	0.0	No
	OXA-9	0.0	No
	IMP-18	0.0	No
meropenem	carA	0.0	No
	mexL	0.0	No
	OXA-301	0.0	No
piperacillin-tazobactam	PDC-2	0.5	No
	Pseudomonas gyrA	0.3	No
	OKP-A-14	0.2	No
tobramycin	ANT(3``)-IIa	0.4	No
	nalC	0.2	No
	cmlB	0.0	No

DISCUSSION

In this study, we built MIC prediction models using different features (chromosome and plasmid-borne resistance determinants or mutations), algorithms (lasso LARS CV, ridge regression CV, and linear regression), and filtering and representation methods (Grantham score and Perfect + Strict representation) for *E. coli, N. gonorrhoeae*, and *P. aeruginosa* datasets. While there have been a number of publications that predict antibiotic MICs (Eyre *et al.*, 2019; Eyre *et al.*, 2017; Hicks *et al.*, 2019; Y. Li *et al.*, 2016; B. J. Metcalf *et al.*, 2016; Nguyen *et al.*, 2018; Nguyen *et al.*, 2020; Nguyen *et al.*, 2019; Pataki *et al.*, 2020), to our knowledge none used different evaluation metrics to assess and then select algorithm and features for accurate MIC prediction. We used two evaluation metrics, the coefficient of determination (R^2) and mean squared error (MSE), to show that while oftentimes their values can be correlated, their values are sometimes decoupled as they are measuring different aspects of the model. For example, we have observed scenarios where R^2 values are all very close to 1 for *P. aeruginosa* MIC prediction using known resistance determinants, but the corresponding MSEs vary from 0 to 1. Alternatively, we have also observed MSE values close to 1 for *N. gonorrhoeae* spectinomycin MIC prediction, while the R^2 values are close to -1. We argue that it is useful to consider both R^2 and MSE as metrics of evaluation as R^2 measures the correlation between the genetic features and the MICs, whereas MSE represents the average of the squared difference between the laboratory determined MICs and the predicted MICs. Challenges arise when the R^2 values conflict with the MSE values and in those scenarios a trade-off between the two will have to be made based on external

criteria, such as accuracy, model simplicity, or interpretability. An inherent limitation is that for dataset EC2 (ertapenem, meropenem, nitrofurantoin) and NG1 (spectinomycin) there are only adjacent MICs (e.g., 2 µg/mL and 4 µg/mL), so if the measure of accuracy is ± one two-fold dilution, then the accuracy should always be close to 100%. In addition, for dataset PA2, there were a combination of phenotypic methods (e.g., Vitek 2 and broth microdilution) that were used, which can decrease accuracies.

In addition to evaluation metrics, only one study has compared use of different genetic features on MIC prediction models (Nguyen *et al.*, 2020). The authors showed that using *k*-mers of core genes compared to whole genome assemblies decreased accuracy and increased error rates (Nguyen *et al.*, 2020). We also show that using different features has an effect on MIC prediction models. Specifically, using chromosome and plasmid-borne resistance determinants together generates MIC models that are better than those using plasmid-borne resistance determinants alone in *E. coli* and *N. gonorrhoeae*. This suggests that resistance is both chromosome and plasmid driven in these two pathogens, aligning with previous evidence for multifactorial causes of resistance (Alekshun & Levy, 2007; Lin *et al.*, 2015). Even though we were unable to identify plasmids since we lacked sequencing reads in the *P. aeruginosa* (PA2) dataset, we only identified one plasmid across all samples in our previous work with a different *P. aeruginosa* dataset (PA1, n=102). The low prevalence of plasmids in *P. aeruginosa* has been observed previously (Plesiat, Alkhalaf, & Michel-Briand, 1988), suggesting that perhaps the PA2 dataset may similarly not have had many plasmids.

We additionally show that using different genetic filtering and representation methods for the resistance determinants have little effect on MIC prediction model performance across all species. Yet, we demonstrate that using mutations improves all antibiotic MIC prediction models for *N. gonorrhoeae*, a few for *E. coli*, and very few for *P. aeruginosa*. AMR in *N. gonorrhoeae* is largely driven by mosaic gene sequences (e.g., *penA*), amino acid substitutions in resistance proteins, and promoter region nucleotide mutations (Unemo & Shafer, 2011, 2014), information which our analysis suggests is better captured by mutations than RGI. In contrast, using known resistance determinants curated in CARD generated better prediction models for *E. coli* antibiotic MICs. This suggests that the curation effort in CARD is sufficient for *E. coli* MIC prediction models and performance may even be improved with further refined curation and detection of mutations in regulatory sequences. For *P. aeruginosa* and *N. gonorrhoeae*, not all previously published resistance determinants are curated in CARD, which is a limitation of using a database dependent method. In addition, single nucleotide polymorphisms in promoter regions of particular genes (e.g., 13-bp deletion in *mtrR* promoter region overexpresses a penicillin efflux pump) have also been associated with resistance (Unemo & Shafer, 2014), however RGI is currently unable to predict mutations in promoter sequences. Yet, unlike for *N. gonorrhoeae*, use of mutations or RGI determinants did not yield accurate prediction of phenotype for *P. aeruginosa*, suggesting that uncaptured biological complexity is hampering effective model construction for this pathogen. It is important to note the *P. aeruginosa* mutations we used were generated from simulated reads via genome assemblies and thus may not reflect the full spectrum of *bona fide P.*

aeruginosa mutations. The overall challenges of generating accurate *P. aeruginosa* antibiotic MIC prediction models can be attributed to its phenotypic plasticity driven by genetic and environmental features (Dotsch *et al.*, 2015). In fact, even just identifying the key resistance determinants could not explain all *P. aeruginosa* resistant phenotypes (Kos *et al.*, 2015). One publication illustrated that using presence/absence of genes and gene expression information generated *P. aeruginosa* resistant/susceptible category prediction models with high sensitivity, but did not produce accurate MIC prediction models (Khaledi *et al.*, 2020). Thus, future studies using transcriptomics to generate MIC prediction models for *P. aeruginosa* would be useful.

Instead of selecting for the most parsimonious method of MIC model prediction, the alternative is to select a feature set and algorithm for each antibiotic-pathogen MIC prediction model. This would require more computational effort and is not what has been conventionally performed in MIC prediction studies. However, it may be a fruitful path to stratify all common antibiotics tested across different species to identify if there are genetic features or algorithms that are beneficial for predicting a particular antibiotic MIC. For example, ciprofloxacin resistance across many pathogens (e.g., *E. coli* and *N. gonorrhoeae*) is known to be driven by substitutions in GyrA, thus perhaps using these known mutations and a particular algorithm would be best for this drug regardless of pathogen (Weigel, Steward, & Tenover, 1998).

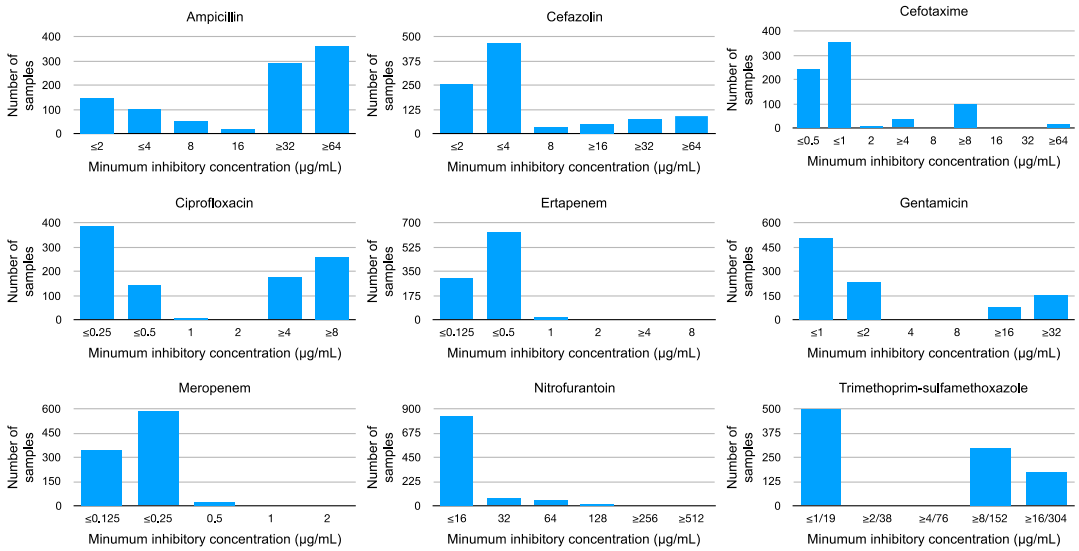
Since we also wanted to interpret MIC prediction models to better understand the mechanisms driving resistance, we used the coefficients assigned to each genetic feature in each prediction model to examine the relative contribution of individual resistance

determinants. The higher value the coefficient, the more important it was for prediction and we examined the three resistance determinants with the highest coefficients for each antibiotic MIC prediction model. For *E. coli*, many of the highest coefficients were assigned to resistance determinants that were known to confer resistance to that particular antibiotic. The substrate activity of many β-lactamases was tested in Chapter 2, which shows the value of having MIC values for each resistance gene to understand the mechanisms driving resistance. We observed that many MIC prediction models assigned the three highest coefficients to resistance determinants that are not known to cause the observed resistance in the literature. This could be explained by other resistance determinants beyond the top three highest coefficients, e.g., NfsA mutations conferring resistance to nitrofurantoin as the fifth highest coefficient in the *E. coli* nitrofurantoin MIC prediction model. Interpretation of these *E. coli* MIC prediction models shows the importance of gene expression studies to understand the drivers of resistance, yet as coefficient values are continuous and not discrete, deciding which to examine and which to ignore is difficult. In *N. gonorrhoeae*, it is evident that understanding the mechanisms driving resistance is difficult for most antibiotics when examining only a few of the highest coefficients for RGI determinants. As use of mutations led to improved *N. gonorrhoeae* models, examination of these coefficients may identify new, polygenic, or under-curated resistance determinants for this pathogen. Lastly and as with overall accuracy, our *P. aeruginosa* MIC prediction models gave few hints on the underlying drivers of resistance. With that being said, if model interpretation is not of value for the end goals, then using other parts of the genome, such as conserved non-AMR genes, has

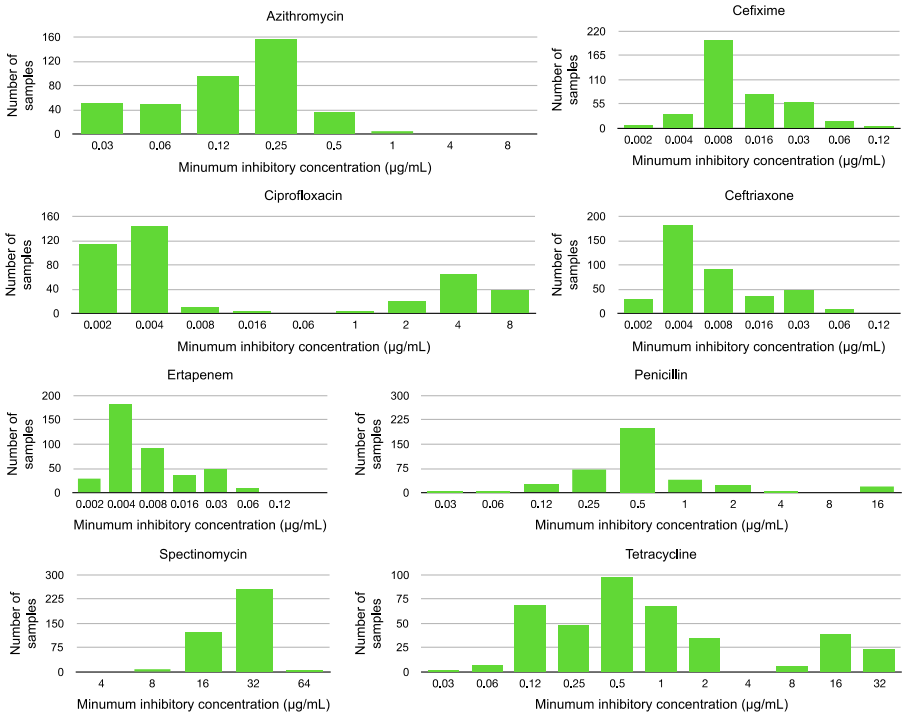
also been shown to be useful for *Mycobacterium tuberculosis, Salmonella enterica,* and *Staphylococcus aureus* resistant/susceptible prediction (Nguyen *et al.*, 2020) and may be useful for *P. aeruginosa* MIC prediction. In general, while we illustrate generation of accurate *N. gonorrhoeae* MIC prediction models and success varied by antibiotic for *E. coli*, pathogens that are known to confer resistance as a result of many interacting factors such as *P. aeruginosa* may benefit from the inclusion of different types of data in combination with machine learning to identify the best parameters to generate accurate MIC prediction models.

SUPPLEMENTARY MATERIAL

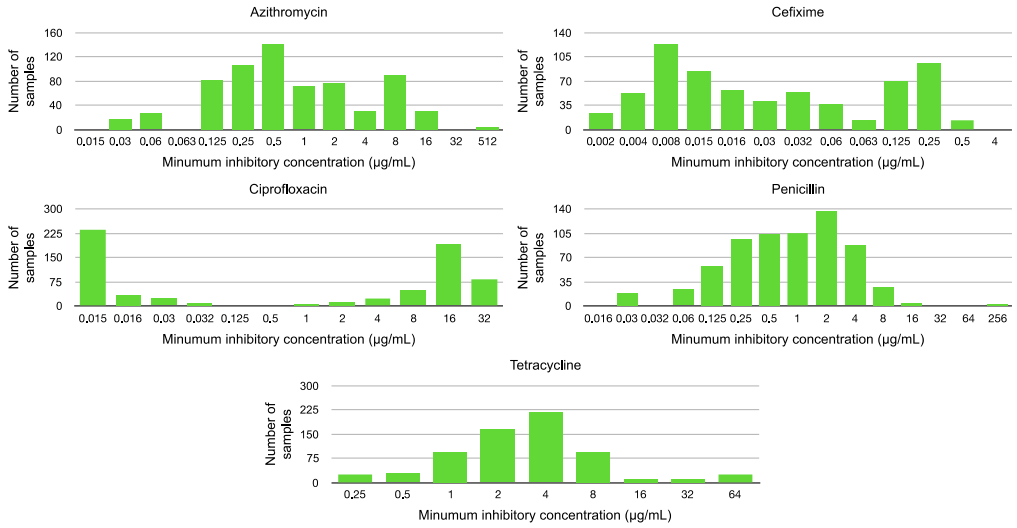
Supplementary Figures



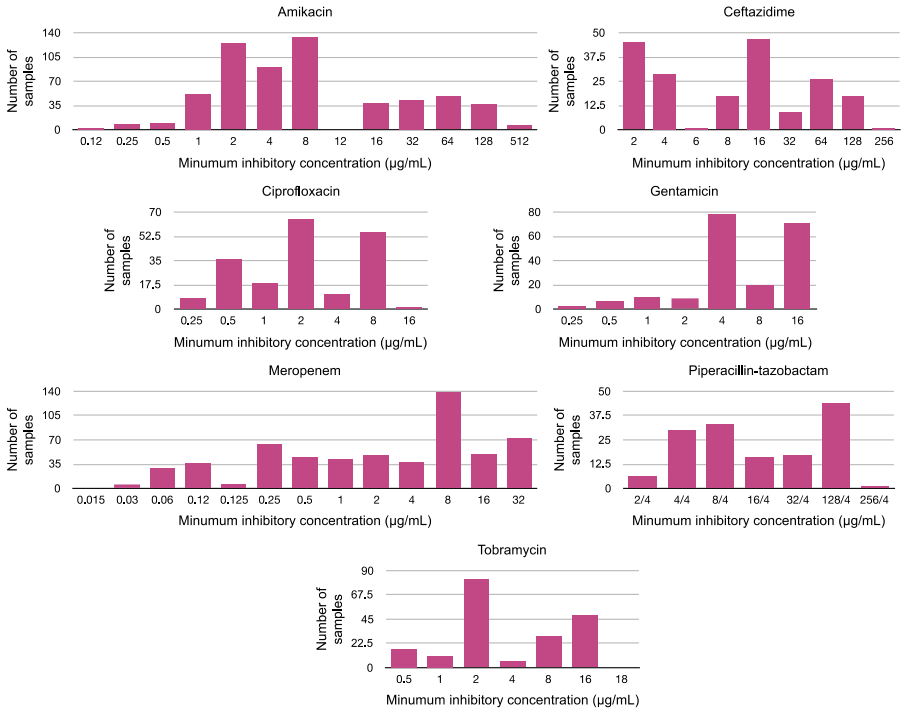
Supplementary Figure 4-1. *E. coli* minimum inhibitory concentration distributions for dataset EC2.



Supplementary Figure 4-2. *N. gonorrhoeae* minimum inhibitory concentration distributions for dataset NG1.



Supplementary Figure 4-3. *N. gonorrhoeae* minimum inhibitory concentration distributions for dataset NG2.



Supplementary Figure 4-4. *P. aeruginosa* minimum inhibitory concentration distributions for dataset PA2.

Supplementary Tables

Supplementary Table 4-1. Mutations identified in simulated reads of PA2 dataset. Mutations were generated using breseq and two reference sequences (*P. aeruginosa* PAO1 and *P. aeruginosa* PA14)

Antibiotic	Number of genomes	Number of mutations (*P. aeruginosa* PAO1)	Number of mutations (*P. aeruginosa* PA14)
amikacin	502	393,606	425,937
ceftazidime	182		
ciprofloxacin	129		
gentamicin	128		
meropenem	534		
piperacillin-tazobactam	155		
tobramycin	127		

Supplementary Table 4-2. Accuracy for all MIC prediction models. Known resistance determinants were identified in the chromosome and plasmid (without the Perfect and Strict representation, Grantham score filter, or only including resistance determinants present in ≥2 genomes). The mutation feature set is filtered to only include mutations in ≥2 genomes. Highlighted in yellow are the models that performed the best for a given dataset.

Species	Dataset	Genetic features	Antibiotic	Accuracy (%)
E. coli	EC2	Known resistance determinants	ampicillin	67
			cefazolin	82
			cefotaxime	90
			ciprofloxacin	96
			ertapenem	75
			gentamicin	96
			meropenem	96
			nitrofurantoin	92
			trimethoprim-sulfamethoxazole	79
		Mutations (*E. coli* O157:H7 str. Sakai)	ampicillin	40
			cefazolin	52
			cefotaxime	47
			ciprofloxacin	4
			ertapenem	90
			gentamicin	57
			meropenem	96
			nitrofurantoin	93
			trimethoprim-sulfamethoxazole	1
N. gonorrhoeae	NG1	Known resistance determinants	azithromycin	66
			cefixime	84
			ciprofloxacin	84
			ceftriaxone	86
			ertapenem	94
			penicillin	85
			spectinomycin	96
			tetracycline	92
		Mutations (*N. gonorrhoeae* ATCC 49226)	azithromycin	99
			cefixime	98
			ciprofloxacin	97
			ceftriaxone	98
			ertapenem	99
			penicillin	97
			spectinomycin	99
			tetracycline	97
	NG2	Known resistance determinants	azithromycin	41
			cefixime	56

			ciprofloxacin	88
			penicillin	57
			tetracycline	72
		Mutations (*N. gonorrhoeae* ATCC 49226)	azithromycin	96
			cefixime	97
			ciprofloxacin	98
			penicillin	96
			tetracycline	98
P. aeruginosa	PA2	Known resistance determinants	amikacin	39
			ceftazidime	36
			ciprofloxacin	53
			gentamicin	58
			meropenem	14
			piperacillin-tazobactam	18
			tobramycin	56
		Mutations (*P. aeruginosa* PAO1)	amikacin	39
			ceftazidime	36
			ciprofloxacin	54
			gentamicin	59
			meropenem	14
			piperacillin-tazobactam	28
			tobramycin	55

CHAPTER FIVE: Discussion and future directions

Discussion

Antimicrobial therapy timing is currently guided by the severity of an infection (Leekha, Terrell, & Edson, 2011). In critically ill patients, empiric therapy should be started in parallel with the collection of specimens for antibiotic susceptibility testing (AST) in a clinical diagnostic laboratory. Where the patient is more stable, antimicrobial therapy should be withheld until definitive therapy (informed by AST) is determined. Since AST results are not typically available for a couple of days, empiric therapy usually includes broad-spectrum antibiotics. However, due to the increased prevalence of antimicrobial resistance (AMR), first-line antibiotics often have to be switched for second- or third-line antibiotics that are more expensive, potentially ineffective, and more toxic (CDC, 2013; Prestinaci, Pezzotti, & Pantosti, 2015). The available arsenal of antibiotics is diminishing and the battle with AMR is exacerbated by lack of funding of antibiotic development (Ventola, 2015). Thus, there is a clear need for rapid alternatives to conventional phenotypic AST for individual patients and antimicrobial stewardship (Smith & Kirby, 2019).

Genomics-based methods are a potential alternative to phenotypic AST because they bypass the need for bacterial culturing. Prior to the advent of genomic technologies, the identification and characterization of individual resistance determinants, e.g., penicillinase (the first β-lactamase discovered), required carefully planned experiments (Abraham & Chain, 1940). Sixty years later in 2021, there are thousands of described β-lactamases that span four classes curated in CARD and ResFinder (Alcock *et al.*, 2020; Bortolaia *et al.*, 2020). Increasingly these are described based on gene sequence alone,

without biochemical characterization (Feldgarden *et al.*, 2019). This increase in identification of new resistance determinants is not only a trend for β-lactamases, but many different classes of resistance determinants. A large source for this increase in data is attributable to the study of bacterial genomics, where bacterial isolates are sequenced and annotated based on sequence similarity. While this increase in data is helpful for genomic epidemiology, high sequence similarity does not necessarily infer the same substrate specificity. For example, a few amino acid substitutions are capable of changing the substrate specificity for TEM-1 (Stojanoski *et al.*, 2015). We have also shown that three CTX-M β-lactamases have broader substrate specificities than reported in the literature (Tsang *et al.*, 2021). This is particularly challenging for use of the resistome to predict resistance phenotype because substrate specificities are often inferred from experimental validation on a similar (but not identical) resistance determinant. Even with new technologies that allow for a broad data-driven resistome analysis, we inherently still rely on experimental validation to increase our depth of knowledge of the resistome. As I have shown in Chapter 2, identifying substrate specificities of known resistance determinants can improve rules-based methods of AMR prediction. Thus, resistome research can be advanced by use of genomics data in an inductive approach to produce specific hypotheses ready for testing via use of deductive experiments.

While next-generation sequencing technologies are being slowly incorporated into clinical microbiology and public health laboratories, researchers have continued to generate and share a plethora of genomics data associated with resistance phenotypes.

The increase in data has allowed a shift from rules-based to machine learning algorithms for AMR prediction. The goal of machine learning is not to identify causative AMR genotype-phenotype relationships, although that is possible using ML, but to build a model that accurately predicts AMR phenotypes using genetic features. The benefit of using interpretable machine learning algorithms is then to generate an accurate model and potentially discover new AMR genotype-phenotype relationships. While we identified novel AMR genotype-phenotype relationships between known resistance genes and resistance phenotypes, others have illustrated that use of *k*-mers as genetic features can potentially identify new mechanisms of resistance (Davis *et al.*, 2016; Drouin *et al.*, 2016; Kavvas *et al.*, 2018; Nguyen *et al.*, 2018; Nguyen *et al.*, 2019). To our knowledge, no new resistance determinants have been characterized through using machine learning models, but there are genetic features in these publications that could be investigated (Aytan-Aktug *et al.*, 2021; Davis *et al.*, 2016; Drouin *et al.*, 2016; Kavvas *et al.*, 2018; Nguyen *et al.*, 2018; Nguyen *et al.*, 2019). Conversely, it could be argued from a clinical perspective that understanding genetic features driving resistance does not matter as long as the prediction models perform as well or more improved than AST methods with head-to-head practical advantages, e.g., sensitivity, specificity, cost, and overall turn-around time. With that being said, the current workflow for genomics sequencing still requires pure cultures of an isolated microorganism and therefore the slow turnaround time for genomics-based diagnostics is still an issue. Perhaps, AMR prediction models can be extended towards culture-free metagenomics sequencing, with the goal being to predict the resistance phenotype from sequences extracted directed from a clinical sample. While

there are still challenges and opportunities ahead, there is benefit in developing accurate AMR prediction models in parallel with inspecting the model, considering the new knowledge about the resistome to be gained and the research avenues it may open.

Since there are currently no standardized methods for building AMR prediction models, researchers are applying and testing new parameters to determine a potential gold standard. Even if the parameters examined reduce the performance of the AMR prediction models, the results should be shared for this new field to organically grow and improve. For example, one study even deliberately excludes all AMR related genes, to only include conserved genes within a species in their AMR prediction models (Nguyen *et al.*, 2020). Even though these models were less accurate than using whole genomes or AMR determinants, they still illustrate that it is possible to achieve ≥80% accuracy across 4 species using this information. Similarly, using partial genome alignments (that predominantly do not encode AMR-related functions) in AMR prediction models achieved at least 70% accuracy across 4 species (Aytan-Aktug *et al.*, 2021). These studies show while the genomic information we have for a given dataset may be finite, there are still innovative methods to partition it for AMR prediction.

Future Directions

I) Even within our current understanding of the resistome, we have incomplete knowledge about known resistance determinants, i.e., substrate specificities. The AMR research community should ambitiously determine the antibiotic minimum inhibitory concentrations of all known resistance determinants against a broad and standardized collection of antibiotics using the Antibiotic Resistance Platform (G. Cox *et al.*, 2017) or similar methods. The idea is to identify the substrate specificity of each resistance gene by controlling the gene copy number in a hyperpermeable, efflux-deficient mutant of *E. coli* that has increased sensitivity to antibiotics. This system isolates the specific minimum inhibitory concentration impact of the gene, but this is one step closer to understanding the impact of that gene in different pathogens. Ultimately, this would shift our annotations of the resistome to individual antibiotics, rather than antibiotic classes.

II) Continue testing the effect of using a variety of parameters for AMR prediction models. Since this is still an active field, there is plenty room for discovery. We should additionally further use of genomic data by incorporating *in silico* structure prediction methods to deduce substrate specify or function. Along with using the genetic aspect of the resistome, there can be further exploration using transcriptomics data for AMR prediction. Different machine learning and deep learning algorithms can continue to be tested alongside a wider variety of features.

Concluding remarks

Innovations in next-generation sequencing have generated a plethora of genomics information that has both guided discovery and potentially masked unknown information in the resistome. Yet, we can still leverage this technology by applying machine learning to predict antibiotic resistance phenotypes. We demonstrate that intricate AMR genotype-phenotype relationships can be modelled for *Escherichia coli* and *Neisseria gonorrhoea,* but these methods require improvement for *Pseudomonas aeruginosa.* This work stands upon the shoulders of giants in this field as a step towards development of genomics-based diagnostic microbiology. We continue to adhere to the euphemism of George E. P. Box, "all models are wrong, but some are useful" (Box, 1976), in that AMR prediction models are not expected to be a complete reflection of the resistome, but hopefully some are useful for AMR phenotype prediction.

REFERENCES

Aanensen, D. M., Feil, E. J., Holden, M. T., Dordel, J., Yeats, C. A., Fedosejev, A., . . . European, S. R. L. W. G. (2016). Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive *Staphylococcus aureus* in Europe. *mBio, 7*(3). doi:10.1128/mBio.00444-16

Abraham, E. P., & Chain, E. (1940). An enzyme from bacteria able to destroy penicillin. *Nature, 146*(3713), 837.

Alam, M. T., Petit, R. A., Crispell, E. K., Thornton, T. A., Conneely, K. N., Jiang, Y., . . . Read, T. D. (2014). Dissecting Vancomycin-Intermediate Resistance in *Staphylococcus aureus* Using Genome-Wide Association. *Genome Biology and Evolution, 6*, 1174-1185. doi:10.1093/gbe/evu092

Alanazi, M. Q., Alqahtani, F. Y., & Aleanizy, F. S. (2018). An evaluation of *E. coli* in urinary tract infection in emergency department at KAMC in Riyadh, Saudi Arabia: Retrospective study. *Annals of Clinical Microbiology and Antimicrobials, 17*, 3. doi:10.1186/s12941-018-0255-z

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., . . . McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*. doi:10.1093/nar/gkz935

Molecular Mechanisms of Antibacterial Multidrug Resistance, 128 1037-1050 (2007).

Ames, S. K., Hysom, D. A., Gardner, S. N., Lloyd, G. S., Gokhale, M. B., & Allen, J. E. (2013). Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics, 29*, 2253. doi:10.1093/BIOINFORMATICS/BTT389

Andrews, S. (2010, August 1, 2019). FastQC: A quality control tool for high throughput sequence data.

Anonymous. (2019). Antibiotic susceptibility diagnostics for the future. *Nature Microbiology, 4*, 1603. doi:10.1038/s41564-019-0577-4

Arredondo-Alonso, S., Willems, R. J., van Schaik, W., & Schurch, A. C. (2017). On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom, 3*(10), e000128. doi:10.1099/mgen.0.000128

Avershina, E., Sharma, P., Taxt, A. M., Singh, H., Frye, S. A., Paul, K., . . . Ahmad, R. (2021). AMR-Diag: Neural network based genotype-to-phenotype prediction of resistance towards beta-lactams in *Escherichia coli* and *Klebsiella pneumoniae*. *Comput Struct Biotechnol J, 19*, 1896-1906. doi:10.1016/j.csbj.2021.03.027

Aytan-Aktug, D., Clausen, P., Bortolaia, V., Aarestrup, F. M., & Lund, O. (2020). Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. *mSystems, 5*(1). doi:10.1128/mSystems.00774-19

Aytan-Aktug, D., Nguyen, M., Clausen, P., Stevens, R. L., Aarestrup, F. M., Lund, O., & Davis, J. J. (2021). Predicting Antimicrobial Resistance Using Partial Genome Alignments. *mSystems*, e0018521. doi:10.1128/mSystems.00185-21

Baker, S. J., Payne, D. J., Rappuoli, R., & De Gregorio, E. (2018). Technologies to address antimicrobial resistance. *Proc Natl Acad Sci U S A, 115*(51), 12887-12895. doi:10.1073/pnas.1717160115

Banerjee, R., & Humphries, R. (2021). Rapid Antimicrobial Susceptibility Testing Methods for Blood Cultures and Their Clinical Impact. *Front Med (Lausanne), 8*, 635831. doi:10.3389/fmed.2021.635831

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology, 19*, 455-477. doi:10.1089/cmb.2012.0021

Belaaouaj, A., Lapoumeroulie, C., CaniÃ§a, M. M., Vedel, G. r., NÃ©vot, P., Krishnamoorthy, R., & Paul, G. r. (1994). Nucleotide sequences of the genes coding for the TEM-like β-lactamases IRT-1 and IRT-2 (formerly called TRI-1 and TRI-2). *FEMS Microbiology Letters*. doi:10.1111/j.1574-6968.1994.tb07010.x

Belland, R. J., Morrison, S. G., Ison, C., & Huang, W. M. (1994). *Neisseria gonorrhoeae* acquires mutations in analogous regions of gyrA and parC in fluoroquinolone-resistant isolates. *Mol Microbiol, 14*(2), 371-380. doi:10.1111/j.1365-2958.1994.tb01297.x

Boisvert, S., Laviolette, F., & Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol, 17*(11), 1519-1533. doi:10.1089/cmb.2009.0238

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England), 30*, 2114-2120. doi:10.1093/bioinformatics/btu170

Boolchandani, M., D'Souza, A. W., & Dantas, G. (2019). Sequencing-based methods and resources to study antimicrobial resistance. *Nat Rev Genet, 20*(6), 356-370. doi:10.1038/s41576-019-0108-4

Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., . . . Aarestrup, F. M. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother, 75*(12), 3491-3500. doi:10.1093/jac/dkaa345

Bottery, M. J., Pitchford, J. W., & Friman, V. P. (2021). Ecology and evolution of antimicrobial resistance in bacterial communities. *ISME J, 15*(4), 939-948. doi:10.1038/s41396-020-00832-7

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*. doi:10.2307/2286841

Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., . . . Iqbal, Z. (2015). Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis. Nature Communications, 6*, 10063. doi:10.1038/ncomms10063

Brankin, A. E., & Fowler, P. W. (2019). Predicting Resistance Is (Not) Futile. *ACS Central Science, 5*, 1312-1314. doi:10.1021/acscentsci.9b00791

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32. doi:10.1023/A:1010933404324

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods, 12*(1), 59-60. doi:10.1038/nmeth.3176

Burge, S., Attwood, T. K., Bateman, A., Berardini, T. Z., Cherry, M., O'Donovan, C., . . . Gaudet, P. (2012). Biocurators and biocuration: surveying the 21st century challenges. *Database (Oxford), 2012*, bar059. doi:10.1093/database/bar059

Burnham, C.-A. D., Leeds, J., Nordmann, P., O'Grady, J., & Patel, J. (2017). Diagnosing antimicrobial resistance. *Nature Reviews Microbiology, 15*, 697-703. doi:10.1038/nrmicro.2017.103

Bush, K. (2013). Proliferation and significance of clinically relevant beta-lactamases. *Ann N Y Acad Sci, 1277*, 84-90. doi:10.1111/nyas.12023

Antibiotics in the clinical pipeline in October 2019, 73, Springer Nature 329-364 (2020).

Cantu, C., Huang, W., & Palzkill, T. (1997). Cephalosporin substrate specificity determinants of TEM-1 β-lactamase. *Journal of Biological Chemistry*. doi:10.1074/jbc.272.46.29144

Urinary tract infections in women: Diagnosis and management in primary care, 332 94-97 (2006).

Cassini, A., Högberg, L. D., Plachouras, D., Quattrocchi, A., Hoxha, A., Simonsen, G. S., . . . Hopkins, S. (2019). Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *The Lancet. Infectious diseases, 19*, 56-66. doi:10.1016/S1473-3099(18)30605-4

CDC. (2013). *Antibiotic resistance threats in the United States, 2013*. Retrieved from https://stacks.cdc.gov/view/cdc/20705

CDC. (2019). *Antibiotic Resistance Threats in the United States, 2019.* Retrieved from Atlanta, GA, USA: http://dx.doi.org/10.15620/cdc:82532

Chan, K.-G. (2016). Whole-genome sequencing in the prediction of antimicrobial resistance. *Expert review of anti-infective therapy, 14*, 617-619. doi:10.1080/14787210.2016.1193005

Chavez-Jacobo, V. M., Hernandez-Ramirez, K. C., Romo-Rodriguez, P., Perez-Gallardo, R. V., Campos-Garcia, J., Gutierrez-Corona, J. F., . . . Ramirez-Diaz, M. I. (2018). CrpP Is a Novel Ciprofloxacin-Modifying Enzyme Encoded by the *Pseudomonas aeruginosa* pUM505 Plasmid. *Antimicrob Agents Chemother, 62*(6). doi:10.1128/AAC.02629-17

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chen, M. L., Doddi, A., Royer, J., Freschi, L., Schito, M., Ezewudo, M., . . . Farhat, M. (2019). Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. *EBioMedicine, 43*, 356-369. doi:10.1016/j.ebiom.2019.04.016

Chen, Y., Sun, M., Wang, M., Lu, Y., & Yan, Z. (2014). Dissemination of IMP-6-producing *Pseudomonas aeruginosa* ST244 in multiple cities in China. *European*

journal of clinical microbiology & infectious diseases: official publication of the European Society of Clinical Microbiology, 33, 1181-1187. doi:10.1007/s10096-014-2063-5

Chiou, J., Leung, T. Y. C., & Chen, S. (2014). Molecular mechanisms of substrate recognition and specificity of New Delhi metallo-β-lactamase. *Antimicrobial Agents and Chemotherapy*. doi:10.1128/AAC.01977-13

Chowdhury, A. S., Call, D. R., & Broschat, S. L. (2019). Antimicrobial Resistance Prediction for Gram-Negative Bacteria via Game Theory-Based Feature Evaluation. *Scientific Reports, 9*, 14487. doi:10.1038/s41598-019-50686-z

Clausen, P., Aarestrup, F. M., & Lund, O. (2018). Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics, 19*(1), 307. doi:10.1186/s12859-018-2336-6

Clausen, P. T. L. C., Zankari, E., Aarestrup, F. M., & Lund, O. (2016). Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *The Journal of antimicrobial chemotherapy, 71*, 2484-2488. doi:10.1093/jac/dkw184

CLSI. (2018). Performance Standards for Antimicrobial Susceptibility Testing.

Coelho, J. R., Carriço, J. A., Knight, D., Martínez, J.-L. L., Morrissey, I., Oggioni, M. R., & Freitas, A. T. (2013). The Use of Machine Learning Methodologies to Analyse Antibiotic and Biocide Susceptibility in Staphylococcus aureus. *PLoS One, 8*, e55582. doi:10.1371/journal.pone.0055582

Coll, F., McNerney, R., Preston, M. D., Guerra-Assuncao, J. A., Warry, A., Hill-Cawthorne, G., . . . Clark, T. G. (2015). Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med, 7*(1), 51. doi:10.1186/s13073-015-0164-0

Council of Canadian Academies. (2019). *When Antibiotics Fail*. Retrieved from Ottawa, ON: https://cca-reports.ca/wp-content/uploads/2018/10/When-Antibiotics-Fail-1.pdf

Cox, C. E. (1973). Cefazolin therapy of urinary tract infections. *Journal of Infectious Diseases, 128*, S378-S397. doi:10.1093/infdis/128.Supplement_2.S397

Cox, G., Sieron, A., King, A. M., De Pascale, G., Pawlowski, A. C., Koteva, K., & Wright, G. D. (2017). A Common Platform for Antibiotic Dereplication and Adjuvant Discovery. *Cell chemical biology, 24*, 98-109. doi:10.1016/j.chembiol.2016.11.011

Cox, G., & Wright, G. D. (2013). Intrinsic antibiotic resistance: Mechanisms, origins, challenges and solutions. *International Journal of Medical Microbiology, 303*, 287-292. doi:10.1016/J.IJMM.2013.02.009

Crofts, T. S., Gasparrini, A. J., & Dantas, G. (2017). Next-generation approaches to understand and combat the antibiotic resistome. *Nature Reviews Microbiology, 15*, 422-434. doi:10.1038/nrmicro.2017.28

Cusack, T. P., Ashley, E. A., Ling, C. L., Rattanavong, S., Roberts, T., Turner, P., . . . Dance, D. A. B. (2019). Impact of CLSI and EUCAST breakpoint discrepancies on reporting of antimicrobial susceptibility and AMR surveillance. *Clin Microbiol Infect, 25*(7), 910-911. doi:10.1016/j.cmi.2019.03.007

Cusack, T. P., Ashley, E. A., Ling, C. L., Roberts, T., Turner, P., Wangrangsimakul, T., & Dance, D. A. B. (2019). Time to switch from CLSI to EUCAST? A Southeast Asian perspective. *Clin Microbiol Infect, 25*(7), 782-785. doi:10.1016/j.cmi.2019.03.016

Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., . . . Stevens, R. (2016). Antimicrobial Resistance Prediction in PATRIC and RAST. *Scientific reports, 6*, 27930. doi:10.1038/srep27930

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R. M. R., Butler, R. M. R., . . . Stevens, R. (2020). The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. *Nucleic Acids Research, 48*, D606-D612. doi:10.1093/nar/gkz943

Day, M. R., Doumith, M., Do Nascimento, V., Nair, S., Ashton, P. M., Jenkins, C., . . . Godbole, G. (2018). Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of *Salmonella enterica* serovars Typhi and Paratyphi. *J Antimicrob Chemother, 73*(2), 365-372. doi:10.1093/jac/dkx379

de Kraker, M. E. A., Stewardson, A. J., & Harbarth, S. (2016). Will 10 Million People Die a Year due to Antimicrobial Resistance by 2050? *PLoS Medicine, 13*, 1002184. doi:10.1371/journal.pmed.1002184

Deatherage, D. E., & Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods in molecular biology (Clifton, N.J.), 1151*, 165-188. doi:10.1007/978-1-4939-0554-6_12

Demczuk, W., Martin, I., Peterson, S., Bharat, A., Van Domselaar, G., Graham, M., . . . Mulvey, M. R. (2016). Genomic Epidemiology and Molecular Resistance Mechanisms of Azithromycin-Resistant *Neisseria gonorrhoeae* in Canada from 1997 to 2014. *Journal of clinical microbiology, 54*, 1304-1313. doi:10.1128/JCM.03195-15

Demczuk, W., Martin, I., Sawatzky, P., Allen, V., Lefebvre, B., Hoang, L., . . . Mulvey, M. R. (2020). Equations to predict antimicrobial MICs in *Neisseria gonorrhoeae* using molecular antimicrobial resistance determinants. *Antimicrobial Agents and Chemotherapy*. doi:10.1128/AAC.02005-19

Deng, X., Memari, N., Teatero, S., Athey, T., Isabel, M., Mazzulli, T., . . . Gubbay, J. B. (2016). Whole-genome Sequencing for Surveillance of Invasive Pneumococcal Diseases in Ontario, Canada: Rapid Prediction of Genotype, Antibiotic Resistance and Characterization of Emerging Serotype 22F. *Front Microbiol, 7*, 2099. doi:10.3389/fmicb.2016.02099

Dotsch, A., Schniederjans, M., Khaledi, A., Hornischer, K., Schulz, S., Bielecka, A., . . . Haussler, S. (2015). The *Pseudomonas aeruginosa* Transcriptional Landscape Is Shaped by Environmental Heterogeneity and Genetic Variation. *mBio, 6*(4), e00749. doi:10.1128/mBio.00749-15

Doyle, R. M., O'sullivan, D. M., Aller, S. D., Bruchmann, S., Clark, T., Pelegrin, A. C., . . . Harris, K. A. (2020). Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: An inter-laboratory study. *Microbial Genomics, 6*. doi:10.1099/mgen.0.000335

Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., Loo, V. G., . . . Corbeil, J. (2016). Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, *17*, 754. doi:10.1186/s12864-016-2889-6

Drouin, A., Letarte, G., Raymond, F., Marchand, M., Corbeil, J., & Laviolette, F. (2019). Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci Rep*, *9*(1), 4071. doi:10.1038/s41598-019-40561-2

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of statistics*, *32*(2), 407-499. doi:10.1214/009053604000000067

Empel, J., Filczak, K., Mrowka, A., Hryniewicz, W., Livermore, D. M., & Gniadkowski, M. (2007). Outbreak of *Pseudomonas aeruginosa* Infections with PER-1 Extended-Spectrum β-Lactamase in Warsaw, Poland: Further Evidence for an International Clonal Complex. *Journal of Clinical Microbiology*, *45*, 2829-2834. doi:10.1128/JCM.00997-07

EUCAST. (2015). EUCAST. *EUCAST*.

Eyre, D. W., Golparian, D., Unemo, M. (2019). Prediction of Minimum Inhibitory Concentrations of Antimicrobials for *Neisseria gonorrhoeae* Using Whole-Genome Sequencing. *Humana*, New York, NY, 59-76.

Eyre, D. W., Silva, D. D., Cole, K., Peters, J., Cole, M. J., Grad, Y. H., . . . Paul, J. (2017). WGS to predict antibiotic MICs for *Neisseria gonorrhoeae*. *Journal of Antimicrobial Chemotherapy*, *72*, 1937-1947. doi:10.1093/jac/dkx067

Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., . . . Klimke, W. (2019). Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob Agents Chemother*, *63*(11). doi:10.1128/AAC.00483-19

Fernández, L., & Hancock, R. E. W. (2012). Adaptive and mutational resistance: Role of porins and efflux pumps in drug resistance. *Clinical Microbiology Reviews*, *25*, 661-681. doi:10.1128/CMR.00043-12

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., . . . et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, *269*(5223), 496-512. doi:10.1126/science.7542800

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., . . . Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, *270*(5235), 397-403. doi:10.1126/science.270.5235.397

Freund, Y., & Schapire, R. E. (1996). *Experiments with a new boosting algorithm.* Paper presented at the Machine Learning: Proceedings of the Thirteenth International Conference.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*(5), 1189-1232, 1144.

Geisinger, E., & Isberg, R. R. (2017). Interplay between antibiotic resistance and virulence during Disease promoted by multidrug-resistant bacteria. *Journal of Infectious Diseases*, *215*, S9-S17. doi:10.1093/infdis/jiw402

Ghodbane, R., Raoult, D., & Drancourt, M. (2014). Dramatic reduction of culture time of Mycobacterium tuberculosis. *Sci Rep, 4*, 4236. doi:10.1038/srep04236

Gibson, M. K., Forsberg, K. J., & Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal, 9*, 207-216. doi:10.1038/ismej.2014.106

Girlich, D., Naas, T., & Nordmann, P. (2004). Biochemical Characterization of the Naturally Occurring Oxacillinase OXA-50 of Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy, 48*, 2043-2048. doi:10.1128/AAC.48.6.2043-2048.2004

Goig, G. A., Blanco, S., Garcia-Basteiro, A. L., & Comas, I. (2020). Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. *BMC Biol, 18*(1), 24. doi:10.1186/s12915-020-0748-z

Golparian, D., Rose, L., Lynam, A., Mohamed, A., Bercot, B., Ohnishi, M., . . . Unemo, M. (2018). Multidrug-resistant *Neisseria gonorrhoeae* isolate, belonging to the internationally spreading japanese FC428 clone, with ceftriaxone resistance and intermediate resistance to azithromycin, Ireland, August 2018. *Eurosurveillance*. doi:10.2807/1560-7917.ES.2018.23.47.1800617

Gordon, N. C., Price, J. R., Cole, K., Everitt, R., Morgan, M., Finney, J., . . . Golubchik, T. (2014). Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *Journal of clinical microbiology, 52*, 1182-1191. doi:10.1128/JCM.03117-13

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science, 185*, 862-864. doi:10.1126/science.185.4154.862

Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., & Rolain, J.-M. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial agents and chemotherapy, 58*, 212-220. doi:10.1128/AAC.01310-13

Hakenbeck, R., Bruckner, R., Denapaite, D., & Maurer, P. (2012). Molecular mechanisms of beta-lactam resistance in *Streptococcus pneumoniae*. *Future Microbiol, 7*(3), 395-410. doi:10.2217/fmb.12.2

Hanson, N. D., & Sanders, C. C. (1999). Regulation of inducible AmpC beta-lactamase expression among *Enterobacteriaceae*. *Curr Pharm Des, 5*(11), 881-894.

Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C. A., Frimodt-Moller, N., & Aarestrup, F. M. (2014). Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J Clin Microbiol, 52*(1), 139-146. doi:10.1128/JCM.02452-13

Hicks, A. L., Wheeler, N., Sánchez-Busó, L., Rakeman, J. L., Harris, S. R., & Grad, Y. H. (2019). Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLOS Computational Biology, 15*, e1007349. doi:10.1371/journal.pcbi.1007349

Hjort, K., Nicoloff, H., & Andersson, D. I. (2016). Unstable tandem gene amplification generates heteroresistance (variation in resistance within a population) to colistin in *Salmonella enterica*. *Mol Microbiol, 102*(2), 274-289. doi:10.1111/mmi.13459

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hombach, M., Mouttet, B., & Bloemberg, G. V. (2013). Consequences of revised CLSI and EUCAST guidelines for antibiotic susceptibility patterns of ESBL- and AmpC β-lactamase-producing clinical *Enterobacteriaceae* isolates. *Journal of Antimicrobial Chemotherapy, 68*, 2092-2098. doi:10.1093/jac/dkt136

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics, 28*(4), 593-594. doi:10.1093/bioinformatics/btr708

Hughes, D., & Andersson, D. I. (2017). Environmental and genetic modulation of the phenotypic expression of antibiotic resistance. *FEMS Microbiol Rev, 41*(3), 374-391. doi:10.1093/femsre/fux004

Huseby, D. L., Pietsch, F., Brandis, G., Garoff, L., Tegehall, A., & Hughes, D. (2017). Mutation Supply and Relative Fitness Shape the Genotypes of Ciprofloxacin-Resistant *Escherichia coli*. *Mol Biol Evol, 34*(5), 1029-1039. doi:10.1093/molbev/msx052

Antibiotics: past, present and future, 51, Elsevier Ltd 72-80 (2019).

Hyun, J. C., Kavvas, E. S., Monk, J. M., & Palsson, B. O. (2020). Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS Comput Biol, 16*(3), e1007608. doi:10.1371/journal.pcbi.1007608

Jacoby, G. A. (2005). Mechanisms of resistance to quinolones. *Clin Infect Dis, 41 Suppl 2*, S120-126. doi:10.1086/428052

Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., . . . Tenaillon, O. (2013). Capturing the mutational landscape of the beta-lactamase TEM-1. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.1215206110

Johnson, J. R., Johnston, B., Clabots, C., Kuskowski, M. A., & Castanheira, M. (2010). *Escherichia coli* sequence type ST131 as the major cause of serious multidrug-resistant *E. coli* infections in the United States. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 51*, 286-294. doi:10.1086/653932

Jolley, K. A., Bray, J. E., & Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome open research, 3*, 124. doi:10.12688/wellcomeopenres.14826.1

Jorgensen, J. H., & Ferraro, M. J. (2000). Antimicrobial susceptibility testing: special needs for fastidious organisms and difficult-to-detect resistance mechanisms. *Clin Infect Dis, 30*(5), 799-808. doi:10.1086/313788

Kashanian, J., Hakimian, P., Blute, M., Wong, J., Khanna, H., Wise, G., & Shabsigh, R. (2008). Nitrofurantoin: The return of an old friend in the wake of growing resistance. *BJU International, 102*(11), 1634-1637. doi:10.1111/j.1464-410X.2008.07809.x

Kassim, A., Omuse, G., Premji, Z., & Revathi, G. (2016). Comparison of Clinical Laboratory Standards Institute and European Committee on Antimicrobial

Susceptibility Testing guidelines for the interpretation of antibiotic susceptibility at a University teaching hospital in Nairobi, Kenya: a cross-sectional study. *Annals of clinical microbiology and antimicrobials, 15*, 21. doi:10.1186/s12941-016-0135-3

Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., . . . Palsson, B. O. (2018). Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat Commun, 9*(1), 4306. doi:10.1038/s41467-018-06634-y

Khaledi, A., Weimann, A., Schniederjans, M., Asgari, E., Kuo, T. H., Oliver, A., . . . Haussler, S. (2020). Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. *EMBO Mol Med, 12*(3), e10264. doi:10.15252/emmm.201910264

Khan, S., Sallum, U. W., Zheng, X., Nau, G. J., & Hasan, T. (2014). Rapid optical determination of β-lactamase and antibiotic activity. *BMC Microbiology*. doi:10.1186/1471-2180-14-84

Kim, J., Greenberg, D. E., Pfifer, R., Jiang, S., Xiao, G., Xie, Y., . . . Zhan, X. (2019). VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. *bioRxiv*. doi:10.1101/537381

Kos, V. N., Deraspe, M., McLaughlin, R. E., Whiteaker, J. D., Roy, P. H., Alm, R. A., . . . Gardner, H. (2015). The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrob Agents Chemother, 59*(1), 427-436. doi:10.1128/AAC.03954-14

Koutsogiannou, M., Drougka, E., Liakopoulos, A., Jelastopulu, E., Petinaki, E., Anastassiou, E. D., . . . Christofidou, M. (2013). Spread of multidrug-resistant *Pseudomonas aeruginosa* clones in a university hospital. *Journal of clinical microbiology, 51*, 665-668. doi:10.1128/JCM.03071-12

Kröse, B., Krose, B., van der Smagt, P., & Smagt, P. (1993). An introduction to neural networks.

Lai, T. L., Robbins, H., & Wei, C. Z. (1978). Strong consistency of least squares estimates in multiple regression. *Proc Natl Acad Sci U S A, 75*(7), 3034-3036. doi:10.1073/pnas.75.7.3034

Land, M., Hauser, L., Jun, S. R., Nookaew, I., Leuze, M. R., Ahn, T. H., . . . Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics, 15*(2), 141-161. doi:10.1007/s10142-015-0433-4

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., . . . Robles, V. (2006). Machine learning in bioinformatics. *Brief Bioinform, 7*(1), 86-112. doi:10.1093/bib/bbk007

Lee, R. S., Seemann, T., Heffernan, H., Kwong, J. C., Gonçalves da Silva, A., Carter, G. P., . . . Williamson, D. A. (2018). Genomic epidemiology and antimicrobial resistance of *Neisseria gonorrhoeae* in New Zealand. *The Journal of antimicrobial chemotherapy, 73*, 353-364. doi:10.1093/jac/dkx405

Leekha, S., Terrell, C. L., & Edson, R. S. (2011). General principles of antimicrobial therapy. *Mayo Clin Proc, 86*(2), 156-167. doi:10.4065/mcp.2010.0639

Lemaître, G., Nogueira, F., Learning, C. A.-T. J. o. M., & 2017, U. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research, 18*, 559-563.

Lewis, D. D. (1998). Naive(Bayes)at forty: The independence assumption in information retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Springer, Berlin, Heidelberg, 4-15.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, Y., Metcalf, B. J., Chochua, S., Li, Z., Gertz, R. E., Jr., Walker, H., . . . Active Bacterial Core surveillance, t. (2017). Validation of beta-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences. *BMC Genomics, 18*(1), 621. doi:10.1186/s12864-017-4017-7

Li, Y., Metcalf, B. J., Chochua, S., Li, Z., Gertz, R. E., Jr., Walker, H., . . . Beall, B. W. (2016). Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting beta-Lactam Resistance Levels in Streptococcus pneumoniae. *mBio, 7*(3). doi:10.1128/mBio.00756-16

Improving the estimation of the global burden of antimicrobial resistant infections, 19, Lancet Publishing Group e392-e398 (2019).

Lin, J., Nishino, K., Roberts, M. C., Tolmasky, M., Aminov, R. I., & Zhang, L. (2015). Mechanisms of antibiotic resistance. *Front Microbiol, 6*, 34. doi:10.3389/fmicb.2015.00034

Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., . . . Lewis, K. (2015). A new antibiotic kills pathogens without detectable resistance. *Nature, 517*(7535), 455-459. doi:10.1038/nature14098

Liu, Y. Y., Wang, Y., Walsh, T. R., Yi, L. X., Zhang, R., Spencer, J., . . . Shen, J. (2016). Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect Dis, 16*(2), 161-168. doi:10.1016/S1473-3099(15)00424-7

Liu, Z., Deng, D., Lu, H., Sun, J., Lv, L., Li, S., . . . Wang, G. (2020). Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of *Actinobacillus pleuropneumoniae* From Whole Genome Sequences. *Frontiers in Microbiology, 11*, 48. doi:10.3389/fmicb.2020.00048

Livermore, D. M., Sefton, A. M., & Scott, G. M. (2003). Properties and potential of ertapenem. *Journal of Antimicrobial Chemotherapy, 52*, 331-344. doi:10.1093/jac/dkg375

Lobanovska, M., & Pilla, G. (2017). Penicillin's Discovery and Antibiotic Resistance: Lessons for the Future? *Yale J Biol Med, 90*(1), 135-145.

Loman, N. J., & Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nat Rev Microbiol, 13*(12), 787-794. doi:10.1038/nrmicro3565

Long, S. W., Olsen, R. J., Eagar, T. N., Beres, S. B., Zhao, P., Davis, J. J., . . . Musser, J. M. (2017). Population Genomic Analysis of 1,777 Extended-Spectrum Beta-Lactamase-Producing *Klebsiella pneumoniae* Isolates, Houston, Texas: Unexpected Abundance of Clonal Group 307. *mBio, 8*(3). doi:10.1128/mBio.00489-17

Lv, J., Deng, S., Zhang, L. J. B., & Health. (2020). A review of artificial intelligence applications for antimicrobial resistance.

Macesic, N., Polubriaginof, F., & Tatonetti, N. P. (2017). Machine learning: novel bioinformatics approaches for combating antimicrobial resistance. *Curr Opin Infect Dis, 30*(6), 511-517. doi:10.1097/QCO.0000000000000406

MacFadden, D. R., Melano, R. G., Coburn, B., Tijet, N., Hanage, W. P., & Daneman, N. (2019). Comparing Patient Risk Factor-, Sequence Type-, and Resistance Locus Identification-Based Approaches for Predicting Antibiotic Resistance in *Escherichia coli* Bloodstream Infections. *Journal of Clinical Microbiology, 57*, 1780-1798. doi:10.1128/JCM.01780-18

Madden, T. (2013). The BLAST Sequence Analysis Tool. *NCBI Handbook* [Internet]. 2nd edition. National Center for Biotechnology Information (US).

Maguire, F., Rehman, M. A., Carrillo, C., Diarra, M. S., & Beiko, R. G. (2019). Identification of Primary Antimicrobial Resistance Drivers in Agricultural Nontyphoidal *Salmonella enterica* Serovars by Using Machine Learning. *mSystems, 4*(4). doi:10.1128/mSystems.00211-19

Mahfouz, N., Ferreira, I., Beisken, S., von Haeseler, A., & Posch, A. E. (2020). Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. *J Antimicrob Chemother, 75*(11), 3099-3108. doi:10.1093/jac/dkaa257

Majiduddin, F. K., & Palzkill, T. (2005). Amino acid residues that contribute to substrate specificity of class a β-lactamase SME-1. *Antimicrobial Agents and Chemotherapy*. doi:10.1128/AAC.49.8.3421-3427.2005

Marchand, M., Shawe-Taylor, J., Brodley, C. E., & Danyluk, A. (2002). *The Set Covering Machine. Journal of Machine Learning Research*, 3(4-5), 723-746.

Marcusson, L. L., Frimodt-Moller, N., & Hughes, D. (2009). Interplay in the selection of fluoroquinolone resistance and bacterial fitness. *PLoS Pathog, 5*(8), e1000541. doi:10.1371/journal.ppat.1000541

Mason, A., Foster, D., Bradley, P., Golubchik, T., Doumith, M., Gordon, N. C., . . . Peto, T. (2018). Accuracy of Different Bioinformatics Methods in Detecting Antibiotic Resistance and Virulence Factors from *Staphylococcus aureus* Whole-Genome Sequences. *J Clin Microbiol, 56*(9). doi:10.1128/JCM.01815-17

Maugeri, G., Lychko, I., Sobral, R., & Roque, A. C. A. (2019). Identification and Antibiotic-Susceptibility Profiling of Infectious Bacterial Agents: A Review of Current and Future Trends. *Biotechnology journal, 14*, e1700750. doi:10.1002/biot.201700750

Maurer, F. P., Christner, M., Hentschke, M., & Rohde, H. (2017). Advances in Rapid Identification and Susceptibility Testing of Bacteria in the Clinical Microbiology

Laboratory: Implications for Patient Care and Antimicrobial Stewardship Programs. *Infectious disease reports, 9*, 6839. doi:10.4081/idr.2017.6839

McArthur, A. G., & Tsang, K. K. (2017). Antimicrobial resistance surveillance in the genomic age. *Annals of the New York Academy of Sciences, 1388*, 78-91. doi:10.1111/nyas.13289

McDermott, P. F., & Davis, J. J. (2021). Predicting antimicrobial susceptibility from the bacterial genome: A new paradigm for one health resistance monitoring. *J Vet Pharmacol Ther, 44*(2), 223-237. doi:10.1111/jvp.12913

McDermott, P. F., Tyson, G. H., Kabera, C., Chen, Y., Li, C., Folster, J. P., . . . Zhao, S. (2016). Whole-Genome Sequencing for Detecting Antimicrobial Resistance in Nontyphoidal *Salmonella*. *Antimicrobial agents and chemotherapy, 60*, 5515-5520. doi:10.1128/AAC.01030-16

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference, 455*, 51-56.

Melendez, J. H., Hardick, J., Barnes, M., Page, K. R., & Gaydos, C. A. (2018). Antimicrobial Susceptibility of *Neisseria gonorrhoeae* Isolates in Baltimore, Maryland, 2016: The Importance of Sentinel Surveillance in the Era of Multi-Drug-Resistant Gonorrhea. *Antibiotics (Basel), 7*(3). doi:10.3390/antibiotics7030077

Mensah, N., Tang, Y., Cawthraw, S., AbuOun, M., Fenner, J., Thomson, N. R., . . . Petrovska-Holmes, L. (2019). Determining antimicrobial susceptibility in *Salmonella enterica* serovar Typhimurium through whole genome sequencing: a comparison against multiple phenotypic susceptibility testing methods. *BMC Microbiol, 19*(1), 148. doi:10.1186/s12866-019-1520-9

Metcalf, B. J., Chochua, S., Gertz, R. E., Jr., Li, Z., Walker, H., Tran, T., . . McGee, L. (2016). Using whole genome sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes for year 2015 invasive pneumococcal disease isolates recovered in the United States. *Clin Microbiol Infect, 22*(12), 1002 e1001-1002 e1008. doi:10.1016/j.cmi.2016.08.001

Metcalf, B. J., Chochua, S., Gertz, R. E., Li, Z., Walker, H., Tran, T., . . . Langley, G. (2016). Using whole genome sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes for year 2015 invasive pneumococcal disease isolates recovered in the United States. *Clinical Microbiology and Infection, 22*, 1002.e1001-1002.e1008. doi:10.1016/j.cmi.2016.08.001

Minh, N. N., Thuong, T. C., Khuong, H. D., Nga, T. V., Thompson, C., Campbell, J. I., . . . Baker, S. (2012). The co-selection of fluoroquinolone resistance genes in the gut flora of Vietnamese children. *PLoS One, 7*(8), e42919. doi:10.1371/journal.pone.0042919

Miotto, P., Tessema, B., Tagliani, E., Chindelevitch, L., Starks, A. M., Emerson, C., . . . Rodwell, T. C. (2017). A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*. *Eur Respir J, 50*(6). doi:10.1183/13993003.01354-2017

Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., & Parts, L. (2018). Prediction of antibiotic resistance in *Escherichia coli* from large-scale

pan-genome data. *PLoS Comput Biol, 14*(12), e1006258. doi:10.1371/journal.pcbi.1006258

Moran, R. A., Anantham, S., Holt, K. E., & Hall, R. M. (2016). Prediction of antibiotic resistance from antibiotic resistance genes detected in antibiotic-resistant commensal *Escherichia coli* using PCR or WGS. *Journal of Antimicrobial Chemotherapy*, dkw511. doi:10.1093/jac/dkw511

Müller, R., & Chauve, C. (2019). HyAsP, a greedy tool for plasmids identification. *Bioinformatics, 35*(21), 4436-4439. doi:10.1093/bioinformatics/btz413

Naghavi, M., Abajobir, A. A., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abera, S. F., . . . Murray, C. J. L. (2017). Global, regional, and national age-sex specifc mortality for 264 causes of death, 1980-2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet, 390*, 1151-1210. doi:10.1016/S0140-6736(17)32152-9

NCBI Resource Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res, 46*(D1), D8-D13. doi:10.1093/nar/gkx1095

Neuert, S., Nair, S., Day, M. R., Doumith, M., Ashton, P. M., Mellor, K. C., . . . Dallman, T. J. (2018). Prediction of Phenotypic Antimicrobial Resistance Profiles From Whole Genome Sequences of Non-typhoidal *Salmonella enterica*. *Front Microbiol, 9*, 592. doi:10.3389/fmicb.2018.00592

Ng, L. K., Martin, I., Liu, G., & Bryden, L. (2002). Mutation in 23S rRNA associated with macrolide resistance in *Neisseria gonorrhoeae*. *Antimicrob Agents Chemother, 46*(9), 3020-3025. doi:10.1128/AAC.46.9.3020-3025.2002

Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., Olson, R., . . . Davis, J. J. (2018). Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae. Scientific Reports, 8*, 421. doi:10.1038/s41598-017-18972-w

Nguyen, M., Olson, R., Shukla, M., VanOeffelen, M., & Davis, J. J. (2020). Predicting antimicrobial resistance using conserved genes. *PLoS Computational Biology, 16*. doi:10.1371/journal.pcbi.1008319

Nguyen, M., Wesley Long, S., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., . . . Davisa, J. J. (2019). Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *Journal of Clinical Microbiology, 57*. doi:10.1128/JCM.01260-18

Niehaus, K. E., Walker, T. M., Crook, D. W., Peto, T. E. A., & Clifton, D. A. (2014, 1-4 June 2014). *Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis.* Paper presented at the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 618-621.

O'Neill, J. (2016). Tackling drug-resistant infections globally: final report and recommendations. *the Review on Antimicrobial Resistance*, 84. doi:10.1016/j.jpha.2015.11.005

Ocampo-Sosa, A. A., Cabot, G., Rodríguez, C., Roman, E., Tubau, F., Macia, M. D., . . . Martínez-Martínez, L. (2012). Alterations of OprD in carbapenem-intermediate and -susceptible strains of *Pseudomonas aeruginosa* isolated from patients with

bacteremia in a spanish multicenter study. *Antimicrobial Agents and Chemotherapy, 56*, 1703-1713. doi:10.1128/AAC.05451-11

Oliphant, T. E. (2006). A guide to NumPy. 85.

Olsen, I. (2015). Biofilm-specific antibiotic tolerance and resistance. *Eur J Clin Microbiol Infect Dis, 34*(5), 877-886. doi:10.1007/s10096-015-2323-z

Pankhurst, L. J., del Ojo Elias, C., Votintseva, A. A., Walker, T. M., Cole, K., Davies, J., . . . Kong, C. J. T. L. R. M. (2016). Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *4*(1), 49-58.

Pataki, B. Á., Matamoros, S., van der Putten, B. C. L., Remondini, D., Giampieri, E., Aytan-Aktug, D., . . . McDermott, P. (2020). Understanding and predicting ciprofloxacin minimum inhibitory concentration in *Escherichia coli* with machine learning. *Scientific Reports, 10*, 1-9. doi:10.1038/s41598-020-71693-5

Patel, R. (2005). Biofilms and antimicrobial resistance. *Clin Orthop Relat Res*(437), 41-47. doi:10.1097/01.blo.0000175714.68624.74

Pedregosa, F., Varoquaux, G., Gramfort, A., & Michel, V. (2011). Scikit-learn: Machine Learning in Python. *Journal of machine learning research, 12*, 2825-2830.

Pesesky, M. W., Hussain, T., Wallace, M., Patel, S., Andleeb, S., Burnham, C.-A. D., & Dantas, G. (2016). Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome Sequence Data. *Frontiers in microbiology, 7*, 1887. doi:10.3389/fmicb.2016.01887

Piddock, L. J. V. (2014). Understanding the basis of antibiotic resistance: A platform for drug discovery. *Microbiology (United Kingdom), 160*, 2366-2373. doi:10.1099/mic.0.082412-0

Pitout, J. D. D., & DeVinney, R. (2017). *Escherichia coli* ST131: a multidrug-resistant clone primed for global domination. *F1000Research, 6*, 195. doi:10.12688/f1000research.10609.1

Plesiat, P., Alkhalaf, B., & Michel-Briand, Y. (1988). Prevalence and profiles of plasmids in *Pseudomonas aeruginosa*. *Eur J Clin Microbiol Infect Dis, 7*(2), 261-264. doi:10.1007/BF01963098

Poirel, L., Gniadkowski, M., & Nordmann, P. (2002). Biochemical analysis of the ceftazidime-hydrolysing extended-spectrum β-lactamse CTX-M-15 and of its structurally related β-lactamase CTX-M-3. *Journal of Antimicrobial Chemotherapy*. doi:10.1093/jac/dkf240

Prestinaci, F., Pezzotti, P., & Pantosti, A. (2015). Antimicrobial resistance: a global multifaceted phenomenon. *Pathog Glob Health, 109*(7), 309-318. doi:10.1179/2047773215Y.0000000030

Public Health Agency of Canada. (2020). *Canadian Antimicrobial System Report Resistance Surveillance System Report* (6139572991). Retrieved from Ottawa, ON: https://www.canada.ca/en/public-health/services/publications/drugs-health-products/canadian-antimicrobial-resistance-surveillance-system-2020-report.html

Quan, T. P., Bawa, Z., Foster, D., Walker, T., Del Ojo Elias, C., Rathod, P., . . . Smith, E. G. (2018). Evaluation of Whole-Genome Sequencing for Mycobacterial Species

Identification and Drug Susceptibility Testing in a Clinical Setting: a Large-Scale Prospective Assessment of Performance against Line Probe Assays and Phenotyping. *J Clin Microbiol, 56*(2). doi:10.1128/JCM.01480-17

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81-106. doi:10.1007/bf00116251

Rammelkamp, C. H. & Maxon, T. (1942). Resistance of *Staphylococcus aureus* to the Action of Penicillin. *51*(3), 386-389.

Ransom, E. M., Potter, R. F., Dantas, G., & Burnham, C. D. (2020). Genomic Prediction of Antimicrobial Resistance: Ready or Not, Here It Comes! *Clin Chem, 66*(10), 1278-1289. doi:10.1093/clinchem/hvaa172

Reller, L. B., Weinstein, M., Jorgensen, J. H., & Ferraro, M. J. (2009). Antimicrobial susceptibility testing: a review of general principles and contemporary practices. *Clin Infect Dis, 49*(11), 1749-1755. doi:10.1086/647952

Robertson, J., & Nash, J. H. E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom, 4*(8). doi:10.1099/mgen.0.000206

Robicsek, A., Strahilevitz, J., Jacoby, G. A., Macielag, M., Abbanat, D., Hye Park, C., . . . Hooper, D. C. (2006). Fluoroquinolone-modifying enzyme: a new adaptation of a common aminoglycoside acetyltransferase. *Nature Medicine, 12*, 83-88. doi:10.1038/nm1347

Rodríguez-Baño, J., Picón, E., Navarro, M. D., López-Cerero, L., & Pascual, Á. (2012). Impact of changes in CLSI and EUCAST breakpoints for susceptibility in bloodstream infections due to extended-spectrum β-lactamase-producing *Escherichia coli*. *Clinical Microbiology and Infection, 18*, 894-900. doi:10.1111/j.1469-0691.2011.03673.x

Rossen, J. W. A., Friedrich, A. W., & Moran-Gilad, J. (2018). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin Microbiol Infect, 24*(4), 355-360. doi:10.1016/j.cmi.2017.11.001

Sadouki, Z., Day, M. R., Doumith, M., Chattaway, M. A., Dallman, T. J., Hopkins, K. L., . . . Jenkins, C. (2017). Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of *Shigella sonnei* isolated from cases of diarrhoeal disease in England and Wales, 2015. *J Antimicrob Chemother, 72*(9), 2496-2502. doi:10.1093/jac/dkx170

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One, 10*, e0118432. doi:10.1371/journal.pone.0118432

Santerre, J. W., Davis, J. J., Gov, J. a., Xia, F., Gov, F. a., & Stevens, R. (2016). Machine Learning for Antimicrobial Resistance. *arXiv, 1607.01224*.

Sawatzky, P., Liu, G., Dillon, J. A., Allen, V., Lefebvre, B., Hoang, L., . . . Martin, I. (2015). Quality Assurance for Antimicrobial Susceptibility Testing of *Neisseria gonorrhoeae* in Canada, 2003 to 2012. *J Clin Microbiol, 53*(11), 3646-3649. doi:10.1128/JCM.02303-15

Scaria, J., Chandramouli, U., & Verma, S. K. (2005). Antibiotic Resistance Genes Online (ARGO): a Database on vancomycin and beta-lactam resistance genes. *Bioinformation, 1*(1), 5-7. doi:10.6026/97320630001005

Schmutz, E., Mühlenweg, A., Li, S. M., & Heide, L. (2003). Resistance genes of aminocoumarin producers: Two type II topoisomerase genes confer resistance against coumermycin A1 and clorobiocin. *Antimicrobial Agents and Chemotherapy*. doi:10.1128/AAC.47.3.869-877.2003

Shafer, W. M., & Folster, J. P. (2006). Towards an understanding of chromosomally mediated penicillin resistance in *Neisseria gonorrhoeae*: evidence for a porin-efflux pump collaboration. *J Bacteriol, 188*(7), 2297-2299. doi:10.1128/JB.188.7.2297-2299.2006

Shelburne, S. A., Kim, J., Munita, J. M., Sahasrabhojane, P., Shields, R. K., Press, E. G., . . . Greenberg, D. E. (2017). Whole-Genome Sequencing Accurately Identifies Resistance to Extended-Spectrum beta-Lactams for Major Gram-Negative Bacterial Pathogens. *Clin Infect Dis, 65*(5), 738-745. doi:10.1093/cid/cix417

Shi, J., Yan, Y., Links, M. G., Li, L., Dillon, J.-A. R., Horsch, M., & Kusalik, A. (2019). Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection. *BMC Bioinformatics, 20*, 535. doi:10.1186/s12859-019-3054-4

Silver, L. L. (2011). Challenges of antibacterial discovery. *Clin Microbiol Rev, 24*(1), 71-109. doi:10.1128/CMR.00030-10

Smith, K. P., & Kirby, J. E. (2019). Rapid Susceptibility Testing Methods. *Clin Lab Med, 39*(3), 333-344. doi:10.1016/j.cll.2019.04.001

Stoesser, N., Batty, E. M., Eyre, D. W., Morgan, M., Wyllie, D. H., Del Ojo Elias, C., . . . Crook, D. W. (2013). Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. *The Journal of antimicrobial chemotherapy, 68*, 2234-2244. doi:10.1093/jac/dkt180

Stojanoski, V., Chow, D. C., Hu, L., Sankaran, B., Gilbert, H. F., Prasad, B. V., & Palzkill, T. (2015). A triple mutant in the Omega-loop of TEM-1 beta-lactamase changes the substrate profile via a large conformational change and an altered general base for catalysis. *J Biol Chem, 290*(16), 10382-10394. doi:10.1074/jbc.M114.633438

Su, M., Satola, S. W., & Read, T. D. (2018). Genome-based prediction of bacterial antibiotic resistance. *Journal of Clinical Microbiology, 57*. doi:10.1128/jcm.01405-18

Sutcliffe, J. G. (1978). Nucleotide sequence of the ampicillin resistance gene of *Escherichia coli* plasmid pBR322. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.75.8.3737

Suthaharan, S. (2016). Support vector machine. In *Machine learning models and algorithms for big data classification* (pp. 207-235): Springer.

Tapsall, J. W. (2009). *Neisseria gonorrhoeae* and emerging resistance to extended spectrum cephalosporins. *Curr Opin Infect Dis, 22*(1), 87-91. doi:10.1097/QCO.0b013e328320a836

Tchesnokova, V. L., Rechkina, E., Larson, L., Ferrier, K., Weaver, J. L., Schroeder, D. W., . . . Sokurenko, E. V. (2019). Rapid and Extensive Expansion in the United States of a New Multidrug-resistant *Escherichia coli* Clonal Group, Sequence Type 1193. *Clinical Infectious Diseases, 68*, 334-337. doi:10.1093/cid/ciy525

Thulin, E., Sundqvist, M., & Andersson, D. I. (2015). Amdinocillin (Mecillinam) resistance mutations in clinical isolates and laboratory-selected mutants of *Escherichia coli*. *Antimicrob Agents Chemother, 59*(3), 1718-1727. doi:10.1128/AAC.04819-14

Ting, K. M. (2017). Confusion Matrix. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 260-260). Boston, MA: Springer US.

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet, 13*(1), 36-46. doi:10.1038/nrg3117

Treepong, P., Kos, V. N., Guyeux, C., Blanc, D. S., Bertrand, X., Valot, B., & Hocquet, D. (2018). Global emergence of the widespread *Pseudomonas aeruginosa* ST235 clone. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases, 24*, 258-266. doi:10.1016/j.cmi.2017.06.018

Tsang, K. K., Maguire, F., Zubyk, H. L., Chou, S., Edalatmand, A., Wright, G. D., . . . McArthur, A. G. (2021). Identifying novel β-lactamase substrate activity through *in silico* prediction of antimicrobial resistance. *Microbial Genomics, 7*, 1-13. doi:10.1099/mgen.0.000500

Tuan, V. P., Narith, D., Tshibangu-Kabamba, E., Dung, H. D. Q., Viet, P. T., Sokomoth, S., . . . Yamaoka, Y. (2019). A Next-Generation Sequencing-Based Approach to Identify Genetic Determinants of Antibiotic Resistance in Cambodian *Helicobacter pylori* Clinical Isolates. *J Clin Med, 8*(6). doi:10.3390/jcm8060858

Tyson, G. H., McDermott, P. F., Li, C., Chen, Y., Tadesse, D. A., Mukherjee, S., . . . Zhao, S. (2015). WGS accurately predicts antimicrobial resistance in *Escherichia coli*. *J Antimicrob Chemother, 70*(10), 2763-2769. doi:10.1093/jac/dkv186

Tyson, G. H., Sabo, J. L., Rice-Trujillo, C., Hernandez, J., & McDermott, P. F. (2018). Whole-genome sequencing based characterization of antimicrobial resistance in *Enterococcus*. *Pathog Dis, 76*(2). doi:10.1093/femspd/fty018

U.S. Department of Health and Human Services. (2009). *Antimicrobial Susceptibility Test (AST) Systems - Class II Special Controls Guidance for Industry and FDA*. Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm080564.htm

Unemo, M., Golparian, D., Sanchez-Buso, L., Grad, Y., Jacobsson, S., Ohnishi, M., . . . Harris, S. R. (2016). The novel 2016 WHO *Neisseria gonorrhoeae* reference strains for global quality assurance of laboratory investigations: phenotypic, genetic and reference genome characterization. *J Antimicrob Chemother, 71*(11), 3096-3108. doi:10.1093/jac/dkw288

Unemo, M., & Shafer, W. M. (2011). Antibiotic resistance in *Neisseria gonorrhoeae*: origin, evolution, and lessons learned for the future. *Ann N Y Acad Sci, 1230*, E19-28. doi:10.1111/j.1749-6632.2011.06215.x

Unemo, M., & Shafer, W. M. (2014). Antimicrobial resistance in *Neisseria gonorrhoeae* in the 21st Century: Past, evolution, and future. *Clinical Microbiology Reviews, 27*, 587-613. doi:10.1128/CMR.00010-14

Uppala, A., King, E. A., & Patel, D. (2019). Cefazolin versus fluoroquinolones for the treatment of community-acquired urinary tract infections in hospitalized patients. *European Journal of Clinical Microbiology and Infectious Diseases, 38*, 1533-1538. doi:10.1007/s10096-019-03582-3

Utturkar, S. M., Klingeman, D. M., Hurt, R. A., Jr., & Brown, S. D. (2017). A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies. *Front Microbiol, 8*, 1272. doi:10.3389/fmicb.2017.01272

van Belkum, A., Bachmann, T. T., Lüdke, G., Lisby, J. G., Kahlmeter, G., Mohess, A., . . . Testing, J. A.-R. W. G. o. A. R. a. R. D. (2019). Developmental roadmap for antimicrobial susceptibility testing systems. *Nature Reviews Microbiology, 17*, 51-62. doi:10.1038/s41579-018-0098-9

van Rossum, G., & Drake, F. L. (2003). Python Language Reference Manual. *Python Language Reference Manual*.

Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and therapeutics, 40*(4), 277-283.

Walker, T. M., Kohl, T. A., Omar, S. V., Hedge, J., Del Ojo Elias, C., Bradley, P., . . . Modernizing Medical Microbiology Informatics, G. (2015). Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis, 15*(10), 1193-1202. doi:10.1016/S1473-3099(15)00062-6

Wang, S.-C. (2003). Artificial neural network. In *Interdisciplinary computing in java programming* (pp. 81-100): Springer.

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., . . . Stevens, R. L. (2017). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic acids research, 45*, D535-D542. doi:10.1093/nar/gkw1017

Webber, M., & Piddock, L. J. (2001). Quinolone resistance in *Escherichia coli*. *Vet Res, 32*(3-4), 275-284. doi:10.1051/vetres:2001124

Weigel, L. M., Steward, C. D., & Tenover, F. C. (1998). gyrA mutations associated with fluoroquinolone resistance in eight species of *Enterobacteriaceae*. *Antimicrob Agents Chemother, 42*(10), 2661-2667. doi:10.1128/AAC.42.10.2661

Wolfensberger, A., Sax, H., Weber, R., Zbinden, R., Kuster, S. P., & Hombach, M. (2013). Change of Antibiotic Susceptibility Testing Guidelines from CLSI to EUCAST: Influence on Cumulative Hospital Antibiograms. *PLoS One, 8*, e79130. doi:10.1371/journal.pone.0079130

World Health Organization. (2014). *Antimicrobial resistance: global report on surveillance.* Paper presented at the World Health Organization Report, Geneva.

World Health Organization. (2015). *Global action plan on antimicrobial resistance.* Paper presented at the WHO, Geneva.

World Health Organization. (2017a). *Global antimicrobial resistance surveillance system (GLASS) report: early implementation 2016-2017.*, Geneva.

World Health Organization. (2017b). *Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics*, Geneva.

World Health Organization. (2019). *Molecular methods for antimicrobial resistance (AMR) diagnostics to enhance the Global Antimicrobial Resistance Surveillance System*, Geneva.

World Health Organization. (2021). 2020 antibacterial agents in clinical and preclinical development: an overview and analysis.

Wright, G. D. (2007). The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol, 5*(3), 175-186. doi:10.1038/nrmicro1614

Wright, G. D. (2011). Molecular mechanisms of antibiotic resistance. *Chem Commun (Camb), 47*(14), 4055-4061. doi:10.1039/c0cc05111j

Wright, R. E. (1995). Logistic regression. In *Reading and understanding multivariate statistics.* (pp. 217-244). Washington, DC, US: American Psychological Association.

Wu, J., Lan, F., Lu, Y., He, Q., & Li, B. (2017). Molecular Characteristics of ST1193 Clone among Phylogenetic Group B2 Non-ST131 Fluoroquinolone-Resistant *Escherichia coli*. *Frontiers in Microbiology, 8*, 2294. doi:10.3389/fmicb.2017.02294

Xavier, B. B., Das, A. J., Cochrane, G., De Ganck, S., Kumar-Singh, S., Aarestrup, F. M., . . . Malhotra-Kumar, S. (2016). Consolidating and Exploring Antibiotic Resistance Gene Data Resources. *J Clin Microbiol, 54*(4), 851-859. doi:10.1128/JCM.02717-15

Xia, L., Liu, Y., Xia, S., Kudinha, T., Xiao, S.-N., Zhong, N.-S., . . . Zhuo, C. (2017). Prevalence of ST1193 clone and IncI1/ST16 plasmid in *E. coli* isolates carrying blaCTX-M-55 gene from urinary tract infections patients in China. *Scientific reports, 7*, 44866. doi:10.1038/srep44866

Yang, Y., Niehaus, K. E., Walker, T. M., Iqbal, Z., Walker, A. S., Wilson, D. J., . . . Clifton, D. A. (2018). Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics, 34*, 1666-1671. doi:10.1093/bioinformatics/btx801

Yong, D., Toleman, M. A., Giske, C. G., Cho, H. S., Sundman, K., Lee, K., & Walsh, T. R. (2009). Characterization of a new metallo-beta-lactamase gene, bla(NDM-1), and a novel erythromycin esterase gene carried on a unique genetic structure in *Klebsiella pneumoniae* sequence type 14 from India. *Antimicrob Agents Chemother, 53*(12), 5046-5054. doi:10.1128/AAC.00774-09

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., . . . Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *The Journal of antimicrobial chemotherapy, 67*, 2640-2644. doi:10.1093/jac/dks261

Zankari, E., Hasman, H., Kaas, R. S., Seyfarth, A. M., Agerso, Y., Lund, O., . . . Aarestrup, F. M. (2013). Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J Antimicrob Chemother, 68*(4), 771-777. doi:10.1093/jac/dks496

Zhao, L., Wang, S., Li, X., He, X., & Jian, L. (2020). Development of in vitro resistance to fluoroquinolones in *Pseudomonas aeruginosa*. *Antimicrob Resist Infect Control, 9*(1), 124. doi:10.1186/s13756-020-00793-8

Zhao, S., Tyson, G. H., Chen, Y., Li, C., Mukherjee, S., Young, S., . . . McDermott, P. F. (2015). Whole-Genome Sequencing Analysis Accurately Predicts Antimicrobial Resistance Phenotypes in *Campylobacter* spp. *Applied and environmental microbiology, 82*, 459-466. doi:10.1128/AEM.02873-15

Zorzet, A. (2014). Overcoming scientific and structural bottlenecks in antibacterial discovery and development. *Ups J Med Sci, 119*(2), 170-175. doi:10.3109/03009734.2014.897277

Zubyk, H. L., & Wright, G. D. (2021). CrpP is not a Fluoroquinolone Inactivating Enzyme. *Antimicrob Agents Chemother*. doi:10.1128/AAC.00773-21

APPENDICES

APPENDIX 1: Tsang, K. K., Maguire, F., Zubyk, H. L., Chou, S., Edalatmand, A., Wright, G. D., . . . McArthur, A. G. (2021). Identifying novel β-lactamase substrate activity through in silico prediction of antimicrobial resistance. ***Microbial Genomics*****, 7, 1-13. doi:10.1099/mgen.0.000500**

MICROBIAL GENOMICS

RESEARCH ARTICLE

Tsang *et al.*, *Microbial Genomics* 2021;7:000500
DOI 10.1099/mgen.0.000500

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Identifying novel β-lactamase substrate activity through *in silico* prediction of antimicrobial resistance

Kara K. Tsang[1,2,3], Finlay Maguire[4], Haley L. Zubyk[1,2,3], Sommer Chou[1,2,3], Arman Edalatmand[1,2,3], Gerard D. Wright[1,2,3], Robert G. Beiko[4] and Andrew G. McArthur[1,2,3,*]

Abstract

Diagnosing antimicrobial resistance (AMR) in the clinic is based on empirical evidence and current gold standard laboratory phenotypic methods. Genotypic methods have the potential advantages of being faster and cheaper, and having improved mechanistic resolution over phenotypic methods. We generated and applied rule-based and logistic regression models to predict the AMR phenotype from *Escherichia coli* and *Pseudomonas aeruginosa* multidrug-resistant clinical isolate genomes. By inspecting and evaluating these models, we identified previously unknown β-lactamase substrate activities. In total, 22 unknown β-lactamase substrate activities were experimentally validated using targeted gene expression studies. Our results demonstrate that generating and analysing predictive models can help guide researchers to the mechanisms driving resistance and improve annotation of AMR genes and phenotypic prediction, and suggest that we cannot solely rely on curated knowledge to predict resistance phenotypes.

DATA SUMMARY

All genomic data analysed in this work are available through National Center for Biotechnology Information (NCBI) BioProject PRJNA532924. All conda environments, code and intermediate data files required to generate this analysis are available at: https://github.com/karatsang/rulesbased_logisticregression, https://doi.org/10.5281/zenodo.3988480.

INTRODUCTION

Antimicrobial resistance (AMR) is a global health crisis accelerated by overuse and misuse of antimicrobials. Amongst Gram-negative pathogens, AMR *Escherichia coli* and *Pseudomonas aeruginosa* are of urgent and critical concern. The World Health Organization has reported high resistance to fluoroquinolones and third-generation cephalosporins when treating urinary tract *E. coli* infections, leading to reliance on carbapenems as a last-resort treatment option [1], while the US Centers for Disease Control and Prevention estimates nearly 32600 antibiotic-resistant *P. aeruginosa* infection-related hospitalizations in the USA alone in 2017, to which 2700 deaths were attributed [2].

Currently, the gold standards for diagnosing antibiotic resistance are culture-based phenotypic methods. However, the turnaround time for antibiotic susceptibility tests often surpasses the optimal time for life-threatening infection treatment [3, 4]. Furthermore, phenotypic tests do not reveal the genetic underpinnings of resistance. As such, genotypic methods that exploit high-throughput DNA sequencing technology combined with bioinformatics resources have the potential to be faster and more accurate and informative than the current phenotypic paradigm [5]. There is growing momentum toward whole-genome sequencing of clinical infections, but there is a lag in the development of bioinformatic platforms that can accurately predict phenotypes such as virulence and AMR, which is essential for the full application of rapid pathogen sequencing as a robust diagnostic tool. Most sequencing pipelines rely on an AMR sequence database

Received 19 March 2020; Accepted 08 December 2020; Published 08 January 2021
Author affiliations: [1]David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario, Canada; [2]M.G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada; [3]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada; [4]Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada.
***Correspondence:** Andrew G. McArthur, mcarthua@mcmaster.ca
Keywords: antimicrobial resistance; bioinformatics; genotype–phenotype; prediction.
Abbreviations: AMR, antimicrobial resistance; AST, antibiotic susceptibility testing; CARD, Comprehensive Antibiotic Resistance Database; CLSI, Clinical and Laboratory Standards Institute; EPI, efflux pump identifier; LR, logistic regression; MIC, minimum inhibitory concentration; MLST, multilocus sequence type; RGI, resistance gene identifier.
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Five supplementary tables and six supplementary figures are available with the online version of this article.

to predict functional AMR genes from DNA sequences [6], of which there are many. For example, the Comprehensive Antibiotic Resistance Database (CARD) is an ontology-driven genomics database used by the Resistance Gene Identifier (RGI) software to predict intrinsic and acquired resistance determinants in genome sequences [7]. The Antibiotic Resistance Gene-ANNOTation database [8] and Pathosystems Resource Integration Center [9] store a similar breadth of resistance determinants to CARD and also use BLAST-based tools for resistome annotations. Antibiotic Resistance Genes Online [8] only catalogues β-lactam and vancomycin resistance determinants, in comparison to ResFinder [10], which primarily annotates acquired resistance genes using BLASTN, while ResFams [11] is a database of protein domain hidden Markov models associated with AMR function.

Despite our dependence upon curated AMR databases for genotype analysis and prediction of phenotype, maintaining and developing AMR databases and tools are challenging due to the ever evolving AMR genetic landscape, inconsistencies in AMR gene nomenclature, sparsity of phenotypic data and lack of funding for biocuration [12, 13]. Without comprehensiveness in phenotypic testing, such as antibiotic susceptibility testing using a broad panel of antibiotics, all of these databases will inherently be missing the full range of a resistance determinant's substrate specificity. Yet, as β-lactams are the most commonly used antibiotic [14], there is strong motivation in the AMR field to identify the substrate specificity of clinically prevalent β-lactamases [14–20], particularly with regard to β-lactams new to the marketplace. Despite the development of gene-based antibiotic susceptibility testing tools such as the Antibiotic Resistance Platform [21], when novel β-lactamases emerge in clinical settings they are often only characterized using a limited selection of β-lactams, or are assumed to have similar substrate activity to a related β-lactamase. This leads to knowledge gaps in AMR databases for β-lactamase substrate specificity. In the face of missing experimental data, the prediction of novel substrate specificities for known β-lactamases can be performed using statistical modelling and machine learning methods [22–24]. While these statistical models can be used to discover novel genotype–phenotype relationships, they often require large and diverse datasets to be effective. Previous studies have used rule-based and statistical models to predict antibiotic resistance phenotypes from genotypes, but only a few studies provide genotype–phenotype associations [22, 24].

Here we report the *in silico* prediction of genotype–phenotype associations and substrate specificities for AMR determinants from multidrug-resistant *E. coli* and *P. aeruginosa* clinical isolates using two computational approaches (rules-based and logistic regression) based upon CARD's RGI [7]. The rules-based method uses new software (the Efflux Pump Identifier) to account for overexpressed multi-component efflux pumps as well as hand-curated knowledge encoded by CARD's Antibiotic Resistance Ontology (ARO). This method helped identify that gaps in CARD's curated knowledge of β-lactam substrate activity contributed to poor β-lactam resistance phenotype prediction. We then performed logistic regression on the same data, observing higher prediction accuracy across most antibiotic resistance phenotypes. We were then able to experimentally validate the predicted genotype–phenotype relationships (i.e. learned weights) used by logistic regression to identify previously unknown β-lactamase substrate activities.

Impact Statement

Antimicrobial resistance (AMR) is an increasingly global crisis and there is need for technologies that can diagnose and surveil it. We compare statistical modelling and rules-based approaches to predict AMR phenotypes for *Escherichia coli* and *Pseudomonas aeruginosa* clinical isolates based on genome sequence. With an emphasis on the substrate activities of clinically important β-lactamases, our algorithms predict previously unknown β-lactamase substrate activities. We validate these novel substrate activities using a robust experimental target gene expression system. Our work illustrates that known clinical AMR gene threats have a broader range of antibacterial activity than previously thought, with important implications for antibiotic stewardship.

RESULTS

Bacterial isolates, antibiotic susceptibility testing (AST), and whole-genome sequencing

In total, 115 *E. coli* and 102 *P. aeruginosa* putative multidrug-resistant clinical isolates were obtained from Hamilton Health Sciences hospitals (Hamilton, Ontario, Canada) and submitted for both genome sequencing and AST, i.e. categorized as 'resistant' or 'susceptible' for 18 antibiotics under Clinical and Laboratory Standards Institute (CLSI) guidelines. Among the isolates, 20 *E. coli* had no resistance to any of the tested antibiotics and all of the *P. aeruginosa* strains were resistant to at least 1 drug. Seventy-four *E. coli* and 101 *P. aeruginosa* isolates were resistant to 3 or more antibiotics. The antibiotics tested and the full AST results are summarized in https://github.com/karatsang/rulesbased_logisticregression/tree/v1.0.0/AST. In the *E. coli* dataset there were 30 unique multilocus sequence types (MLSTs) and 5 isolates with unresolved MLST allele(s). The 2 most prevalent *E. coli* MLSTs in the dataset were ST131 and ST1193, which 39 and 10 clinical isolates belonged to, respectively. Notably, ST131 is known to be a major cause of multidrug-resistant *E. coli* infections in the USA [25] and a globally dominant clone [26] associated with CTX-M β-lactamases, while ST1193 is a newer multidrug-resistant *E. coli* clonal group (2017–2019) associated with both CTX-M β-lactamases, plasmid-borne TEM-1 and aminoglycoside acetyltransferases (AACs) [27–29]. In the *P. aeruginosa* dataset there were 59 unique MLSTs (43 known and 16 novel MLSTs) and 3 isolates with unresolved MLST allele(s). The three most prevalent MLSTs,

ST244, ST235 and ST253, were identified in five *P. aeruginosa* isolates each. *P. aeruginosa* ST244 is an international clone, many isolates of which are multidrug-resistant [30, 31], ST235 is amongst the most prevalent of international clones originating from Europe, with regional acquisition of AMR genes [32], and ST253 a less common clone associated with multidrug resistance in Spain and Greece [33]. The full MLST results are summarized in https://github.com/karatsang/rulesbased_logisticregression/tree/v1.0.0/MLST. Raw Illumina DNA sequencing reads for each isolate are available through National Center for Biotechnology Information (NCBI) BioProject PRJNA532924.

Rules-based interpretation leads to poor β-lactam phenotype prediction

Our rules-based algorithm relies on the resistome predictions of CARD's RGI and the genotype–phenotype relationships curated in CARD's ARO. RGI uses four bioinformatics models to predict the resistome of a clinical isolate, which are the protein homology, protein variant, rRNA variant and protein overexpression models (detailed at https://github.com/arpcard/rgi). The protein homology model detects a protein sequence based on its similarity to a curated reference sequence in CARD. The protein variant model builds on the protein homologue model to identify curated mutations that are shown to confer resistance in antibiotic targets, while the rRNA variant model performs the same function for mutations conferring resistance to antibiotics targeting ribosomal RNAs. The protein overexpression model identifies proteins with or without mutations which reflects regulatory proteins that are functional without a mutation, but confer overexpression of their targets with a mutation. As CARD's RGI software is unable to predict multi-component efflux pump systems important for AMR, we developed the Efflux Pump Identifier (EPI) software to interpret RGI results for the prediction of overexpressed efflux pump systems, classifying them into three categories: Perfect, Partial and Putative. The Perfect category identifies sequence matches to CARD for all components of a predicted efflux multi-component system. The Partial category identifies all components of an efflux multi-component system, but at least one component is a sequence variant of CARD's reference sequence. The Putative category predicts potential efflux multi-component systems with missing components or otherwise entirely composed of previously uncharacterized sequence variants.

For our analyses we used the above models and RGI's Perfect and Strict criteria, supplemented with the EPI's interpretation of efflux complexes, to predict resistomes from isolate genome sequences. RGI's Perfect criterion requires that a query protein sequence be identical to a curated reference sequence in CARD, while Strict detects variants of known resistance determinants that pass a curated bit-score cut-off (protein homologue model) or a known AMR-conferring mutation (protein variant model) that can be found curated within CARD (card.mcmaster.ca). The predicted resistomes of the individual *P. aeruginosa* and *E. coli* isolates were generally unique and contained a large diversity of resistance determinants (Table 1, also see https://git.io/JJFh3), with the exceptions being two groups of three *P. aeruginosa* isolates and five *E. coli* isolates that had the same predicted resistome, respectively.

In the *P. aeruginosa* clinical isolate dataset, RGI detected 4 Perfect and 38 Strict, non-efflux, unique resistance genes (protein homologue models) across 34 of CARD's drug classes, plus 4 unique, non-efflux mutations (protein variant models) known to confer resistance to particular antibiotics (ParE A473V, GyrA T83I, BasR L71R and EF-Tu R234F). In the *E. coli* dataset, RGI detected 31 Perfect and 59 Strict non-efflux, unique resistance genes (protein homologue models), plus 15 unique, non-efflux mutations or combinations of mutations (protein variant models) known to confer resistance to particular antibiotics (UhpT E350Q; ParC S80I, E84G; EF-Tu R234F; PBP3 D350N, S357N; GlpT E448K; GyrB S464Y; GyrA D87Y, D87G, D87N, S83L; CyaA S352T; PtsI V25I; NfsA Y45C). For efflux, in *P. aeruginosa* there were 2 unique Perfect and 14 Strict and in *E. coli* there were 11 unique Perfect and 34 Strict protein homologue models representing single-component efflux resistance genes. EPI additionally detected one Perfect or Partial efflux complex with an overexpression mutation (*E. coli* AcrAB-TolC with MarR mutation Y137H conferring resistance to ciprofloxacin and tetracycline) in two different *E. coli* isolates; otherwise, EPI identified six unique Partial efflux pump complexes without an overexpression mutation among the *E. coli* isolates. In contrast, EPI did not identify any Perfect efflux pump complexes among *P. aeruginosa* isolates; however, three unique Partial efflux pump complexes with an overexpression mutation were identified in three different clinical isolates (MexEF-OprN with MexS F253L,V73A; MexAB-OprM with MexR R91C; MexAB-OprM with NalC S209R, G71E, A186T). Supplementary information and citations for all variants predicted by RGI/EPI can be found at CARD.

Comparing the above RGI and EPI resistome predictions, phenotypically classified by CARD's ARO, to the laboratory ASTs, we observed instances of true-positive, true-negative, false-positive and false-negative predictions of AMR phenotype for both *E. coli* and *P. aeruginosa* (Figs 1 and 2).

No antibiotic resistance phenotypes were predicted with 100% accuracy (defined as the percentage of correctly classified phenotypes). Most of the penicillin and cephalosporin (amoxicillin/clavulanic acid, piperacillin/tazobactam, cefazolin, ceftriaxone, ceftazidime, cefixime and meropenem) resistance phenotype predictions resulted in false negatives for both *E. coli* and *P. aeruginosa* (i.e. we failed to predict the observed resistance based on genome sequence). In particular, the prediction of both cefazolin and cefixime resistance phenotypes was less than 2% accurate in the *P. aeruginosa* dataset and less than 57% accurate in the *E. coli* dataset. In addition, for *E. coli* the rules-based algorithm failed to predict any of the observed cefazolin and cefixime resistance based on genome sequence (i.e. not a single true-positive result was obtained).

Table 1. The prevalence of Perfect and Strict resistance determinants detected by the Resistance Gene Identifier, organized by the Antibiotic Resistance Ontology (ARO) drug class designations. Columns show the number and percentage of sampled isolates with at least one AMR determinant associated with resistance to each drug class, broken down as harbouring efflux or non-efflux determinants, or both. For example, 98% of all *P. aeruginosa* isolates had a least one resistance gene for rifamycin resistance, with 99 isolates predicted to have only efflux gene(s) conferring resistance to rifamycin and a single isolate predicted to have only a non-efflux determinant of rifamycin resistance. The total number of *E. coli* and *P. aeruginosa* isolates is 115 and 102, respectively.

ARO drug class	No. of *E. coli* isolates (non-efflux+efflux+both)	% of *E. coli* isolates	No. of *P. aeruginosa* isolates (non-efflux+efflux+both)	% of *P. aeruginosa* isolates
Acridine dye	0+115+0	100.0%	0+102+0	100.0%
Aminocoumarin antibiotic	0+114+1	100.0%	0+101+1	100.0%
Aminoglycoside antibiotic	0+44+71	100.0%	0+0+102	100.0%
Benzalkonium chloride	0+115+0	100.0%	0+1+0	1.0%
Bicyclomycin	0+1+0	0.9%	0+102+0	100.0%
Carbapenem	0+0+115	100.0%	0+0+102	100.0%
Cephalosporin	0+0+115	100.0%	0+0+102	100.0%
Cephamycin	0+0+115	100.0%	0+101+1	100.0%
iaminopyrimidine antibiotic	50+1+3	47.0%	0+101+1	100.0%
Elfamycin antibiotic	115+0+0	100.0%	2+0+0	2.0%
Fluoroquinolone antibiotic	0+42+73	100.0%	0+67+35	100.0%
Fosfomycin	0+111+4	100.0%	102+0+0	100.0%
Fusidic acid	0+1+0	0.9%	0+0+0	0.0%
Glycopeptide antibiotic	0+111+4	3.5%	2+0+0	2.0%
Glycylcycline	0+115+0	100.0%	0+100+0	98.0%
Lincosamide antibiotic	4+68+3	65.2%	3+1+0	3.9%
Macrolide antibiotic	0+60+55	100.0%	0+0+102	100.0%
Monobactam	0+0+115	100.0%	0+0+102	100.0%
Mupirocin	0+0+0	0.0%	1+0+0	1.0%
Nitrofuran antibiotic	115+0+0	100.0%	0+2+0	2.0%
Nitroimidazole antibiotic	0+115+0	100.0%	0+0+0	0.0%
Nucleoside antibiotic	0+112+3	100.0%	0+1+0	1.0%
Nybomycin	72+0+0	62.6%	21+0+0	20.6%
Oxazolidinone antibiotic	0+0+0	0.0%	1+0+0	1.0%
Penam	0+0+115	100.0%	0+0+102	100.0%
Penem	0+65+50	100.0%	0+99+3	100.0%
Peptide antibiotic	0+0+115	100.0%	0+0+0	100.0%
Phenicol antibiotic	0+91+24	100.0%	0+1+101	100.0%
Pleuromutilin antibiotic	39+0+0	33.9%	1+0+0	1.0%
Rhodamine	0+115+0	100.0%	0+1+1	1.0%
Rifamycin antibiotic	0+115+0	100.0%	0+99+1	98.0%
Streptogramin antibiotic	42+0+0	36.5%	3+0+0	2.9%

Continued

Table 1. Continued

ARO drug class	No. of *E. coli* isolates (non-efflux+efflux+both)	% of *E. coli* isolates	No. of *P. aeruginosa* isolates (non-efflux+efflux+both)	% of *P. aeruginosa* isolates
Sulfonamide antibiotic	67+0+0	58.3%	0+94+8	100.0%
Sulfone antibiotic	67+0+0	58.3%	8+0+0	7.8%
Tetracycline antibiotic	0+112+3	100.0%	0+99+3	100.0%
Triclosan	0+114+1	100.0%	0+102+0	100.0%

Logistic regression improves AMR phenotype prediction accuracy

A limitation of the rules-based method is that it only uses known and curated information to predict resistance and is thus inherently blind to any unknown AMR genotype–phenotype relationships. To overcome this limitation, we used logistic regression (LR) to independently identify patterns between RGI-predicted AMR determinants and observed AMR phenotypes. For the *E. coli* dataset (n=115) it was possible to train LR classification models, optimized via cross-validation, and test them on a set of withheld isolates for 14 out of 18 antibiotics (Fig. 1). Due to the relative imbalance of resistant versus susceptible isolates for amikacin, ertapenem, meropenem and nitrofurantoin, models trained for these antibiotics required the use of all isolates, preventing the evaluation of model generalizability on a held-out test set. In the *P. aeruginosa* dataset, piperacillin/tazobactam, ceftazidime, meropenem, ciprofloxacin and gentamicin resistance prediction models were trained and tested on separate isolates, while nitrofurantoin and tetracycline required use of 'dummy' models (i.e. all isolates were intrinsically resistant) and the remainder of the AMR prediction models were trained on all isolates due to unbalanced sampling of resistant and susceptible isolates (Fig. 2).

We evaluated model performance using test set average precision (i.e. trapezoidal area under the precision–recall curve) and a model was categorized as very precise if the test set average precision was ≥0.85, relative to previous studies. Generally, our models were very precise with our *E. coli* data, with a test set average precision of ≥0.85 for all antibiotics except amoxicillin/clavulanic acid (0.811), piperacillin/tazobactam (0.435) and cefoxitin (0.385). In contrast, the *P. aeruginosa* dataset was particularly problematic for LR, with the majority of resistance phenotypes being either ubiquitous (tetracycline and nitrofurantoin) or the less-frequent phenotype representing fewer than 10% of isolates (10/17 antibiotics; ertapenem was not evaluated for these isolates) (Fig. 2). Only five antibiotics had properly fitted and evaluated models for *P. aeruginosa*: ceftazidime, ciprofloxacin, gentamicin, meropenem and piperacillin/tazobactam. These models had either moderate (ciprofloxacin:~0.650), poor (ceftazidime, piperacillin/tazobactam: 0.512, 0.403), or extremely poor (meropenem: 0.227, gentamicin C: 0.196) test set average precision.

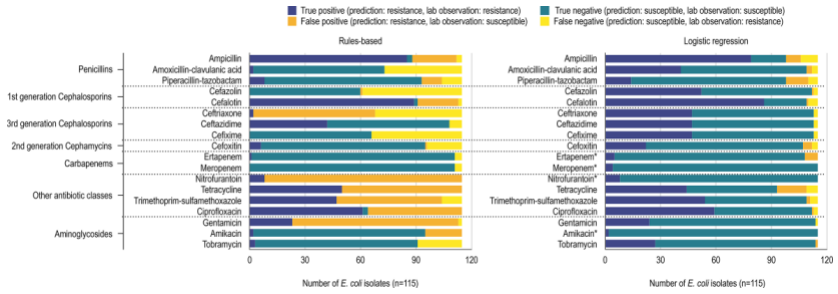


Fig. 1. True-positive, true-negative, false-positive and false-negative predictions of *E. coli* resistance phenotype using a rules-based (left) and logistic regression (right) method. Antibiotic susceptibility tests used 18 antibiotics organized into their respective drug classes. True positives (dark blue) and true negatives (teal) indicate that the classifier predicted resistance and susceptibility correctly. False positives (orange) indicates classifier prediction of resistant but an AST of susceptible. Similarly, false negatives (yellow) indicates classifier prediction of susceptible but an AST of resistant. The rules-based method uses RGI, EPI and the Antibiotic Resistance Ontology to predict resistance phenotypes. Logistic regression classifiers use RGI-detected AMR determinants to predict resistance phenotypes. Logistic regression models for antibiotics for which <10% of a species' isolates displayed susceptible or resistant phenotypes could not be properly validated and tested and as such were trained using all the data (indicated by an asterisk).

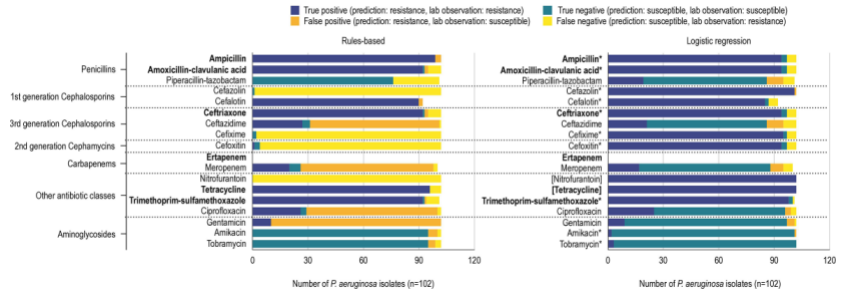


Fig. 2. True-positive, true-negative, false-positive and false-negative predictions of *P. aeruginosa* resistance phenotype using a rules-based (left) and logistic regression (right) method. Antibiotic susceptibility tests used 17 antibiotics (ertapenem was not tested in *P. aeruginosa*) organized into their respective drug classes. Prediction performances for antibiotic logistic regression classifiers using RGI detected AMR determinants to predict resistance phenotypes for *E. coli* and *P. aeruginosa*. True positives (dark blue) and true negatives (teal) indicate that the classifier predicted resistance and susceptibility correctly. False positives (orange) indicates classifier prediction of resistant but an AST of susceptible. Similarly, false negatives (yellow) indicates classifier prediction of susceptible but an AST of resistant. The rules-based method uses RGI, EPI and the Antibiotic Resistance Ontology to predict resistance phenotypes. Logistic regression classifiers use RGI-detected AMR determinants to predict resistance phenotypes. Logistic regression models for antibiotics for which <10% of a species' isolates displayed susceptible or resistant phenotypes could not be properly validated and tested and as such were trained using all the data (indicated by an asterisk). Similarly, when all isolates were resistant or susceptible a 'dummy' model was used, which always returns the relevant label (placed in square brackets). The bolded antibiotics represent antibiotics that *P. aeruginosa* confer intrinsic resistance towards, according to the Clinical and Laboratory Standards Institute (CLSI). The total of *P. aeruginosa* phenotype predictions does not always equal the total number of isolates (*n*=102) because not all isolates were tested against every antibiotic.

Overall, using LR reduced problems of false-positive and false-negative prediction of AMR phenotypes (Figs 1 and 2). For *P. aeruginosa* cefazolin and cefixime resistance phenotypes, where the rules-based approach had very few accurate predictions, LR was able to improve accuracy by 92 and 98%, respectively. Similarly, the rules-based method could not predict any true-positive *E. coli* cefazolin and cefixime resistance phenotypes, whereas LR improved accuracy by 45 and 41%, respectively. In both *P. aeruginosa* and *E. coli* datasets, LR reduced the number of false positives in most tested antibiotic resistance phenotypes compared to the rules-based method. Even in the antibiotic resistance phenotypes where the number of false positives increased, prediction accuracy still improved, e.g. *P. aeruginosa* piperacillin/tazobactam resistance and *E. coli* tobramycin resistance (Figs 1 and 2).

LR models predict novel *β*-lactamase activity

For every antibiotic resistance phenotype, LR assigns every resistance determinant a weight to estimate its relative contribution to overall resistance. We investigated the five most highly weighted predictors for each antibiotic and pathogen to examine the predicted AMR genotype–phenotype relationships. LR weights that confirmed a known relationship (i.e. supported by the published literature and already curated in CARD) for *E. coli* included *CTX-M-15* for ceftazidime resistance, *tet(C)* for tetracycline resistance, *aac(3)-IIb* for gentamicin and tobramycin resistance, *dfrA17* for trimethoprim/sulfamethoxazole resistance, and *gyrA* for ciprofloxacin resistance (Fig. 3a–j) and for *P. aeruginosa* included *mexD* for amoxicillin/clavulanic acid, ceftriaxone, and cefoxitin resistance, *gyrA* for ciprofloxacin resistance, and *mexB* for amikacin resistance (Fig. 3k–o).

A number of the most highly weighted predictors suggested a previously undocumented substrate specificity for a known *β*-lactamase, most notably *CMY-2* conferring resistance to amoxicillin/clavulanic acid and cefazolin, along with *CTX-M-15* conferring resistance to cefixime. To independently test these highly weighted associations, we tested the substrate activity of 11 resistance genes predicted in either the *E. coli* isolates (*aac(6')-Ib-cr*, *CMY-2*, *CTX-M-15*, *CTX-M-3*, *CTX-M-27*, *OXA-1*, *OXA-50*, *TEM-1* and *TEM-30*) or *P. aeruginosa* isolates (*PDC-3* and *PDC-5*) using the Antibiotic Resistance Platform (ARP) [21], concluding clinical resistance based on a ≥2-fold elevation in minimum inhibitory concentration (MIC) compared to control that also passed the CLSI Resistant MIC breakpoint value. In total, 22 previously unknown activities between 7 AMR genes and an antibiotic were experimentally validated as clinically relevant in at least 1 pathogen using the ARP and CLSI breakpoints (Table 2). These included new knowledge for resistance to ampicillin (*CMY-2*, *CTX-M-3*, *CTX-M-27*, *OXA-1* and *TEM-30*), amoxicillin/clavulanic acid (*CMY-2*, *CTX-M-3*, *OXA-1* and *TEM-1*), cefazolin (*CMY-2*, *CTX-M-3*, *CTX-M-15*, *CTX-M-27* and *TEM-1*), cefixime (*CMY-2* and *CTX-M-3*), ceftazidime (*CMY-2*, *CTX-M-3* and *CTX-M-27*), ertapenem (*CTX-M-27*) and

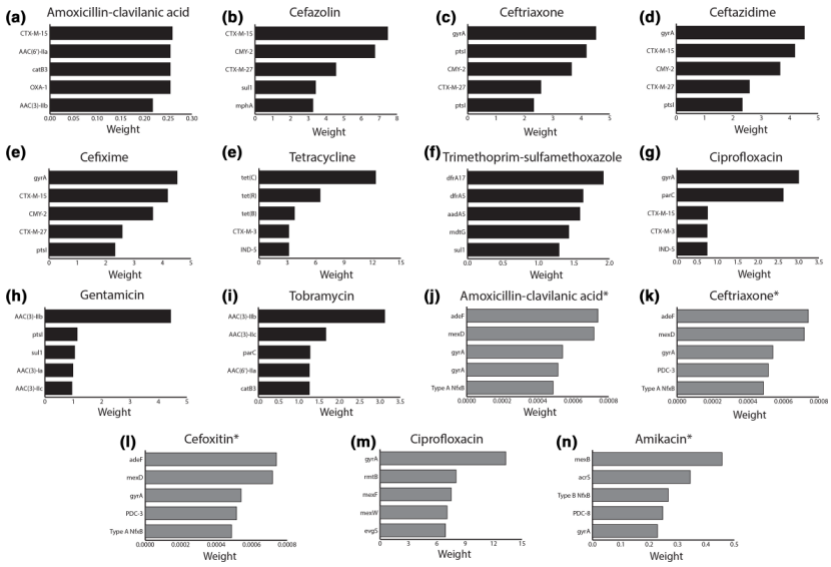


Fig. 3. Logistic regression and RGI identify resistance determinants for predicting *E. coli* and *P. aeruginosa* resistance phenotypes that are supported by the literature. The *x*-axes indicate assigned logistic regression weights for individual AMR phenotype predictions, while the *y*-axes list the top five weighted AMR determinants. Black and grey bars represent *E. coli* and *P. aeruginosa* resistance phenotypes, respectively. An asterisk indicates that <10% of a species' isolates displayed a susceptible or resistant phenotype to amikacin and therefore could not be properly validated and tested, so were trained using all of the data. Models identifying resistance determinants inconsistent with the literature are shown in Figs S4 and S5.

ceftriaxone (*CMY-2* and *CTX-M-3*). However, none of the tested resistance genes explained the observed resistance to meropenem and an additional four genes only confirmed previous knowledge: AAC(6′)-Ib-cr conferring resistance to tobramycin [34], TEM-1 conferring resistance to ampicillin [35], TEM-30 conferring resistance to amoxicillin/clavulanic acid [36] and CTX-M-15 conferring resistance to ceftriaxone (Table S1, available in the online version of this article) [37]. ASTs also invalidated some predictions, e.g. *CTX-M-15* conferring clinically relevant resistance towards cefixime and ceftazidime. Notably, while *OXA-50* is reported to elevate the MIC towards ampicillin and cefotaxime when cloned into a multicopy plasmid and expressed in *P. aeruginosa,* like others [38], we did not observe any appreciable elevation in MIC compared to control in *E. coli* (data not shown). Overall, LR combined with AST validation provided a wealth of new knowledge on antibiotic specificities for β-lactamases appearing in clinical isolates. Interestingly, incorporation of these results into the rules-based algorithm improved resistance prediction in *E. coli* for cefazolin (75% improvement in true-positive results) and cefixime (31% improvement in true-positive results) (Fig. 4) plus in *P. aeruginosa* for cefixime (34% improvement in true-positive results) and cefoxitin (35% improvement in true-positive results) (Fig. S6), illustrating the sensitivity of rules-based methods to available knowledge. Yet, even with this new knowledge, the rules-based algorithm was still outperformed by the LR approach.

DISCUSSION

Fast and accurate prediction of AMR phenotypes from genotypes would improve AMR surveillance, patient outcomes and antibiotic stewardship. Currently, our ability to diagnose bacterial infections is costly and slow, contributing to the misuse and overuse of antibiotics, as well as to poor clinical outcomes. Genotypic approaches using whole-genome sequencing paired with bioinformatics resources have the potential to be a faster and more accurate method. The goal of this study was to identify and elucidate β-lactamase substrate activity, a limiting factor in AMR phenotype prediction, by using two different *in silico* AMR phenotype prediction algorithms, subsequently validated using targeted gene expression experiments. In the rules-based method, we developed EPI to be used in combination with RGI to better identify overexpressed multi-component efflux pumps, while the LR method only used the resistance determinants predicted by RGI as its

Table 2. Antibiotic susceptibility testing (AST) of known resistance genes predicted to have previously undescribed activity. As per the Antibiotic Resistance Platform, AMR genes were cloned into the pGDP plasmid series and transformed into wild-type *E. coli* BW25113, which is representative of a clinical isolate. AST was performed for each construct using the microdilution broth method, with the inoculum prepared using the growth method following CLSI guidelines.

Antibiotic	Resistance gene	Plasmid	MIC (μg ml^{-1}) wild-type *E. coli* BW25113	CLSI resistant MIC (μg ml^{-1}) breakpoint for *Enterobacteriaceae*	CLSI resistant MIC (μg ml^{-1}) breakpoint for *P. aeruginosa*
Ampicillin	None	None	64	≥32	–
	CMY-2	pGDP1	>256	≥32	–
	CTX-M-3	pGDP1	>256	≥32	–
	CTX-M-27	pGDP1	>256	≥32	–
	OXA-1	pGDP1	>256	≥32	–
	TEM-30	pGDP1	>256	≥32	–
Amoxicillin/clavulanic acid	None	None	8–16	≥32/16	–
	CMY-2	pGDP1	256	≥32/16	–
	CTX-M-3	pGDP1	64	≥32/16	–
	CTX-M-15	pGDP1	16	≥32/16	–
	OXA-1	pGDP1	64	≥32/16	–
	TEM-1	pGDP1	128	≥32/16	–
Cefazolin	None	None	4	≥8/≥32 (urine only)	–
	CMY-2	pGDP1	>256	≥8/≥32 (urine only)	–
	CTX-M-3	pGDP1	>256	≥8/≥32 (urine only)	–
	CTX-M-27	pGDP1	>256	≥8/≥32 (urine only)	–
	TEM-1	pGDP1	256	≥8/≥32 (urine only)	–
Cefixime	None	None	0.25	≥4	–
	CMY-2	pGDP1	>256	≥4	–
	CTX-M-3	pGDP1	32	≥4	–
Ceftazidime	None	None	0.5	≥16	≥32
	CMY-2	pGDP1	256	≥16	NR
	CTX-M-3	pGDP1	16–32	≥16	NR
	CTX-M-27	pGDP1	128	≥16	NR
Ertapenem	None	None	0.25	≥2	–
	CTX-M-27	pGDP1	128	≥2	–
Ceftriaxone	None	None	0.25	≥4	–
	CMY-2	pGDP1	128	≥4	–
	CTX-M-3	pGDP1	>256	≥4	–

–, no CLSI breakpoint for *P. aeruginosa* due to intrinsic resistance; NR, not relevant as CMY-2, CTX-M-3, and CTX-M-27 were only identified in *P. aeruginosa*.

starting point. While naïve about the relative contribution of individual resistance determinants to overall resistance and sensitive to any gaps in knowledge for β-lactamase activity, the rules-based method nonetheless was able to accurately predict a number of resistance phenotypes when they involved well-characterized resistance determinants that confer resistance surpassing clinical breakpoints, e.g. AAC(6′)-Ib-cr for tobramycin. In terms of false-positive predictions using this approach, we hypothesize that CARD contains incorrect genotype–phenotype information,

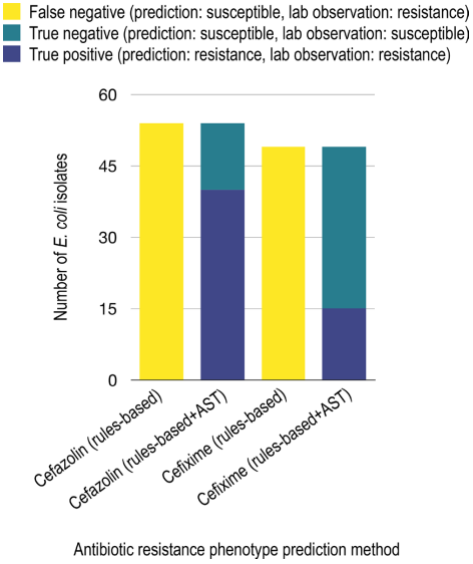


Fig. 4. Improvement of *E. coli* cefazolin and cefixime resistance prediction using rules-based algorithm and substrate activity knowledge gained from antibiotic susceptibility testing (AST). Through antibiotic susceptibility testing, we observed CTX-M-3, CTX-M-27 and CMY-2 conferring clinically relevant resistance to cefazolin and cefixime. Curating this knowledge into CARD would improve cefazolin and cefixime true positive resistance prediction in *E. coli* by 74.1 and 30.6%, respectively.

an environmental factor is altering the expression of a predicted resistance determinant, or that CARD has a knowledge gap regarding repressors. With the first scenario, removal of incorrect curation could decrease instances of false positives, highlighting one of the limitations of human biocuration for AMR phenotype prediction. The second scenario, i.e. adaptive resistance, should not be a concern for our study, since our antibiotic susceptibility tests were standardized and automated, notwithstanding potential inconsistencies affecting gene expression [39]. The third scenario suggests that there are gaps in the literature, as CARD only includes information published in peer-reviewed literature with clear experimental evidence of elevated resistance. Genetic determinants that decrease the expression or change the substrate profile of a resistance determinant, such as mutations within regulatory regions or active sites, would result in false-positive predictions. Alternatively, entirely unknown resistance genes or mutations could explain false-negative predictions of AMR phenotypes.

To identify relationships between known resistance genes and resistance phenotypes without relying on CARD's ARO for curated genotype–phenotype relationships, we used RGI in combination with LR. It is important to note that accurate and generalizable LR-based prediction of susceptibility or resistance to an antibiotic from detected AMR determinants is only feasible when there are relatively large numbers of genomes exemplifying each phenotype, which was not always the case in our data. Even with stratified sampling and methods, such as SMOTE [40], to resample datasets and improve balance (e.g. the relative proportion of susceptible and resistant isolates) there are limitations to what can be achieved with small datasets that are predominantly resistant or susceptible to a given antibiotic. Models that are not properly tested are likely to overfit to the data and are unlikely to generalize well for new data, in our case samples from outside the Hamilton, Ontario area. Additional validation of our models using publicly available data is important for future studies; models may be dependent on feature selection, taxonomic distribution, resistance mechanism and algorithm choice. Yet, despite the models not being appropriately tested properly due to imbalance, LR proved a useful tool for improving prediction of resistance from genomic features, even without the rules-based algorithm's additional consideration of overexpressed, multi-component efflux pumps. LR substantially decreased instances of false positives or false negatives, and the poor performance for predicting particular resistance phenotypes (e.g. tetracycline resistance in *E. coli*, ceftazidime resistance in *P. aeruginosa* and piperacillin/tazobactam resistance in both species) could either represent a failure of the LR algorithm to capture the combination of resistance determinants required to predict resistance due to additive or synergistic resistance or to recognize undiscovered resistance determinants not in CARD and thus not predicted by RGI.

While bioinformatics tools such as breseq [41] or *k*-mer approaches combined with LR could be used to potentially identify unknown mutations or functional gene loss (e.g. OprD loss is associated with imipenem, meropenem and doripenem resistance [42]), our prediction of CLSI [43] 'resistant' and 'susceptible' resistance phenotypes places limits upon interpretation, as other clinical breakpoint guidelines exist, e.g. the European Committee on Antimicrobial Susceptibility Testing (EUCAST) [44] breakpoint guidelines are based on interpretation of quantitative MIC values, which unfortunately are not recorded in CARD or any other database for the breadth of known resistance genes and mutations. As such, detection of a CARD resistance determinant in a clinical isolate was interpreted as 'resistant', even though in reality the MIC value generated by the gene may not have reached the CLSI or EUCAST breakpoints for resistant. Nonetheless, aligning with George E. P. Box's aphorism, 'all models are wrong, but some are useful' [45], our goal was to identify the LR models with 'useful' or logical biological relevance with a focus on prevalent clinical β-lactamases. Prediction of genomic determinants responsible for resistance based on the feature weights of the LR only made biological sense in some cases based on the literature and knowledge. For example, *novA* was the highest weighted predictor for

P. aeruginosa trimethoprim/sulfamethoxazole resistance, but is known to instead be involved in the transport of and resistance to novobiocin [46]. Failure to predict logical determinants could be attributed to high levels of divergence from the canonical sequence or an unknown resistance determinant with prevalence correlated with *novA*. In the balanced datasets, known relationships in CARD, such as *tet(C)* conferring resistance to tetracycline in *E. coli* and *P. aeruginosa gyrA* mutation conferring resistance to ciprofloxacin, were predicted by both the rules-based and LR methods (Fig. 3f, n). Beyond this, LR was additionally able to predict genotype–phenotype relationships that were useful in that they were new findings not predicted by the rules-based method and not published in the literature, yet consistent with known resistance mechanisms. Indeed, there is value in looking beyond the most highly weighted LR predictor, since analysis of a model can garner major insights into AMR genotype–phenotype relationships. We were able to experimentally validate many of the top five most highly weighted candidates, illustrating that systematic screening of a broad selection of antibiotics against known resistance genes using molecular AST platforms such as the ARP [21], perhaps guided by LR, or at minimum community adoption of standard panels of antibiotics for AST characterization of newly reported resistance genes, could be adopted to fill these gaps in the literature and improve antibiotic resistance phenotype prediction.

We have illustrated that completely accurate AMR phenotype prediction is not achievable using either rules-based or LR methods. There are likely unknown genomic determinants leading to both false-positive and false-negative prediction of resistance phenotypes, such as mutations in regulatory regions that change expression of a resistance gene. Overall, our results suggest that LR is capable of predicting resistance phenotypes and identifying substrate specificities of known resistance genes when there are sufficiently balanced datasets. Evaluating learned weights for each LR model led to novel hypotheses, illustrating the use of LR as an inductive approach to guide deductive research. Yet, our results also illustrate that full prediction of resistome and resistance phenotype will require careful examination of genome feature space and clinical breakpoints, plus broad and balanced sampling of diverse susceptible and resistant strains. It is our hope that collective advances in these methods will result in tools for clinical prediction of resistance, aiding antimicrobial stewardship and improving patient outcomes. Elucidating AMR genotype–phenotype relationships will reveal the genetic and mechanistic underpinnings of resistance to guide both public health surveillance and future drug discovery.

METHODS

Bacterial isolates, antibiotic susceptibility testing, and DNA extraction

Clinical bacterial isolates were obtained from the IIDR Clinical Isolate Collection, which consists of isolates from the core clinical laboratory at Hamilton Health Sciences, Hamilton, Ontario. Samples were collected between 2015 and 2018 and were resistant to 3 or more antibiotics based on antimicrobial susceptibility to 18 and 17 antibiotics for *E. coli* and *P. aeruginosa*, respectively. As ertapenem lacks activity against *P. aeruginosa* [47], it was not included in *P. aeruginosa* antibiotic susceptibility tests. Initial culture and antibiotic susceptibility testing (AST) were performed by Hamilton General Hospital General Microbiology Laboratory using a VITEK 2 Automated System and its Advanced Expert System (BioMérieux, Marcy-l'Étoile, France), compliant with the Clinical and Laboratory Standards Institute (CLSI) [43] antibiotic susceptibility testing formulations, reporting CLSI breakpoint-determined susceptible (S), intermediate (I), or resistant (R). For DNA extraction, isolates were provided on blood agar plates and single colonies were restreaked onto brain heart infusion (BHI) agar. After overnight incubation, single colonies of each isolate were used to inoculate Luria–Bertani (LB) broth. Overnight broth cultures were used to prepare glycerol stocks for long-term storage at −80 °C. One millilitre of the same overnight cultures was centrifuged, the supernatant was removed and the pellet was stored at −80 °C for genomic DNA extraction. The Invitrogen Pure Link Genomic DNA Mini kit (K182002) was used for DNA extraction from pellets. DNA was eluted with water and stored at 4 °C.

Whole-genome sequencing, assembly and species identification

DNA sequencing library construction (Illumina Nextera XT DNA Library Preparation kit or NEBNext Ultra II DNA Library Preparation kit) and all sequencing runs were performed at the Farncombe Metagenomics Facility at McMaster University using 2×150 bp paired-end sequencing on an Illumina HiSeq 1500 platform (*E. coli n*=115, *P. aeruginosa n*=92) or 2×250 bp paired-end sequencing on an Illumina MiSeq v3 platform (*P. aeruginosa n*=10). Paired sequencing reads were trimmed using Trimmomatic (v0.36) [48], checked for quality using FASTQC (v0.11.8, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [49] and *de novo* assembled using SPAdes (v3.9.0) [50]. The Livermore Metagenomics Analysis Toolkit (LMAT, v1.2.6) [51] was used to confirm bacterial species and screen for contamination or mixed culture. For *E. coli*, after quality trimming of the sequencing reads by Trimmomatic, sequencing isolate read coverage averaged 207.5-fold, assembly size averaged ~5163879 bp and N50s averaged 231879 bp. For *P. aeruginosa*, quality-trimmed sequencing read coverage averaged 100.6-fold, assembly sizes averaged 6680703 bp and assembly N50s averaged 260849 bp. Diversity of isolates for both *E. coli* and *P. aeruginosa* was assessed by multilocus sequence typing (MLST) via comparison to the reference sequences available at pubMLST (https://github.com/agmcarthur/pubMLST) [52].

Curation of CARD

At minimum, CARD requires the curation of a '*confers_resistance_to_drug_class*' relationship between an AMR gene family and a drug class in the ARO. However, to predict specific drug resistance phenotypes we needed curation of a '*confers_resistance_to_antibiotic*' relationship between an individual resistance gene or mutation and a specific antibiotic. The curation of '*confers_resistance_to_antibiotic*'

relationships is incomplete in CARD and is determined by experimental evidence of an elevation of MIC in the published literature [7]. Using extensive literature review, we curated '*confers_resistance_to_antibiotic*' relationships for all resistance determinants identified as RGI Perfect or Strict RGI hits for our *E. coli* and *P. aeruginosa* isolates: an additional 250 '*confers_resistance_to_antibiotic*' relationships (152 *E. coli* and 98 *P. aeruginosa*) were added to CARD (available as of v2.0.2). During the curation process we also identified two errors in CARD curation. These included incorrect inclusion of mutation Y45C in the *E. coli* protein *NfsA* as conferring resistance to nitrofurantoin and the *β*-lactamase gene *SHV-1* as conferring resistance to cefazolin. In both cases, the original publications lacked clear experimental support for these claims.

To additionally improve efflux pump prediction and facilitate the functionality of the Efflux Pump Identifier (EPI), *E. coli* and *P. aeruginosa* efflux meta-models (a combination of individual models) were curated into CARD v1.1.9, based on review of the literature. Efflux meta-models comprise protein homologue and/or protein overexpression models to represent a known efflux pump complex and its regulatory network. For example, the *AcrAB-TolC* efflux system (ARO:3000384) is encoded along with its regulatory network: *marR*, *marA*, *acrR*, *sdiA*, *soxS*, *soxR*, *rob*. In this meta-model, each component is a protein homologue model with the exception of *marR*, *acrR* and *soxS*, which are protein overexpression models. We curated 21 *P. aeruginosa* efflux pump meta-models, 10 *E. coli* efflux pump meta-models and 2 plasmid-borne efflux pump meta-models known to confer resistance to the 18 antibiotics tested in this study for analysis by EPI.

Rules-based prediction of antibiotic susceptibility phenotypes

Isolate genomes were analysed using the Comprehensive Antibiotic Resistance Database (v2.0.2) and Resistance Gene Identifier (v4.1.0) [7], plus the new EPI (v1.0.0) software developed by KKT, to predict resistance determinants. The EPI predicts multi-component efflux pumps and their regulatory networks using the efflux meta-models curated in CARD (https://git.io/JJFhT). RGI and EPI results were filtered to only include RGI Perfect and Strict hits, and EPI Perfect and Partial hits, respectively. Antibiotic susceptibility phenotypes were predicted by traversing CARD's Antibiotic Resistance Ontology (ARO) to identify the antibiotic(s) each detected resistance determinant confers resistance to, based on peer-reviewed literature. In this rules-based method, the detection of a resistant determinant by RGI or EPI that had a '*confers_resistance_to_antibiotic*' relationship to an antibiotic in the ARO resulted in a 'resistant' phenotype prediction, otherwise a 'susceptible' phenotype was predicted. Computational antibiotic susceptibility predictions were then compared to clinical ASTs. As AST 'intermediate' resistances were rare (2.2% of *P. aeruginosa* resistance phenotypes and 3.6% of *E. coli* resistance phenotypes), we treated them as 'resistant' in our analyses.

Using logistic regression to predict antibiotic resistance phenotypes

To prepare the datasets, all RGI results for each species were collated into count matrices X_{ij} where i represents each genome of that species and j represents a specific AMR determinant detected by RGI at either Strict or Perfect cut-offs. The most appropriate algorithm for phenotype prediction was determined using the *E. coli* data, as these comprised the more balanced dataset. For each antibiotic, the resampled training data were used to fit four interpretable binary classification models: logistic regression, multinomial naïve Bayes, decision tree and random forest classifiers [53]. For each model the hyperparameters were then tuned using a threefold stratified shuffle split cross-validation scheme and evaluated using a negative log loss scoring function [53], as negative log loss considers prediction uncertainty in relation to the divergence of the predicted probabilities and the actual AMR phenotype. Logistic regression and random forest classifiers had the highest performance of all tested modelling methods, so we chose logistic regression, a simpler algorithm, as our classification paradigm under the principle of parsimony. To predict each antibiotic resistance phenotype, antibiotic-specific LR models were trained, optimized via cross-validation and tested separately for each species dataset. To determine whether each species and antibiotic dataset was phenotypically balanced enough for LR, the relative proportion of resistant predictions to susceptible predictions was evaluated. If the less frequent phenotype represented <10% of all genomes it was considered inappropriate to train and properly test a model due to extreme class imbalance and low signal. For these antibiotics an 'unbalanced classifier' was trained and evaluated using all genomes of that species. Some antibiotics displayed an even more extreme case of imbalance where only a single phenotype was observed. For these, a 'dummy' model was used that only returned the observed phenotype (i.e. all observed isolates were resistant to an antibiotic and therefore the model always predicts resistance). For the remaining species-antibiotics combinations with greater label balance, 20% of the genomes were randomly selected with stratification (i.e. maintaining the relative proportion of susceptible to resistant) and withheld as a test set. The training set was then rebalanced using the synthetic minority over-sampling technique (SMOTE) [40] as implemented in imbalanced-learn (v0.3.3) [54] to generate a training set with equal proportions of susceptible and resistant genomes. After training of the *E. coli* models, the *P. aeruginosa* training data were used to fit and optimize logistic regression models via the same threefold stratified cross-validation scheme.

The individual trained antibiotic–species logistic regression models (including unbalanced and dummy classifiers) were evaluated against the test set to see if they could predict AMR phenotype, with evaluation using precision–recall curves (summarized as average precision) and the receiver operating characteristic (summarized as area under the curve) (Figs S1–S3) [55]. A test with perfect discrimination between resistance and susceptible resistance phenotypes would have a receiver operating characteristic curve that

passes through the upper-left corner (Figs S1 and S2). For each species the number of true positives, true negatives, false positives and false negatives was tallied and plotted for each antibiotic. To evaluate which resistance determinants within each classifier were important for predicting resistance phenotypes, we considered the estimated coefficients (scikit-learn's coef_attribute) as the 'weight of importance' for each resistance determinant. Thus, given two resistance determinants, each with an estimated coefficient value, the resistant determinant with a larger estimated coefficient value was interpreted as more important for predicting a particular resistance phenotype. The five most highly weighted predictors of each resistance phenotype were examined (Figs S4 and S5), but all feature weights of importance and their *P*-values were inspected and are listed in Tables S2–S5.

Antibiotic susceptibility testing (AST) using the Antibiotic Resistance Platform

In cases where we wished to perform AST for individual resistance genes, we cloned these genes into pGDP1/pGDP3 from the Antibiotic Resistance Platform [21] and transformed into wild-type *E. coli* BW25113. AST was performed for *E. coli* BW25113 using the microdilution broth method, with the inoculum prepared using the growth method following CLSI guidelines [43]. Plates were sealed in a bag and incubated for 18 h at 37 °C, 250 r.p.m. before the optical density at 600 nm was measured using the Spectramax microplate reader.

Software availability

CARD data and RGI software are available at the CARD website, http://card.mcmaster.ca. CARD (v2.0.2) and RGI (v4.1.0) were used for all resistome prediction, and RGI (v.5.1.0) was used for creating the heatmaps. The EPI software is available at https://github.com/karatsang/rulesbased_logisticregression/tree/v1.0.0/rulesbased/EffluxPumpIdentifier. LR and dataset partitioning were performed using scikit-learn (v0.20.0) [53] with data otherwise manipulated using numpy (v1.17.2) [56] and pandas (v0.25.1) [57]. For both datasets, the code, conda environments (using python v3.7.2 [58]), and intermediate data files required to generate this analysis are available: https://github.com/karatsang/rulesbased_logisticregression, https://doi.org/10.5281/zenodo.3988480.

Funding information
This research was funded by the Canadian Institutes of Health Research (PJT-156214 to A. G. M., MT-14981 to G. D. W.), the Ontario Research Fund (to G. D. W.), Genome Canada (to R. G. B.), a Canada Research Chair to G. D. W. and a Cisco Research Chair in Bioinformatics to A. G. M., supported by Cisco Systems Canada, Inc. K. K. T. was supported by an Ontario Graduate Scholarship, McMaster University's MacDATA Institute Graduate Fellowship and Michael G. DeGroote Institute for Infectious Disease Research Michael Kamin Hart Memorial Scholarship. F. M., was supported by a Donald Hill Family Fellowship in Computer Science. Computer resources were supplied by the McMaster Service Lab and Repository computing cluster, funded in part by grants to A. G. M. from the Canadian Foundation for Innovation (34531).

Acknowledgements
We would like to acknowledge Linda Ejim for clinical isolate culturing, Susan McCusker for genomic DNA isolation from clinical isolates, Biren Dave for the initial development of the isolate assembly pipeline, Arjun Sharma for developing the parser for species identification, Amos Raphenya for Resistance Gene Identifier software development and Brian Alcock for Comprehensive Antibiotic Resistance Database curation.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
Not required as all samples were from an existing clinical bacterial isolate collection without any associated patient information.

References

1. World Health Organization. Antimicrobial resistance: global report on surveillance. World Health organization report. *Geneva* 2014.
2. U.S. Department of Health and Human Services CDC. Antibiotic resistance threats in the United States, 2019. *Atlanta, GA, USA* 2019.
3. Maugeri G, Lychko I, Sobral R, Roque ACA. Identification and antibiotic-susceptibility profiling of infectious bacterial agents: a review of current and future trends. *Biotechnol J* 2019;14:e1700750.
4. Maurer FP, Christner M, Hentschke M, Rohde H. Advances in rapid identification and susceptibility testing of bacteria in the clinical microbiology laboratory: implications for patient care and antimicrobial stewardship programs. *Infect Dis Rep* 2017;9:6839.
5. Chan K-G. Whole-genome sequencing in the prediction of antimicrobial resistance. *Expert Rev Anti Infect Ther* 2016;14:617–619.
6. Crofts TS, Gasparrini AJ, Dantas G. Next-generation approaches to understand and combat the antibiotic resistome. *Nat Rev Microbiol* 2017;15:422–434.
7. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M *et al.* CARD 2020: antibiotic resistome surveillance with the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res* 2019;10:D517–D525.
8. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M *et al.* ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother* 2014;58:212–220.
9. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T *et al.* Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res* 2017;45:D535–D542.
10. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S *et al.* Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012;67:2640–2644.
11. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* 2015;9:207–216.
12. McArthur AG, Tsang KK. Antimicrobial resistance surveillance in the genomic age. *Ann N Y Acad Sci* 2017;1388:78–91.
13. van Belkum A, Bachmann TT, Lüdke G, Lisby JG, Kahlmeter G *et al.* Developmental roadmap for antimicrobial susceptibility testing systems. *Nat Rev Microbiol* 2019;17:51–62.
14. Cantu C, Huang W, Palzkill T. Cephalosporin substrate specificity determinants of TEM-1 β-lactamase. *J Biol Chem* 1997;272:29144–29150.
15. Chiou J, Leung TYC, Chen S. Molecular mechanisms of substrate recognition and specificity of New Delhi metallo-β-lactamase. *Antimicrob Agents Chemother* 2014;58:5372–5378.
16. Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E *et al.* Capturing the mutational landscape of the β-lactamase TEM-1. *Proc Natl Acad Sci U S A* 2013;110:13067–13072.
17. Khan S, Sallum UW, Zheng X, Nau GJ, Hasan T. Rapid optical determination of β-lactamase and antibiotic activity. *BMC Microbiol* 2014;14:84.
18. Lee D, Das S, Dawson NL, Dobrijevic D, Ward J *et al.* Novel computational protocols for functionally classifying and characterising serine beta-lactamases. *PLoS Comput Biol* 2016;12:e1004926.

19. Majiduddin FK, Palzkill T. Amino acid residues that contribute to substrate specificity of class a β-lactamase SME-1. *Antimicrob Agents Chemother* 2005;49:3421–3427.
20. Palzkill T. Structural and mechanistic basis for extended-spectrum drug-resistance mutations in altering the specificity of TEM, CTX-M, and KPC β-lactamases. *Front Mol Biosci* 2018;5:16.
21. Cox G, Sieron A, King AM, De Pascale G, Pawlowski AC *et al.* A common platform for antibiotic dereplication and adjuvant discovery. *Cell Chem Biol* 2017;24:98–109.
22. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C *et al.* Antimicrobial resistance prediction in PATRIC and RAST. *Sci Rep* 2016;6:27930.
23. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M *et al.* Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics* 2016;17:754.
24. Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S *et al.* Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in Gram-negative bacilli from whole genome sequence data. *Front Microbiol* 2016;7:7.
25. Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M. *Escherichia coli* sequence type ST131 as the major cause of serious multidrug-resistant *E. coli* infections in the United States. *Clin Infect Dis* 2010;51:286–294.
26. Pitout JDD, DeVinney R. *Escherichia coli* ST131: a multidrug-resistant clone primed for global domination. *F1000Res* 2017;6:195.
27. Tchesnokova VL, Rechkina E, Larson L, Ferrier K, Weaver JL *et al.* Rapid and extensive expansion in the United States of a new multidrug-resistant *Escherichia coli* clonal group, sequence type 1193. *Clin Infect Dis* 2019;68:334–337.
28. Wu J, Lan F, Lu Y, He Q, Li B. Molecular characteristics of ST1193 clone among phylogenetic group B2 non-ST131 Fluoroquinolone-Resistant *Escherichia coli*. *Front Microbiol* 2017;8:2294.
29. Xia L, Liu Y, Xia S, Kudinha T, Xiao S-N *et al.* Prevalence of ST1193 clone and IncI1/ST16 plasmid in *E-coli* isolates carrying $bla_{CTX-M-55}$ gene from urinary tract infections patients in China. *Sci Rep* 2017;7:44866.
30. Chen Y, Sun M, Wang M, Lu Y, Yan Z. Dissemination of IMP-6-producing *Pseudomonas aeruginosa* ST244 in multiple cities in China. *Eur J Clin Microbiol Infect Dis* 2014;33:1181–1187.
31. Empel J, Filczak K, Mrowka A, Hryniewicz W, Livermore DM *et al.* Outbreak of *Pseudomonas aeruginosa* infections with PER-1 extended-spectrum beta-lactamase in Warsaw, Poland: further evidence for an international clonal complex. *J Clin Microbiol* 2007;45:2829–2834.
32. Treepong P, Kos VN, Guyeux C, Blanc DS, Bertrand X *et al.* Global emergence of the widespread *Pseudomonas aeruginosa* ST235 clone. *Clin Microbiol Infect* 2018;24:258–266.
33. Koutsogiannou M, Drougka E, Liakopoulos A, Jelastopulu E, Petinaki E *et al.* Spread of multidrug-resistant *Pseudomonas aeruginosa* clones in a university hospital. *J Clin Microbiol* 2013;51:665–668.
34. Robicsek A, Strahilevitz J, Jacoby GA, Macielag M, Abbanat D *et al.* Fluoroquinolone-Modifying enzyme: a new adaptation of a common *aminoglycoside acetyltransferase*. *Nat Med* 2006;12:83–88.
35. Sutcliffe JG. Nucleotide sequence of the ampicillin resistance gene of *Escherichia coli* plasmid pBR322. *Proc Natl Acad Sci U S A* 1978;75:3737–3741.
36. Belaaouaj A, Lapoumeroulie C, Caniça MM, Vedel G, Névot P *et al.* Nucleotide sequences of the genes coding for the TEM-like beta-lactamases IRT-1 and IRT-2 (formerly called TRI-1 and TRI-2). *FEMS Microbiol Lett* 1994;120:75–80.
37. Poirel L, Gniadkowski M, Nordmann P. Biochemical analysis of the ceftazidime-hydrolysing extended-spectrum beta-lactamase CTX-M-15 and of its structurally related beta-lactamase CTX-M-3. *J Antimicrob Chemother* 2002;50:1031–1034.
38. Girlich D, Naas T, Nordmann P. Biochemical characterization of the naturally occurring oxacillinase OXA-50 of *Pseudomonas aeruginosa*. *Antimicrob Agents Chemother* 2004;48:2043–2048.
39. Fernández L, Hancock REW. Adaptive and mutational resistance: role of porins and efflux pumps in drug resistance. *Clin Microbiol Rev* 2012;25:661–681.
40. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Jair* 2002;16:321–357.
41. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol* 2014;1151:165–188.
42. Ocampo-Sosa AA, Cabot G, Rodríguez C, Roman E, Tubau F *et al.* Alterations of oprD in carbapenem-intermediate and -susceptible strains of *Pseudomonas aeruginosa* isolated from patients with bacteremia in a Spanish multicenter study. *Antimicrob Agents Chemother* 2012;56:1703–1713.
43. Clinical and Laboratory Standards Institute. *M100: Performance Standards for Antimicrobial Susceptibility Testing*, 28th ed.
44. The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters, version 10.0 2020.
45. Box GEP. Science and statistics. *J Am Stat Assoc* 1976;71:791–799.
46. Schmutz E, Mühlenweg A, Li S-M, Heide L. Resistance genes of aminocoumarin producers: two type II topoisomerase genes confer resistance against coumermycin A1 and clorobiocin. *Antimicrob Agents Chemother* 2003;47:869–877.
47. Livermore DM, Sefton AM, Scott GM. Properties and potential of ertapenem. *J Antimicrob Chemother* 2003;52:331–344.
48. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
49. Andrews S. FastQC: a quality control tool for high throughput sequence data 2010.
50. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
51. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB *et al.* Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics* 2013;29:2253–2260.
52. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018;3:124.
53. Pedregosa F, Varoquaux G, Gramfort A, Michel V. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–2830.
54. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* 2017;18:559–563.
55. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432.
56. Oliphant TE. *A Guide to NumPy*, 1. Trelgol Publishing; 2006. p. 85.
57. McKinney W. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference* 2010:51–56.
58. van Rossum G, Drake FL. Python language reference manual. python language reference manual. *Network Theory Ltd* 2003.