Assessing computational predictions of antimicrobial resistance phenotypes from microbial genomes

Kaixin Hu[1,2], Fernando Meyer[1,2], Zhi-Luo Deng[1,2], Ehsaneddin Asgari[1,3], Tzu-Hao Kuo[1,2], Philipp C. Münch[1,2,4,5,6] and Alice C. McHardy[1,2*]

[1] Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany

[2] Braunschweig Integrated Centre of Systems Biology (BRICS), Technische Universität Braunschweig, Braunschweig, Germany

[3] Molecular Cell Biomechanics Laboratory, Department of Bioengineering and Mechanical Engineering, University of California, Berkeley, USA

[4] Cluster of Excellence RESIST (EXC 2155), Hannover Medical School, Hannover, Germany

[5] German Center for Infection Research (DZIF), partner site Hannover Braunschweig, Braunschweig, Germany

[6] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

[*] E-mail: Alice.McHardy@helmholtz-hzi.de

Abstract

The advent of rapid whole genome sequencing has created new opportunities for computational prediction of antimicrobial resistance (AMR) phenotypes directly from genomic data. Both rule-based and machine learning (ML) approaches have been explored for this task, but systematic benchmarking to evaluate the performance of these methods is still needed. Here, we evaluated four state-of-the-art ML methods (Kover, PhenotypeSeeker, Seq2Geno2Pheno, Aytan-Aktug), a ML baseline and the rule-based ResFinder, by training and testing each of them across 78 species-antibiotic datasets, using a rigorous benchmarking workflow that integrates three evaluation approaches, each paired with three distinct sample splitting methods. Our analysis revealed considerable performance variation across techniques and datasets. While ML methods generally excelled for closely related strains, ResFinder better handled divergent genomes. Overall, Kover most frequently ranked top among ML approaches, followed by PhenotypeSeeker and Seq2Geno2Pheno. AMR phenotypes for antibiotic classes like macrolides, and sulfonamides, were predicted with the highest accuracies. Prediction quality varied substantially across species-antibiotic combinations particularly for beta-lactams; across species, resistance phenotyping of the beta-lactams compound, aztreonam, amox-clav, cefoxitin, ceftazidime,

and piperacillin/tazobactam, together with tetracyclines demonstrated more variable performances than the other benchmarked antibiotics. By organism, *C. jejuni* and *E. faecium* phenotypes were more robustly predicted than *Escherichia coli*, *Staphylococcus aureus*, *Salmonella enterica*, *Neisseria gonorrhoeae*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *Acinetobacter baumannii*, *Streptococcus pneumoniae*, and Mycobacterium tuberculos*is*. In addition, our study provides software recommendations for each species-antibiotic combination. It furthermore highlights optimization needs for robust clinical application, particularly for strains that diverge substantially from those used for training.

Key words: AMR; antimicrobial resistance; machine learning; benchmarking; phenotype prediction

Introduction

The efficacy of antimicrobials, compounds that can eliminate or inhibit microorganisms, is increasingly undermined by the problem of antimicrobial resistance (AMR) [1–3]. As one of the top ten public health threats identified by the World Health Organization [4,5], bacterial AMR is responsible for an alarming 1.27 million deaths worldwide in 2019 [6]. The conventional culture-based antimicrobial susceptibility testing (AST) currently guiding therapy [7] is time-consuming and limited to pure bacterial cultures [1,7–10]. Marker-based molecular diagnostics offer a faster alternative; but cover a few known loci, making them prone to false-negative results [9,11,12]. In contrast, whole genome sequencing, by providing access to all genes, offers an AMR phenotype determination approach potentially more accurate than current molecular methods [8,13,14], and in combination with metagenome analytics more rapid than culture-based testing for biomedical applications [15], which in future could complement or replace the standard AST for AMR phenotyping in clinical diagnostics, surveillance, and containment [1,8,13,14].

Two types of computational approaches have been developed for predicting AMR phenotypes from genomic data: rule-based AMR catalog mapping (referred to as rule-based) and machine learning (ML) approaches. Rule-based approaches determine AMR phenotypes using catalogs of known single nucleotide polymorphisms (SNPs), indels, and/or genes associated with AMR, comparing the sample to reference catalogs [10,16–32]. Although this approach can indicate the underlying AMR mechanisms by pinpointing SNPs and genes, it is inherently constrained by the scope, quality, and interpretation of the catalog/database [33].

The accumulation of genomes with AMR metadata [34] provides a rich resource for developing ML-based AMR phenotyping approaches, offering a strategy that is potentially more versatile and accurate for AMR prediction. These ML approaches use various classification algorithms for identifying

AMR patterns from the data (**Table S1**). ML models utilizing feature representation of k-mers [35–40], pan-genomes [41–44], and genome-wide or common mutations [41,45–50], can predict AMR phenotype without an understanding of the molecular mechanism. Additionally, some ML-based approaches also encode AMR determinant information into features [43,46,48,49,51–63], and some research utilized ML techniques for AMR determinant detection [41,44,64–69].

A direct consequence of the diverse landscape of AMR prediction methodologies is the inherent difficulty in benchmarking and comparing their effectiveness. While newly developed methods often highlight their advancement compared to existing methods, the evaluations are not standardized, leading to the use of different datasets, different, yet limited combinations of species and antibiotics, and lacking comprehensive comparisons with the full spectrum of available methods. ML-based AMR phenotyping is assessed with different approaches (**Table S1**): most straightforwardly, performances are compared based on the same set of features across various classifiers [35,36,38,41,45,48,51,55,56,70]. Further, benchmarks compare using features derived from various types of information, such as different AMR reference databases [71], complete sets and subsets of features (SNPs, genes, or k-mers) [39,46,47,49,59–61,63,67,68,72], and combinations of features (pan-genomes, SNPs, population or non-genomic information) [41–43]. Other studies compare ML-based software to rule-based software [42,46,53,59–61,63], or with another ML-based software on only one species [50,52,57,62,73], or assess a larger group of methods only on *M. tuberculosis* [49,57,58,62]. Of these studies, all except two, use at most up to five pathogen species for benchmarking (**Supplemental Methods**). Still lacking is a more comprehensive comparison covering both a wide range of pathogens and the wide spectrum of ML-based AMR phenotyping software, to survey whether ML methods are generally more competitive than the conventional ones. Therefore, establishing standardized benchmarking procedures for a systematic comparison of ML-based methods on a wide range of species, remains a significant research gap in the field of computational AMR phenotype prediction from microbial genomes.

To establish a comprehensive benchmark for evaluating AMR phenotyping, we developed a rigorous workflow in terms of datasets, guidelines, and metrics. We curated 78 datasets from the PATRIC database, representing different combinations of species and antibiotics. We then evaluated AMR phenotyping software using one of three defined approaches (nested CV, iterative bound selection [36] with CV, or iterative evaluation) to accommodate different techniques, each in conjunction with three sample partitioning methods (random folds, phylogeny-aware folds, homology-aware folds). We evaluated ML-based methods in both single-species-antibiotic models and multi-antibiotic/species models. We measured the performance by a set of common metrics, prioritizing the macro F1-score for a comprehensive analysis, and precision and F1-score for the susceptible samples for a clinical-oriented analysis. Within this framework, we conducted a thorough evaluation of state-of-the-art

AMR phenotyping methods, — Aytan-Aktug [53], Seq2Geno2Pheno [41], PhenotypeSeeker [38], and Kover [36]— alongside the rule-based ResFinder 4.0 [19] and ML baseline [36]. Our study aims to guide researchers in selecting the most suitable software for specific pathogens and antimicrobials, and indicating strengths and weaknesses of current techniques in the field.

Results

Development of a rigorous transparent benchmarking workflow for AMR prediction

To perform an unbiased and thorough comparison of AMR prediction software, we developed a benchmarking workflow using three distinct sample splitting methods and providing common performance metrics (**Fig 1A**). Within this workflow, the methods under evaluation are trained and tested on the predefined splits. This comprehensive approach, in terms of software, antimicrobial, and species coverage, combined with our rigorous workflow, provides a balanced assessment, thoroughly examining the generalization capacity across a spectrum of scenarios. To promote transparency and enable replication of our results, we have made the workflow available under an open-source license (http://github.com/hzi-bifo/AMR_benchmarking).

ML-based AMR phenotyping has inherent challenges, notably the ability of models to generalize to new input data. To assess the AMR phenotyping software in various scenarios, we employed three distinct sample splitting methods (random partitioning, phylogeny-based and homology-based approach) to reflect different degrees of evolutionary divergence or dissimilarity between taxa in the training and test data and to balance realistic evaluation and statistical reliability (**Supplemental Fig. S2**, **Materials and Methods, Supplemental Methods**). To ensure reproducibility, and maintain consistency and uniformity across evaluations, each splitting method was applied uniformly to every dataset, resulting in the same three dataset partitions for assessing the evaluated methods. In addition to the three sample partitioning methods for the single-species model (**Fig 1A**), we applied a challenging species-based sample partitioning method for evaluating cross-species models, using Leave-one-species-out (LOSO) evaluation (**Fig 1B, Materials and Methods**). The model hyperparameter optimizations and performance evaluations were implemented using nested CV designed to mitigate the bias in error estimation from CV [74], as well as two iterative evaluations, in order to accommodate different techniques (**Conclusions, Materials and Methods**).

We measured method performance using a set of common metrics: precision, quantifying the ratio of correct predictions to all predictions for a given class; recall, measuring the proportion of correctly assigned instances of a particular class; and F1-score, providing a balanced measure across these metrics for each class. Additionally, we assessed the overall assignment accuracy across the entire

dataset and the macro F1-score, offering a performance estimate across both resistance and susceptible classes, independent of class sizes (**Fig 1A,B**). Considering clinical implications and comprehensiveness (**Supplemental Methods**), we prioritized the macro F1-score for comparison; for overall comparison, we used the mean of macro F1-scores (referred to as F1-macro mean), averaged across the 10 folds (**Materials and Methods**). Turning to the clinical relevance of our study, we placed emphasis on metrics that are vital for antimicrobial susceptibility prediction in the clinical context. Specifically, the clinical-oriented precision and the F1-score for the susceptible (negative) class were used, recognizing their critical role in identifying actionable antibiotic susceptibilities (**Materials and Methods**).

Assessing software for comprehensive AMR phenotype prediction

To select state-of-the-art ML-based software for our study we followed a systematic approach: initially, we limited our search to software released since 2017 and with available source codes, ensuring relevance and accessibility for reproducibility, resulting in 17 methods (**Supplemental Table S1**). We then further considered the universal applicability of the software, excluding those that were species-specific (mostly only on one species, with one study on three species), requiring a reference strain (usually solely SNP-based software [45–48,50,58,60,61,63]) or an AMR gene catalog [46,57,58,60,61,63] for feature generation, resulting in seven methods [36,38,41,42,52,53,72].

We also aimed to cover a wide spectrum of ML classifiers. The selected methods include both common ML classifiers like support vector machine, logistic regression, and random forest (used by Seq2Geno2Pheno [41] and PhenotypeSeeker [38]) and neural networks (used by Aytan-Aktug [53]). Additionally, we include more novel approaches such as bound selection rule-based learning models by Kover [36]. The variety in the selected classifiers provides a broad and representative overview of the different algorithmic approaches available in the field, showcasing both established techniques and innovative methodologies.

Crucially, we furthermore selected methods such that they also provided a comprehensive comparison of ML methods based on various features that can be derived from the genomic dataset that we used for benchmarking. These features can generally be categorized into k-mers, pan-genomes (including gene presence and absence, sequence similarity clusters, SNPs, indels), AMR catalog-encoded features (presence and absence of AMR-associated SNPs and genes), SNPs, and non-genomic features (such as gene expressions, and year of isolation; **Supplemental Table S1**). Our selection thus included ML-based methods Aytan-Aktug (AMR catalog based) [53], Seq2Geno2Pheno (pan-genomes and k-mers counting) [41], PhenotypeSeeker [38] and Kover [36] (both using k-mers presence and absence). Additionally, we incorporated the AMR catalog mapping method ResFinder 4.0 [19],

recognized for its use of comprehensive and regularly updated databases including both resistance genes and SNPs, as well as a methodological ML baseline, predicting the most frequently occurring phenotype in the training data for each experiment [36]. Taken together, this selection strategy reflects the varied ways in which genomic information can be utilized for AMR prediction. To evaluate software within our rigorous workflow, we made some adaptations or refinements of ResFinder, Aytan-Aktug, Seq2Geno2Pheno, and PhenotypeSeeker (**Supplemental Methods**, https://github.com/hzi-bifo/AMR_benchmarking_khu).

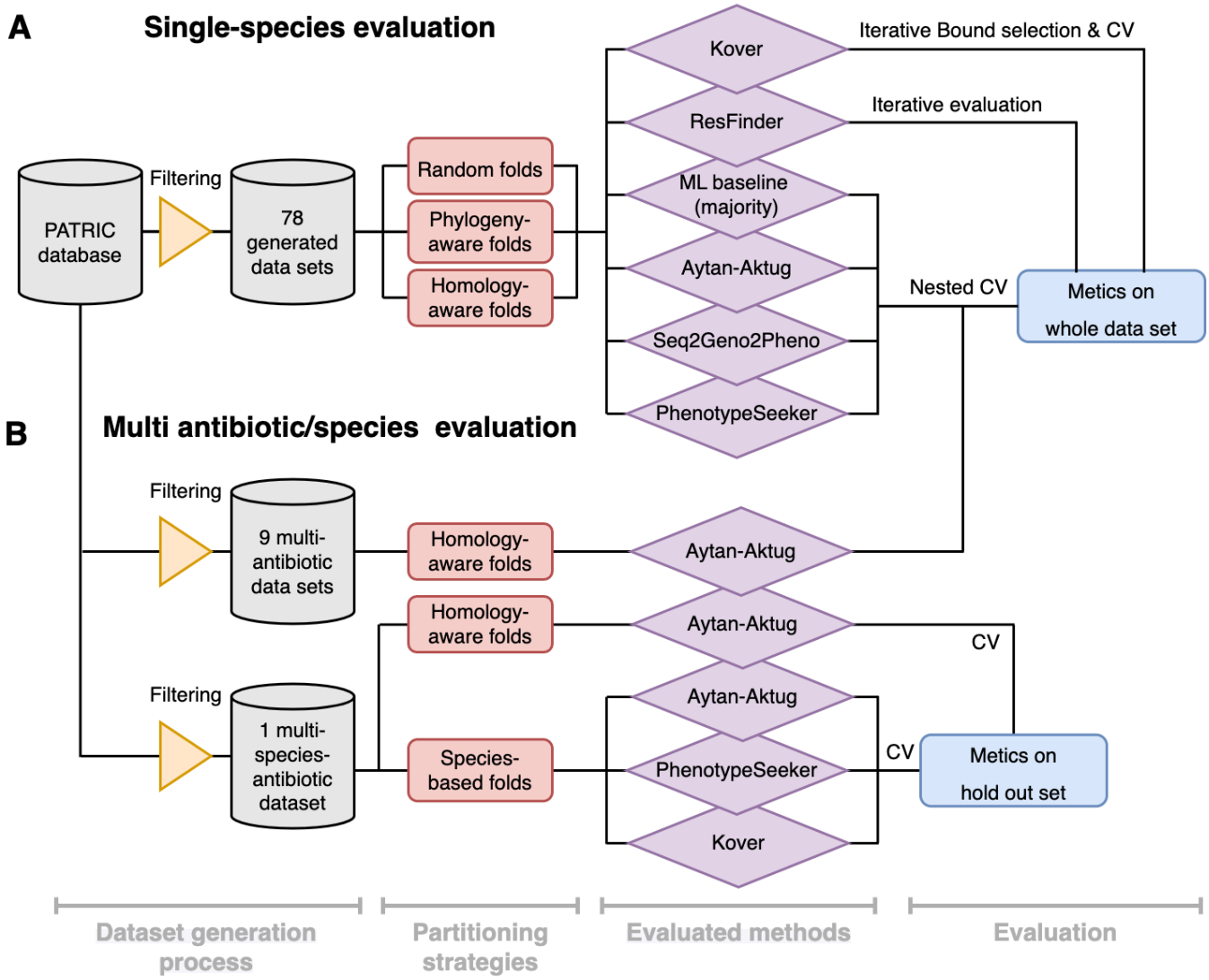

Figure 1. Overview of the evaluation workflow for antimicrobial resistance (AMR) prediction, spanning single-species to multi-species models and varied data processing strategies. A) The single-species model evaluation process, employing 78 datasets created from the PATRIC database. Software for AMR phenotype determination is evaluated through one of three approaches (nested CV, iterative bound selection [36] with CV, or iterative evaluation), each in conjunction with three sample partitioning methods (random folds, phylogeny-aware folds, homology-aware folds). The evaluation outcomes are expressed as mean and standard deviation for various model-oriented metrics (F1-macro, F1-negative, F1-positive, accuracy), and clinical relevant metrics (F1-negative, precision-negative, and recall-negative), benchmarked against in silico inferred phenotype and culture-based antimicrobial susceptibility testing standards. B) Multi-antibiotic or multi-species model evaluation illustrates the evaluation workflow for the Aytan-Aktug single-species multi-antibiotic model, two types of Aytan-Aktug multi-species-antibiotic models, and cross-species models using Aytan-Aktug, Kover and PhenotypeSeeker techniques. Aytan-Aktug single-species multi-antibiotic model was evaluated across 9 multi-antibiotic datasets using nested CV with homology-aware partitioning. Two types of Aytan-Aktug multi-species-antibiotic models are evaluated on a multi-species-antibiotic dataset using conventional CV with a hold-out test set on samples partitioned by sequence similarity, using conventional CV with a hold-out test set obtained by species-based stratification. Aytan-Aktug (multi-antibiotic), Kover (single-antibiotic) and PhenotypeSeeker (single-antibiotic) cross-species models are evaluated by leave-one-species-out on a multi-species-antibiotic dataset. Metrics are reported for each combination in the hold-out set in case of conventional CV.

Creation of comprehensive benchmark datasets for AMR phenotype prediction techniques

We curated a collection of 78 datasets from the PATRIC AMR database [75,76], each including microbial genomes for a distinct combination of species and antibiotics (**Fig. 2A, Supplemental Fig. S1**). These genomes were selected after a filtering process based on specific criteria, ensuring quality and relevance. Specifically, we considered the AST phenotype annotation, sequence type (excluding plasmid-only sequences), contig number, genome length, and PATRIC scores for genome quality, fine consistency, coarse consistency and completeness (**Materials and Methods**). This filtering process was crucial for excluding any inconsistencies, inaccuracies, or irrelevant information that could skew the assessment or reduce the reliability of the evaluation.

The 78 datasets, ranging from 200 and 13,500 genomes each and encompassing 31,195 genomes in total, provide coverage of at least 100 genomes for each evaluated phenotype class (resistant or susceptible to a certain antibiotic). The datasets include genomes from eleven bacterial species, including key pathogens and species with high clinical relevance. This includes eight high priority pathogens linked to the most AMR-related deaths [6], namely *Escherichia coli*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Streptococcus pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Mycobacterium tuberculosis*, and *Enterococcus faecium*, as well as *Salmonella enterica* which is among the eleven most concerning pathogens. We further incorporated genomes of two additional species of clinical and biological relevance, *Campylobacter jejuni*, the most common bacterial cause of gastroenteritis with the rise of AMR [77,78]; *Neisseria gonorrhoeae*, posing a significant clinical challenge with the emergence of pan-resistant isolates [79]. For each species, we compiled genomes along with phenotypic annotations indicating susceptibility or resistance to a given

antimicrobial. The resulting 44 antimicrobials span 14 major drug classes like beta-lactams, fluoroquinolones, aminoglycosides, tetracyclines, folate pathway inhibitors, lipopeptides, oxazolidinones, tuberculosis-specific drugs and combination therapies (**Supplemental Table S2**). The coverage of these combinations aims to reflect the complexity of antimicrobial resistance across diverse species associated with a variety of infectious diseases and potential resistance mechanisms [79,80].

We further created more complex datasets, to extend the evaluation scope from single-species single-label model to multi-label and multi-species AMR phenotyping models (**Supplemental Methods**). First, we generated nine multi-antibiotic datasets, each containing genomes from one of nine species that were annotated with resistance phenotypes from all the antibiotics compiled in that species' single-species-antibiotic datasets (**Fig. 2B,D**). For example, the *E. coli* multi-AMR phenotype dataset encompasses 2,493 genomes annotated with AMR phenotypes for up to 13 antimicrobials. Second, we created a multi-species-antimicrobial dataset covering nine key species (such as *S. aureus*, *K. pneumoniae*, *P. aeruginosa*, and *M. tuberculosis*) annotated with resistance against 20 antimicrobials, encompassing 20,989 genomes (**Fig. 2C, Supplemental File 1**). This covers antimicrobials that occur in at least two of the 78 species-antimicrobial combinations. In total, the multi-species dataset covers 54 species-antimicrobial combinations.

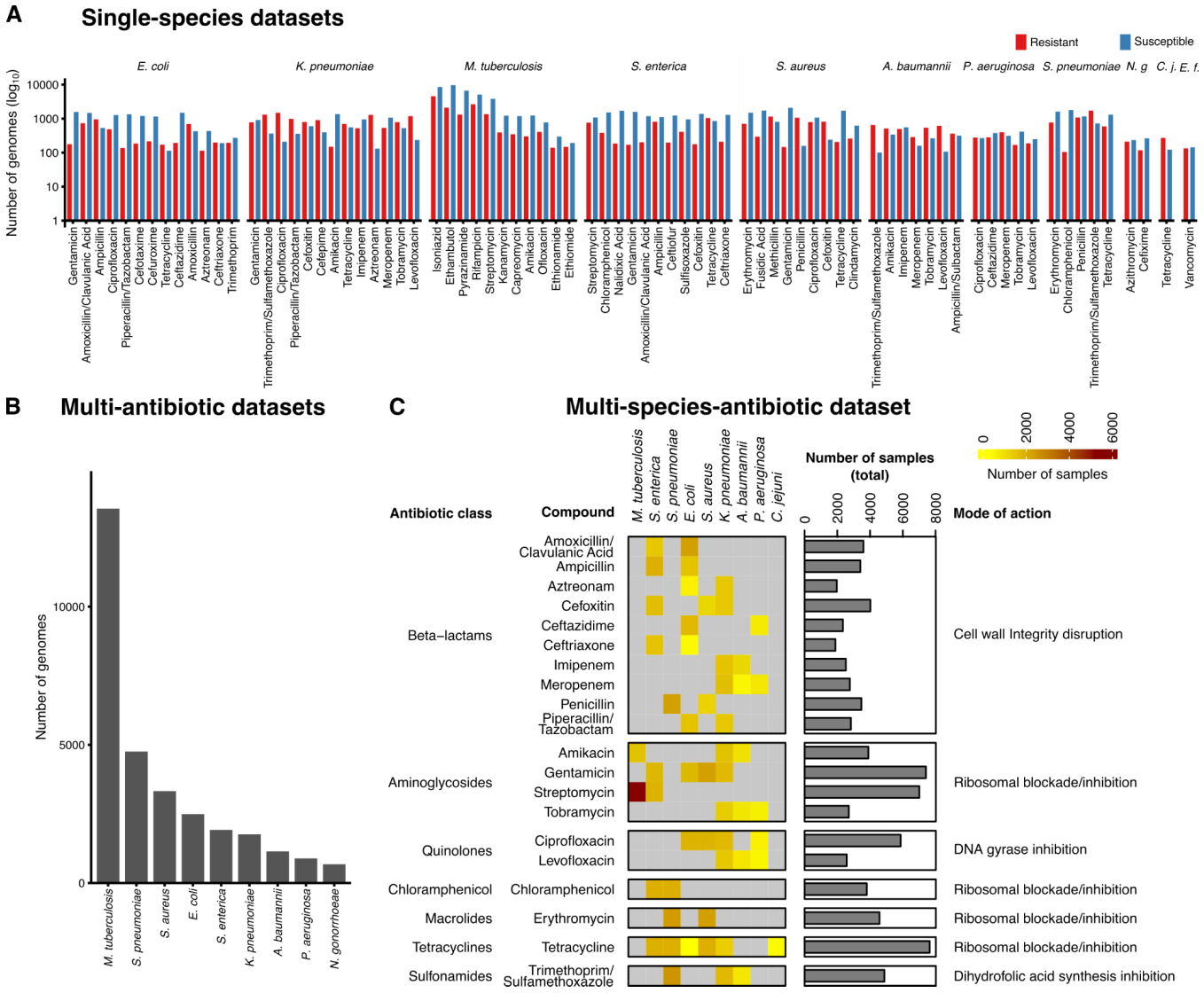

Figure 2: Overview of generated datasets for evaluating AMR phenotype predictions. A) 78 datasets for the single-species evaluation (**Fig. 1A**), spanning eleven species, annotated with resistant (red) and susceptible (blue) genome isolates. Abbreviation: *N. g.*: *N. gonorrhoeae*; *C. j.*: *C. jejuni; E. f.*: *E. faecalis*; B) Number of genomes included in the multi-antibiotic datasets for multi-antibiotic model evaluation (**Fig. 1B**); C) Overview of the multi-species-multi-antibiotic dataset, spanning 9 species and 20 antibiotics for multi-species model evaluation (**Fig. 1B**).

Comparative assessment of AMR phenotyping software: performance, robustness, and applicability

We critically evaluated the selected AMR phenotyping methods by assessing software performances across a variety of conditions, exploring performance quality and robustness, to provide insights into the practical applicability of these methods under different conditions. Software prediction quality and uncertainty were measured by F1-macro mean and F1-macro standard deviation, respectively, across the ten evaluation folds. Software robustness refers to the performance stability when the trained

software is evaluated on test sets with genomes of varied phylogeny or sequence similarity relationships to the training set. The smaller the performance change across different evaluation scenarios, the higher the software robustness and the more the trained ML-based software can be trusted by users to provide stable and accurate predictions on samples with unknown relationships to the training data.

We evaluated the five methods (four ML-based and the rule-based method ResFinder) on 78 datasets with random and homology-aware dataset splitting, as well as phylogeny-aware splitting on 67 datasets (11 datasets splitting failed due to runtime issues). For each method, hyperparameter selection range configurations are specified in the **Supplemental Methods**. Seq2Geno2Pheno failed to generate pan-genome-based features for *M. tuberculosis*, and PhenotypeSeeker could not complete the CV evaluation for four *M. tuberculosis*-antibiotic combination datasets, due to computational time exceeding two months with a maximum of 20 available CPUs. For these experiments, metric values were assigned as 0. Additionally, 13 of the 78 species-antibiotic combinations could not be assessed by ResFinder, owing to absent AMR catalogs, with metric values also consequently set to 0. In total, based on the 78 species-antibiotic combinations (referred to as cases), we conducted 390 (78 datasets x 4 ML methods+ 78 ResFinder) experiments under random folds and homology-aware folds, respectively, and 335 (67 datasets x 4 ML methods+67 ResFinder) experiments under phylogeny-aware folds (see **Supplemental Table S3** for metric scores). To assess method performances, we first compared these methods to the methodological ML baseline for each experiment [36]. With the exception of 6 experiments and 30 failed experiments due to excessive training times, ML methods outperformed (better or tied) the ML baseline in 96% of the experiments. Notably, the performance of ResFinder was comparably worse, primarily due to the absence of AMR catalogs for specific datasets in 37 of 43 underperforming experiments, surpassing the ML baseline only in 78% of the experiments (**Supplemental Table S4A**).

To further compare the performance of the various methods, we ranked them based on the frequency each performed the best (see **Fig. 3** for heatmap by F1-macro metric**, Table 1**). A method performed the best by having the highest F1-macro mean, or the lowest standard deviation in case of ties of the highest mean. Methods were assigned an equal fraction in a tie of mean and standard deviation for a particular species-antibiotic combination. While ML methods' prediction capacity varied, they maintained a generally declining trend across evaluation settings: these methods were performed best in random split evaluations, performed less well in phylogeny-based evaluations with strains in the test set being more divergent from the training data, and the least well in homology-based evaluation – with Kover being the notable exception. Across all settings, Kover most frequently emerged as the top performer, followed by PhenotypeSeeker, Seq2Geno2Pheno and the Aytan-Aktug models (**Table 1**). In

the context of random split evaluations, Kover significantly outperformed other methods (p-value of 0.01 and 3.27×10^{-7}, when compared to PhenotypeSeeker and ResFinder, respectively, using one-sided paired t-tests), attaining the highest F1-macro values for 30% of the species-antibiotic combinations. ResFinder and PhenotypeSeeker followed, with 25% and 20%, respectively (**Table 1**). On the other hand, ResFinder demonstrated a particular strength in predicting AMR phenotypes for evolutionarily distant strains from the training samples. During phylogeny- and homology-based evaluation, ResFinder performed best on 44% and 50% of the datasets, respectively, followed by Kover, with 28% and 34%, respectively (**Table 1**). Despite these findings, the lack of AMR catalogs and the consequent inability to offer predictions for 17% species-drug combinations meant that ResFinder did not prove to be significantly better than Kover overall (**Fig. 3**, **Fig. 4A**). In addition, we evaluated the reproducibility of ML methods, showing that all ML-based software exhibited some variability in their results due to stochastic factors, except for PhenotypeSeeker (**Supplemental Results**).

Table 1: Percentage of species-antibiotic combinations where each method performed best, stratified by the dataset splitting methods. Each dataset covers a specific species-antimicrobial combination. A method is considered to perform best for a species-antimicrobial combination by having the highest F1-macro mean, or the lowest standard deviation in case of ties of the highest mean (see **Supplemental Table S6** for the best-performing method for each species-antimicrobial combination). Methods were assigned a fraction of the count for a mean and standard deviation tie. The fraction in brackets stands for the number of corresponding winning combinations over the total. Bold indicates the first-ranking method.

AMR prediction method	Fraction of experiments where AMR prediction method is performing best (by sample splitting method)		
	Random folds	Phylogeny-aware folds	Homology-aware folds
ResFinder	25% (19.53/78)	**44% (29.25/67)**	**50% (38.7578)**
Aytan-Aktug	8% (6.12/78)	6% (4.08/67)	2% (1.58/78)
Seq2Geno2Pheno	17% (13.28/78)	10% (6.58/67)	4% (3.25/78)
PhenotypeSeeker	20% (15.37/78)	13% (8.58/67)	10% (7.92/78)
Kover	**30% (23.70/78)**	28% (18.50/67)	34% (26.50/78)

Collectively, ML methods demonstrated very good performance, achieving an F1-macro mean ≥ 0.9 in 64% of the experiments and ≥ 0.8 in 81% of the experiments during random evaluation (**Table 2**). They were the best-performing methods in 64% of the species-antibiotic combinations, outpacing the AMR-catalog mapping technique of ResFinder, which was best in 17% of the cases. In 19% of the cases, both approaches were tied (**Supplemental Table S4C**). However, the effectiveness of ML methods dropped noticeably when applied to strains that were more evolutionarily divergent from the model training genomes. This decrease was pronounced in phylogeny-based and even more so in homology-based evaluations. In these settings, ML methods only achieved an F1-macro mean ≥ 0.9 in

33% and 25% of experiments, respectively, and an F1-macro mean ≥ 0.8 in 60% and 50%, respectively. Also, when compared to ResFinder, ML methods attained higher or equal F1-macro scores in fewer experiments (47% or 40%, respectively), with actually higher scores being observed in 40% and 35% of the experiments(**Table 2, Supplemental Table S4B**).

When considering the application to clinical diagnostics and treatment, prediction of a pathogen's susceptibility to a specific drug is of high importance [81]. Thus, high precision in detecting susceptibilities is essential [8]; meanwhile, high recall in detecting susceptibilities is required for identifying viable treatment options. To this end, we evaluated software performances with an emphasis on the F1-score and precision for the negative, i.e., susceptible class (**Material and Methods, Supplemental Table S3**). With ML methods, a precision of at least 0.95 for the negative class was reached for 47% (147/312) of experiments with random split evaluation. However, this decreased to 39% (105/268) and 30% (94/312), under phylogeny-based and homology-based evaluation, respectively. Similarly, an F1-score for the negative class ≥ 0.95 was achieved in 49% (154/312) of random split evaluation experiments, compared to 41% (111/268) and 29% (92/312) in phylogeny-aware and homology-aware evaluation, respectively (**Table 2**). Thus, depending on the evolutionary divergence of strains relative to the training data, specific antibiotic susceptibilities of these pathogens can be predicted using ML methods with high confidence in almost one-third to half of the cases. Similar to evaluations measured by F1-macro mean, by the metrics for the negative class ML methods performed better in random split evaluations, while in the phylogeny-based and homology-based evaluation settings, AMR catalog mapping with ResFinder performed comparably well (**Table 2, Supplemental Results**).

Table 2: Percentage of machine learning (ML) method experiments with the performance above the indicated threshold, stratified by the dataset splitting methods. The fraction in brackets stands for the number of corresponding experiments over the total. In total, there are 312 (78 datasets x 4 ML methods) experiments under random folds and homology-aware folds, respectively, and 268 (67 datasets x 4 ML methods) experiments under phylogeny-aware folds.

Threshold	Fraction of experiments satisfying the indicated threshold (by sample splitting method)		
	Random folds	Phylogeny-aware folds	Homology-aware folds
ML methods F1-macro ≥ 0.8	81% (252/312)	60% (161/268)	50% (155/312)
ML methods F1-macro ≥ 0.9	64% (201/312)	33% (89/268)	25% (79/312)

ML methods F1-macro ≥ ResFinder F1-macro	66% (206/312)	47% (127/268)	40%(126/312)
ML methods precision-negative ≥ 0.95	47% (147/312)	39% (105/268)	30% (94/312)
ML methods F1-negative ≥ 0.95	49% (154 /312)	41% (111/268)	29% (92/312)
ML methods precision-negative ≥ ResFinder precision-negative	62% (194/312)	49% (131/268)	42% (132/312)
ML methods F1-negative ≥ ResFinder F1-negative	69% (214/312)	54% (145/268)	47% (147/312)

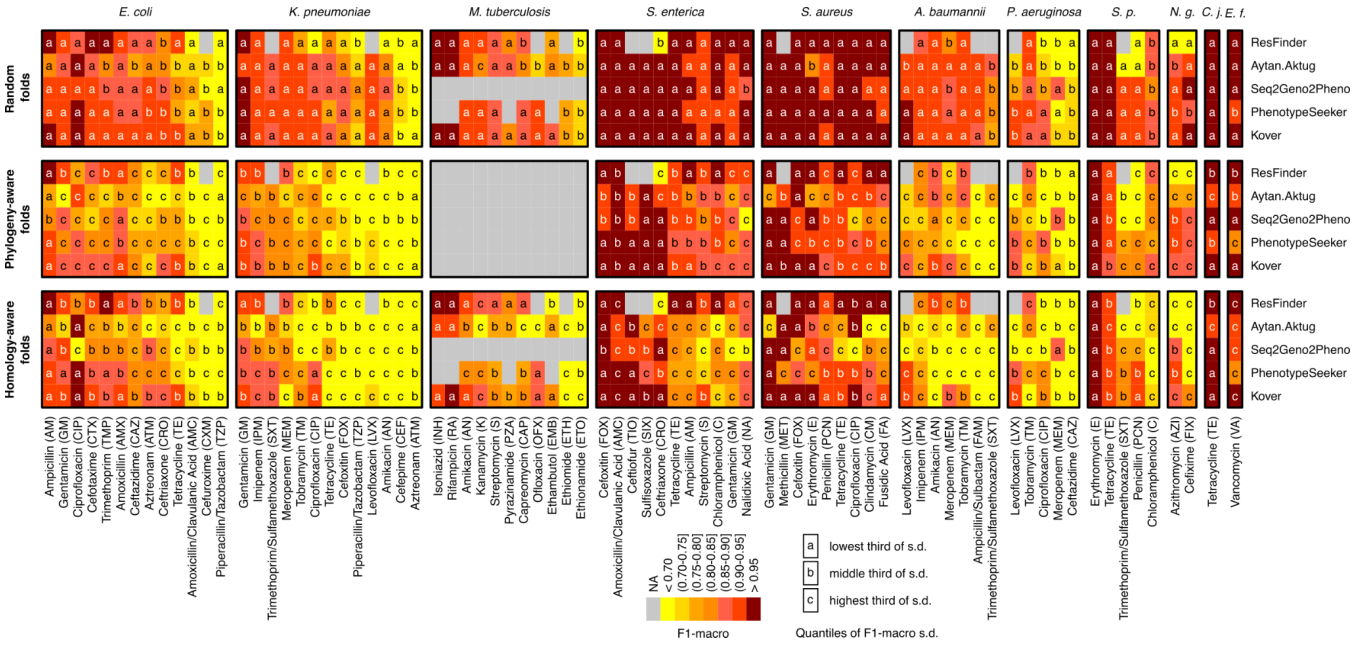

Figure 3. Heatmaps comparing AMR phenotyping performance (F1-macro) under random split folds, phylogeny-aware folds, and homology-aware folds for the evaluated methods on 78 species-antibiotic combinations. The grey cells indicate either the absence of an AMR catalog for ResFinder, or failure of ML-based software evaluation due to running time issues. Colors represent F1-macro mean, while markers 'a,' 'b,' and 'c' denote the F1-macro standard deviation categories of 'lowest', 'middle', and 'highest', respectively, based on their tertiles. Abbreviation: *S. p*: *S. pneumoniae*; *N. g.*: *N. gonorrhoeae*; *C. j.*: *C. jejuni; E. f.*: *E. faecalis*.

Species- and antibiotic-specific trends in AMR prediction performance

For two pathogen species, *C. jejuni* and *E. faecium*, the predictive quality remained consistently high across the three evaluation types. This consistency highlights the capacity of all methods to robustly predict antimicrobial resistance phenotypes, even for strains that are more evolutionarily divergent (**Fig.**

2, Fig. 4B for F1-macro mean; **Supplemental Fig. S10B** for F1-macro standard deviation). In contrast, for the other nine evaluated species, model performances experienced a decline in the more challenging scenarios involving test data from phylogenetically divergent strains relative to the training data. While *S. aureus*, *S. enterica*, *A. baumannii*, *N. gonorrhoeae* and *S. pneumoniae* showed strong performance with ML methods under random evaluation, there was a marked deterioration in the median of F1-macro mean and increased variation in both F1-macro mean and standard deviations in the more challenging settings. For species like *E. coli*, *M. tuberculosis*, *K. pneumonia* and particularly for *P. aeruginosa*, ML models performed not quite as well (median of F1-macro means between 0.89 and 0.79) already with random evaluation, and performances further declined in the other settings. For all species, variance increased for the more challenging settings, indicating the increased uncertainty of trained models. For all species, with an exception for *M. tuberculosis* where the phylogenetic tree could not be generated due to runtime issues, there is an uneven distribution of the misclassification frequencies across the corresponding phylogenetic tree (**Supplemental Results, Supplemental Fig. S13-22**). This indicated that AMR phenotyping across a range of drugs might be complicated by lineage backgrounds.

Our results revealed substantial variations in predictive qualities for different antibiotic classes (**Supplemental Fig. S8**) and across different antibiotics, excluding experiments with computing failures due to runtime issues (**Supplemental Fig. S9**). For the antibiotic classes, macrolides, sulfonamides, rifamycins, glycopeptides and lipoglycopeptides tended to be more accurately predicted than other antibiotic classes across the three evaluation scenarios. For aminoglycosides, quinolones, chloramphenicol, tetracyclines, lincosamides, fusidane, TB drugs, and dihydrofolate reductase inhibitors, AMR prediction methods generally achieved high accuracy in random evaluations, but experienced a decline in the two more challenging scenarios. Intriguingly, performance on beta-lactams class exceptionally exhibited a high variation of F1-macro means across both random splits and the other two more challenging evaluation scenarios. Further, our analysis demonstrated variations among different species treated with the same antibiotic (**Supplemental Fig. S5-7**). This variation was particularly evident in the case of five Beta-lactams — aztreonam, amoxicillin/clavulanic acid, cefoxitin, ceftazidime, and piperacillin/tazobactam — and tetracycline under random folds. These variations, except for piperacillin/tazobactam, still exist under phylogeny-aware and homology-aware folds. Moreover, the variation between individual species-antibiotic combinations was pronounced, with cefoxitin for *S. aureus* and aztreonam for *K. pneumoniae* being predicted with the highest (F1-macro ≥ 0.99) and lowest accuracies (F1-macro average 0.59).

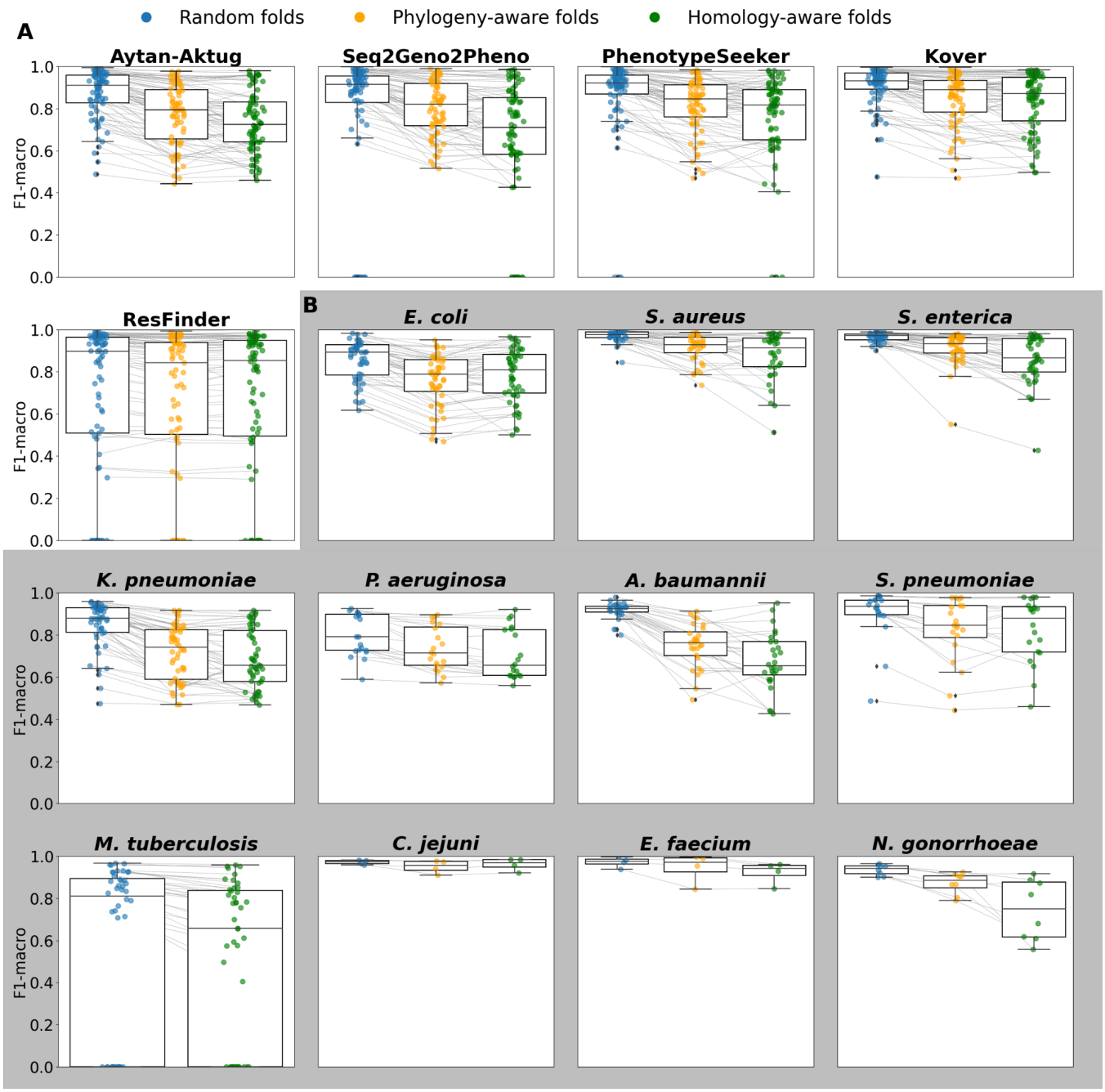

Figure 4: Performance (F1-macro mean) changes under random folds, phylogeny-aware folds, and homology-aware folds. A) Performance changes for each software. Each dot represents one combination of species and antibiotics for the corresponding software. B) Performance changes for each species by ML-based software Kover, PhenotypeSeeker, Seq2Geno2Pheno, and Aytan-Aktug. Each dot represents one combination of ML-based software and antibiotic for the corresponding species.

Assessment of multi-antibiotic and multi-species AMR prediction models

A substantial number of sequenced genomes with AMR metadata, which are vital for training ML-based AMR phenotype prediction models, are not available for all species and antibiotic combinations. Similarly, high-quality AMR catalogs, crucial for developing ML-based AMR phenotyping software relying on AMR catalogs, are not available for all species and antibiotic combinations. Furthermore, multidrug-resistance, seen in strains like methicillin-resistant *S. aureus* and multidrug-resistant *M. tuberculosis*, often arises from interacting mechanisms — primarily intrinsic via multidrug-resistance efflux systems [80], and acquired via mutations or multidrug resistance mobile genetic material [79,80,82–84]; AMR genes coding for multidrug efflux pumps, are highly prevalent in all studied organisms [84]. It is of particular interest therefore to create models generalizing to multiple antibiotics and multiple species, thus potentially improving AMR phenotype predictions for less-studied taxa and taxa with limited data. Therefore, in addition to single-species-antibiotic models (**Fig. 1A**), we also evaluated multi-antibiotic models, that fall under multi-label classification, as well as multi-species models trained on genomes from multiple species (**Fig. 1B**). For the multi-species models, we examined their performance under two conditions, first when applied to a species already included in the training data, and second when applied to a species the model had never encountered before.

To assess the multi-antibiotic model, we retrained and tested the Aytan-Aktug multi-label model on nine species with data available for multiple antibiotics, respectively, using homology-based data stratification and nested CV (**Supplemental Methods**). Similar to Aytan-Aktug, et al. [53], we observe no substantial difference in the predictive performances of multi-antibiotic and single-antibiotic models, by evaluation on 9 species instead of only one in the native study (p-value 0.48 with paired t-test, **Supplemental Fig. S23**).

To access the multi-species model where multiple species were involved in both the training and testing procedures, Aytan-Aktug techniques were employed and evaluated on a multi-species-antibiotic dataset, covering 54 combinations of one of 9 species and one of 20 antibiotics with homology-based data stratification (**Supplemental Methods**, **Supplemental Fig. S25**). We did not find a significant difference between the homology-aware CV evaluation of Aytan-Aktug single-species models and a control multi-species model utilizing species-specific mutation-detection reference databases for training the resistance models (paired t-test; p-value 0.47). However, the second type Aytan-Aktug cross-species model that utilized a combined mutation-detection reference database performed less well than the Aytan-Aktug single-species models (paired t-test; p-value 1.15×10^{-4}).

To assess the multi-species model on a species not included in the training set, we trained and tested cross-species models with the Aytan-Aktug technique (multi-antibiotic model), Kover (single-antibiotic models), and PhenotypeSeeker (single-antibiotic models), using LOSO evaluation, on the same

above-mentioned multi-species-antibiotic dataset (**Supplemental Methods**). For all the LOSO cross-species models, we found each performed significantly less well than the corresponding method's single-species models (**Fig. 5, Supplemental Fig. S24**), with the paired t-test p-value of 1.38×10^{-13}, 7.17×10^{-19}, 3.42×10^{-12}, 4.01×10^{-18} for Aytan-Aktug, Kover set covering machines (SCM), Kover classification and regression trees (CART), and PhenotypeSeeker, respectively. This agrees with a previous study[53], which tested this approach using only Aytan-Aktug technique for only one species.

Compared to all the single models evaluated under homology-aware folds (**Fig. 1A**), at least a multi-model (**Fig. 1B**) performed best or tied best for 13 species-antibiotic combinations by the F1-macro metric (**Supplemental Table S8A,** evaluation results of multi-models for all possible species-antibiotic combinations see **Table S8B-J**). Notably, the control multi-species Aytan-Aktug achieved significantly better performance on *A. baumannii*-trimethoprim/sulfamethoxazole, outperforming the second-best method (Aytan-Aktug single-species-antibiotic model) by F1-macro of 0.18.

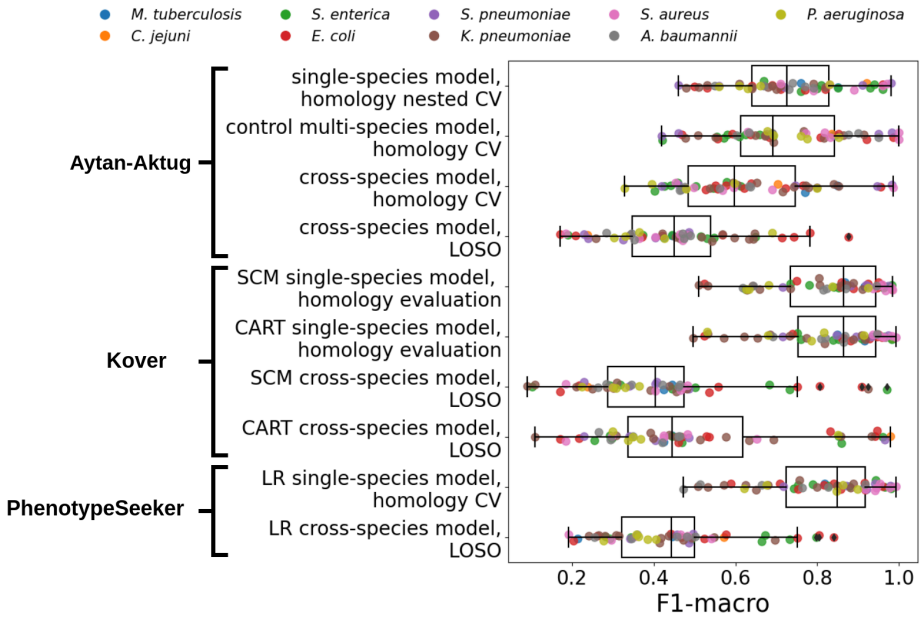

Figure 5: Comparison of the performance (F1-macro) of multi-species models and single-species models. Performance (F1-macro) of LOSO evaluation of cross-species models using Aytan-Aktug, Kover set covering machines (SCM), Kover classification and regression trees (CART), and PhenotypeSeeker logistic regression (LR), respectively, as well as homology-aware evaluation of control multi-species model (Aytan-Aktug) are compared to the performance (F1-macro mean) of corresponding single-species models (Aytan-Aktug, Kover, and PhenotypeSeeker). The evaluations were conducted on the multi-species-antibiotic dataset, covering 9 species (indicated by the color).

A guide for selecting software for specific species-antibiotic combinations

To help users in selecting the most suitable software for their analysis, we compiled a table of recommendations derived from our evaluation across the different scenarios (**Fig. 1A**), where we selected the software with the highest F1-macro mean, resorting to the lowest standard deviation for a tie, across evaluation folds for each species-antibiotic combination (**Supplemental Table S6**). This table reflected the upper limits for AMR predictive accuracies, the performance could be achieved by selectively employing the best-performing software for each species-antibiotic combination (outer "envelopes" of F1-macro mean circles in **Supplemental Fig. S4**). In random evaluation, this approach allowed the prediction of AMR phenotypes with an F1-macro ≥ 0.9 for 79% of the species-antibiotic combinations. Predictions for evolutionarily divergent strains declined to 60% in phylogenetically-aware evaluations and further to 55% in homology-aware evaluations. In 49% of the species-antibiotic combinations, software recommendations contain at least one consistent method across different evaluation settings, indicative of robust behavior of the corresponding models in case of varying evolutionary divergences between training and test genome data. Considering evaluation results based on the “multi-” models (**Fig. 1B**), the F1-macro of the Aytan-Aktug control multi-species model was substantially higher by 0.18 compared to the second-best method for the combination of *A. baumannii* and trimethoprim/sulfamethoxazole. This led to the model's inclusion in the selection guideline (**Supplemental Table S6**) for that antibiotic-species pairing. It should be noted that different data splits and evaluation approaches were employed due to the “multi-” configuration of the models (as indicated in **Fig. 1A, B**).

To facilitate the use of the recommended software, we provided an integrated automated pipeline for AMR phenotype prediction based on the top three models Kover, PhenotypeSeeker and ResFinder (**Table 1**) after training ML models on each of the 78 datasets (**Fig. 2A,** https://github.com/hzi-bifo/AMR_prediction_pipeline). The models were trained using the set of hyperparameters selected via CV under the three above-mentioned scenarios (**Supplemental Methods**). For a particular species-antibiotic combination, in case recommendations differ between scenarios, users can determine the degree of evolutionary divergence between training genomes and genomes they are analyzing. If the genome is evolutionarily close to these training genomes, the best choice would be to use the method that performed best in random evaluation, and otherwise, the method performing best in the other settings.

Conclusions

Addressing the lack of a systematic comparison of ML-based AMR prediction software, our study provides a broad and nuanced evaluation of five AMR prediction methods, including both ML-based techniques and the rule-based AMR catalog mapping approach. Our analysis revealed the capabilities of leading computational methods for AMR phenotyping across species and antibiotics in different scenarios, guiding software selection for specific species-antibiotic pairs. To date, no studies have benchmarked ML-based AMR software across such a broad spectrum.

When comparing ML-based and rule-based catalog mapping methods, a notable performance discrepancy was observed between the two. ML methods' superiority under random folds underscores the strength of ML techniques in identifying unknown or novel AMR determinants, a crucial capability given the lack of comprehensive AMR catalogs for numerous species-antibiotic combinations; the rule-based approach's superiority under more challenging scenarios highlights its advantage in generalizing to evolutionarily divergent genomes when AMR catalogs are available, which was the case for 83% of all evaluated species-antibiotic combinations. Performance discrepancies between the two approaches have also been reported by others [46,59–61,63].

This tradeoff is demonstrated by Kover, which underperformed in the more complex scenarios compared to the rule-based method but consistently outperformed the other evaluated ML-based methods on all evaluated scenarios. Kover's superiority may be attributable to procedures specially designed to address learning generalization issues by mitigating dimensionality problems in ML though compression-theory-based rule-based ML models. The Aytan-Aktug method might be limited by the AMR catalog used to extract features, as its developers have noted no significant improvement over the results obtained by ResFinder [19], and Seq2Geno2Pheno's performance might be weakened by the lack of core gene mutation information in its features, since it provides only a reference strain of *P. aeruginosa* for SNPs detection.

For individual species-antibiotic combinations, the performance of ML methods generally declined for more challenging evaluation scenarios, while ResFinder maintained consistent performance. This trend can potentially be explained by three hypotheses pertaining to the role of lineage backgrounds in ML models [47,59,66]. First is shortcut learning. As homologous isolates from a specific source could predominantly exhibit one AMR phenotype due to factors like environments [80], models learning AMR patterns may pick up spurious clade-specific signals from the dataset. This hypothesis proposes the incorporation of gene expression information for AMR phenotyping. The second hypothesis posits the independent stepwise evolution of strains leading to different AMR variants, exemplified by the emergence of multidrug-resistant *E. coli* ST131 [85,86] and carbapenem-resistant *K. pneumoniae*

ST258 [87] clones. This indicates the importance of comprehensive training data. While evolution can be both contingent and deterministic [88]; a ML-based investigation [89] aligns with the deterministic theory, and some methods [42,90] have achieved moderate AMR phenotyping accuracy using phylogenetic information, indicating potential efficacy of exclusively utilizing ancestral lineages for training ML models in the absence of comprehensive data. Lastly, the third hypothesis suggests that the combined effects of AMR variants and lineage backgrounds could impact the efficacy of ML-derived predictors, as indicated by the genome misclassification frequency. Our misclassification analysis offers a novel insight by uncovering clade-specific variations in AMR prediction accuracy, shedding some light over the complex interactions between an organism's phylogeny and AMR phenotype. Our approach is complementing existing genome-wide association [91] and ML [66] studies.

Among the 14 benchmarked antibiotic classes, predictions for macrolides and sulfonamides yielded the highest accuracies while performance for beta-lactams exhibited high variation, even under random folds. This suggests that it poses more difficulties for ML algorithms to capture beta-lactams' multifaceted resistance mechanisms, including altered penicillin-binding proteins, beta-lactamase production, and changes in porin channels [2]. While we evaluated the performance of both multi-antibiotic and multi-species models, we observed no significant difference in their effectiveness compared to single-species and single-antibiotic models.

Other than most studies in this domain, our workflow incorporates more sophisticated and less biased evaluation techniques and distinct sample partitioning methods. For assessing ML classifiers, studies commonly rely on repeated CV [45,61,63] or nested CV [41,52]. We employed nested CV for the ML methods, which offers a more rigorous and unbiased performance estimate. For the rule-based method ResFinder, where hyperparameter optimization is not needed, we evaluated performance on the same folds as for the other methods, referring to this as iterative evaluation. Kover employs an intrinsic hyperparameter optimization using bound selection. Similarly, we applied CV, calling this iterative bound selection; at each iteration, the classifier maximizing the "F1-macro" on the training set was selected of two classifiers produced by Kover for each dataset. While previous ML-based AMR studies mostly partitioned samples randomly for CV, a few explored phylogeny-aware and random folds [41], homology-aware folds [53], and country-based folds [48]. Uniquely, our research incorporates three distinct sample partitioning methods, aiming to evaluate the generalization power and robustness of ML-based software when confronted with taxa that differ evolutionary from training data.

Our clinical-focused analysis indicated that the tested software provides very accurate antimicrobial susceptibility predictions for many genomes with representative training data available, but requires further optimization for others, especially for evolutionarily divergent strains. Further improvements could arise from including (1) more genome data with AMR phenotype from the wet lab for more

comprehensive and phenotype-balanced training data; (2) further optimizing ML techniques to capture the underlying patterns associated with AMR phenotypes, for example, by using transfer learning, up-sampling and/or sub-sampling before training, and an ensemble of existing AMR bioinformatics methods.

Materials and Methods

Datasets

We compiled 78 datasets from the PATRIC genome database [75,76] (accessed in Dec 2020) through the following steps to ensure quality and relevance: (a) We downloaded genomes with laboratory-determined AMR phenotypes, resulting in 99 species, including 67,836 genomes. We kept species with more than 500 genomes, resulting in 13 species, including 64,738 genomes. (b) We filtered genomes based on PATRIC quality measuring attributes [92,93]: avoiding plasmid-only sequences, ensuring good genome quality, limiting contig count, and checking fine consistency, coarse consistency, completeness, and contamination criteria (see **Supplemental Methods** for thresholds). (c) We computed the average genome length for each species. Then, we kept only those genomes whose length deviated from the average by no more than 5% (one-twentieth of the mean). (d) We retained genomes with phenotype either resistant or susceptible, excluding genomes with intermediate phenotype or without phenotype. (e) We only retain species-antibiotic combinations with a minimum of 100 genomes for both resistant and susceptible phenotypes. (f) We included 402 *P. aeruginosa* genomes, and excluded the *S. pneumoniae*-beta-lactam and *M. tuberculosis*-rifampin combinations (**Supplemental Methods, https://github.com/hzi-bifo/AMR_benchmarking/blob/main/data/README.md**).

Sample partitioning

For phylogeny-aware folds, core gene alignments were created via Seq2Geno using Prokka [94] and a modified Roary version [41,95]; subsequently, joint neighborhood phylogenetic trees were derived from these alignments using the R packages phangorn and phytools; folds were then generated based on the clade structure of the phylogenetic tree for each species, with sample partition annotation through Geno2Pheno. Random folds were generated using Geno2Pheno, excluding *M. tuberculosis* folds, which were created using scikit-learn package “model_selection.KFold” module. Homology-aware folds were produced using Aytan-Aktug techniques [53], refined by us (**Supplemental Methods**). Sample partitions indicated by PATRIC genome IDs (random, phylogeny-aware, and homology-aware folds) and visualized on phylogenetic trees (random and phylogeny-aware folds) are available on Mendeley Data **(doi: 10.17632/6vc2msmsxj.1).**

Evaluation

For PhenotypeSeeker, Seq2Geno2Pheno, Aytan-Aktug, and ML baseline, the ten-fold nested CV was employed. During each iteration, a fold was held out as the test set, while the remaining nine folds were employed for training, with the hyperparameters and classifier chosen via a nine-fold conventional CV (**Supplemental Methods**). This procedure was iterated over the ten folds (**Supplemental Fig. S1A**). For Kover, the ten-fold Iterative bound selection was employed. During each iteration, a fold was held out as the test set, while the remaining nine folds were employed for training either Kover CART or Kover SCM; the classifier was selected via a nine-fold conventional CV; the hyperparameters were chosen via bound selection [36]. This procedure was iterated over the ten folds (**Supplemental Fig. S1B**). For the rule-based software, a ten-fold iterative evaluation was employed. In this process, each fold was assessed sequentially by ResFinder (**Supplemental Fig. S1C**).

Evaluation metrics

The F1-macro (**Eq. 7**) was computed using the Python scikit-learn package f1_score module. Precision-positive (**Eq. 1**), recall-positive (**Eq. 2**), F1-positive (**Eq. 3**), precision-negative (**Eq. 4**), recall-negative (**Eq. 5**), F1-negative (**Eq. 6**), and accuracy (**Eq. 8**) were computed using the Python scikit-learn package classification_report module. The metrics were defined as follows:

$$Recall_{positive} = \frac{TP}{TP+FN} \tag{1}$$

$$Precision_{positive} = \frac{TP}{TP+FP} \tag{2}$$

$$F1_{positive} = 2 \times \frac{Recall_{positive} \times Precision_{positive}}{Recall_{positive} + Precision_{positive}} = \frac{2TP}{2TP+FN+FP} \tag{3}$$

$$Recall_{negative} = \frac{TN}{TN+FP} \tag{4}$$

$$Precision_{negative} = \frac{TN}{TN+FN} \tag{5}$$

$$F1_{negative} = 2 \times \frac{Recall_{negative} \times Precision_{negative}}{Recall_{negative} + Precision_{negative}} = \frac{2TN}{2TN+FN+FP} \tag{6}$$

$$F1_{macro} = \frac{F1_{positive} + F1_{negative}}{2} \quad (7)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (8)$$

In the formula, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives, and TP is the number of true positives.

Statistical analysis

In iterative evaluation approaches, the mean and standard deviation of evaluation metrics were calculated across the 10 folds for each species-antibiotic combination and each method (**Fig. 1**). Sample-weighted mean and standard deviation were used for homology-aware folds, as the distribution of samples was uneven across folds. Arithmetic mean and normal standard deviation were used for random and phylogeny-aware folds. In the conventional CV for multi-species models, we computed the metrics for the hold-out test set (**Fig. 1B**). To avoid ill-defined scores resulting from the uneven distribution of negative samples across folds(**Supplemental Methods**), clinical-oriented F1-negative, precision-negative, and recall-negative were computed after pooling prediction results from 10 folds . All the metric scores were rounded to 2 decimal points for comparison. F1-macro mean, and clinical-oriented F1-negative and precision-negative were used for performance analysis. To test if the performance of two methods was significantly different, a paired t-test was applied on the F1-macro means of the two methods using Python SciPy library stats.ttest_rel module. To test if a method significantly outperformed the other method, a one-sided paired t-test was applied using the same above-mentioned Python module, with the “alternative” parameter set to “greater”.

Supplemental material

Supplemental Texts

- Supplemental Results
- Supplemental Methods

Supplemental Figures

- Figure S1: Dataset overview.

- Figure S2: Three cross-validation approaches in the benchmarking framework.
- Figure S3: The sample partitioning methods for cross-validation.
- Figure S4: Radar plot of the performance of AMR prediction methods.
- Figure S5-7: Performance (F1-macro) of software over antibiotics shared by multiple species under random folds, phylogeny-aware folds, and homology-aware folds.
- Figure S8: Performance (F1-macro mean) changes for each antibiotic class under random folds, phylogeny-aware folds, and homology-aware folds.
- Figure S9: Performance (F1-macro mean) changes for each antibiotic under random folds, phylogeny-aware folds, and homology-aware folds.
- Figure S10: Performance stability (standard deviation of macro F1-score) changes for (A) each software and (B) each species, under random folds, phylogeny-aware folds, and homology-aware folds.
- Figure S11-12: Performance and its stability (mean and standard deviation of negative F1-score) change for each species-antibiotic combination under random folds, phylogeny-aware folds, and homology-aware folds.
- Figure S13-22: *E. coli*, *S. enterica, S. aureus*, *K. pneumoniae*, *S. pneumoniae*, *A. baumannii*, *P. aeruginosa, C. jejuni, N. gonorrhoeae* and *E. faecium* genome misclassification distributions on corresponding phylogenetic trees, respectively.
- Figure S23: Performance comparison (F1-macro mean and standard deviation) between the single-species-antibiotic and single-species multi-antibiotic Aytan-Aktug model.
- Figure S24: Performance (F1-macro) of multi-species model using Aytan-Aktug, Kover and PhenotypeSeeker, in comparison to corresponding single-species model (F1-macro mean with standard deviation error bars).
- Figure S25: Sample partitioning method for Aytan-Aktug control multi-species model and cross-species model.
- Figure S26: The performance (F1-macro) comparison of control multi-species Aytan-Aktug model, cross-species Aytan-Aktug model, and single-species Aytan-Aktug models by Aytan-Aktug native methods.

Supplemental Tables

- **Table S1:** ML-based studies on AMR phenotype determination based on genome data.
- **Table S2:** The antibiotic acronym, class, and mechanism of action.
- **Table S3:** Performance F1-macro, F1-negative, F1-positive, accuracy, precision, recall, clinical_F1-negative, clinical_precision-negative, clinical_recall-negative) of five benchmarked

methods alongside the baseline method w.r.t. random folds, phylogeny-aware folds, and homology-aware folds.

- **Table S4: A.** Fraction of experiments with the performance above the ML-baseline threshold, stratified by the dataset splitting methods. **B.** Fraction of ML-based experiments with the performance above the indicated threshold, stratified by the dataset splitting methods. **C.** Fractions of species-antibiotic combinations the ML-based methods and rule-based method are performing best.
- **Table S5:** Performance (F1-macro, F1-positive, F1-negative, accuracy, precision, recall) of KMA-based ResFinder and BLAST-based ResFinder on each species-antibiotic combination's whole dataset (instead of on folds iteratively).
- **Table S6:** Three sets of software recommendation lists, each corresponding to random, phylogeny-aware, and homology-aware evaluation. For each species-antibiotic combination, we selected the software (or several) with the highest F1-macro mean, and in case of a tie, the lowest F1-macro standard deviation. The lists can guide researchers in selecting software for each of the 78 species-antibiotic combinations.
- **Table S7:** Performance (F1-macro, F1-positive, F1-negative, accuracy, precision, recall) of Aytan-Aktug single-species-antibiotic model using the default hyperparameters, in comparison with our modified version.
- **Table S8: A.** 13 species-antibiotic combinations for which at lease a multi-model performed best or tied best than single-models by F1-macro metric. **B-J.** Performance (F1-macro, F1-positive, F1-negative, accuracy, precision, recall) of Aytan-Aktug single-species multi-antibiotic model, control multi-species model, and cross-species model, as well as LOSO cross-species model for Aytan-Aktug, Kover and PhenotypeSeeker.
- **Table S9:** Benchmarking results of 14 datasets spanning *C. jejuni* and *E. coli* in two repeated independent evaluations via our workflow.
- **Table S10:** Mean, standard deviation, and median of F1-macro mean for each benchmarked method.
- **Table S11:** Dataset folds potentially suffering from ill-defined F1-negative and F1-positive metrics under phylogeny-aware and homology-aware evaluations.

Data and code availability

The benchmarking datasets were established based on the open-source PATRIC (https://patricbrc.org/) and are available at **Mendeley Data, doi: 10.17632/6vc2msmsxj.1**. The source code used for benchmarking AMR software and software adaptations is available on GitHub (https://github.com/hzi-bifo/AMR_benchmarking_khu). Tutorials for evaluating ML-based AMR software

are provided (https://github.com/hzi-bifo/AMR_benchmarking/wiki/AMR-phenotyping-benchmarking-tutorial).

Competing interests

E.A and T.K are two of the co-authors of software Seq2Geno2Pheno[41].

Acknowledgements

P.C.M. received funding from the Cluster of Excellence RESIST (EXC 2155) funded by Deutsche Forschungsgemeinschaft (DFG) - projekt number 390874280. This work was funded in part by the German Center for Infection Research (DZIF) Translational Infrastructure Bioresources, Biodata and Digital Health (TI BBD) - project number TI 12.002 and the Deutsche Forschungsgemeinschaft (DFG) - project number 460129525 (NFDI4Microbiota).

Key Points

- We introduce a new benchmarking standard using advanced, unbiased evaluation techniques to rigorously test the performance and generalizability of AMR prediction models.
- We performed benchmarking by training and testing four ML methods and a rule-based method on 78 species-antibiotic datasets.
- Benchmarking reveals that while ML methods excel with closely related bacterial strains, the rule-based ResFinder is superior for divergent genomes, with Kover emerging as the leading ML tool in performance.
- The study provides explicit software recommendations for AMR phenotype prediction, aiding in the accurate detection of resistance across diverse bacterial species.
- Findings emphasize the necessity for refined prediction models, especially for clinically critical antibiotics like beta-lactams, to enhance their utility in real-world applications.

Author Biographies

Kaixin Hu is a PhD student at the Department Computational Biology of Infection Research of the Helmholtz Centre for Infection Research. She focuses on machine learning and antimicrobial resistance.

Fernando Meyer is a postdoctoral researcher at the Department Computational Biology of Infection Research of the Helmholtz Centre for Infection Research. He focuses on metagenomics.

Zhi-Luo Deng is a postdoctoral researcher at the Department Computational Biology of Infection Research of the Helmholtz Centre for Infection Research. He focuses on microbial genomics and microbiome.

Ehsaneddin Asgari is a postdoctoral researcher at the Department Computational Biology of Infection Research of the Helmholtz Centre for Infection Research. He focuses on natural language processing and bioinformatics.

Tzu-Hao Kuo was a PhD student at the Department Computational Biology of Infection Research of the Helmholtz Centre for Infection Research. He is currently a research associate at Evotec International GmbH.

Philipp C. Münch is a research scientist at the Department Computational Biology of Infection Research of the Helmholtz Centre for Infection Research. He focuses on deep learning and computational biology.

Alice C. McHardy is the head of Department Computational Biology of Infection Research of the Helmholtz Centre for Infection Research. The department studies the human microbiome, viral and bacterial pathogens, and human cell lineages within individual patients by analysis of large-scale biological and epidemiological data sets with computational techniques.

References

1. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet. 2019;20: 356–370.

2. Reygaert WC. An overview of the antimicrobial resistance mechanisms of bacteria. AIMS Microbiology. 2018;4: 482–501.

3. Aljeldah MM. Antimicrobial Resistance and Its Spread Is a Global Threat. Antibiotics (Basel). 2022;11. doi:10.3390/antibiotics11081082

4. Geneva: World Health Organization. Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report 2022. Licence: CC BY-NC-SA 30 IGO. 2022.

5. Ten threats to global health in 2019. In: Geneva: World Health Organization [Internet]. Available: https://www.who.

6. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. 2022;399: 629–655.

7. Vasala A, Hytönen VP, Laitinen OH. Modern Tools for Rapid Diagnostics of Antimicrobial Resistance. Front Cell Infect Microbiol. 2020;10: 308.

8. Su M, Satola SW, Read TD. Genome-Based Prediction of Bacterial Antibiotic Resistance. J Clin Microbiol. 2019;57. doi:10.1128/JCM.01405-18

9. van Belkum A, Dunne WM Jr. Next-generation antimicrobial susceptibility testing. J Clin Microbiol. 2013;51: 2018–2024.

10. Bradley P, Claire Gordon N, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nature Communications. 2015. doi:10.1038/ncomms10063

11. Gajic I, Kabic J, Kekic D, Jovicevic M, Milenkovic M, Mitic Culafic D, et al. Antimicrobial Susceptibility Testing: A Comprehensive Review of Currently Used Methods. Antibiotics (Basel). 2022;11. doi:10.3390/antibiotics11040427

12. Geneva: World Health Organization. Molecular methods for antimicrobial resistance (AMR) diagnostics to enhance the Global Antimicrobial Resistance Surveillance System. Licence: CC BY-NC-SA 30 IGO. 2019 (WHO/WSI/AMR/2019.1).

13. Köser CU, Ellington MJ, Peacock SJ. Whole-genome sequencing to control antimicrobial resistance. Trends Genet. 2014;30: 401–407.

14. Lo SW, Kumar N, Wheeler NE. Breaking the code of antibiotic resistance. Nat Rev Microbiol. 2018;16: 262.

15. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019;37: 783–792.

16. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67: 2640–2644.

17. Zankari E, Allesøe R, Joensen KG, Cavaco LM, Lund O, Aarestrup FM. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. J Antimicrob Chemother. 2017;72: 2764–2768.

18. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genom. 2017;3: e000131.

19. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother. 2020;75: 3491–3500.

20. Clausen PTLC, Zankari E, Aarestrup FM, Lund O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. J Antimicrob Chemother. 2016;71: 2484–2488.

21. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 2014;6: 90.

22. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017;45: D566–D573.

23. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. Antimicrob Agents Chemother. 2019;63. doi:10.1128/AAC.00483-19

24. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Scientific Reports. 2021. doi:10.1038/s41598-021-91456-0

25. de Man TJB, Limbago BM. SSTAR, a Stand-Alone Easy-To-Use Antimicrobial Resistance Gene Predictor. mSphere. 2016;1. doi:10.1128/mSphere.00050-15

26. Metcalf BJ, Chochua S, Gertz RE Jr, Li Z, Walker H, Tran T, et al. 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. 2016;22: 1002.e1–1002.e8.

27. McDermott PF, Tyson GH, Kabera C, Chen Y, Li C, Folster JP, et al. Whole-Genome Sequencing for Detecting Antimicrobial Resistance in Nontyphoidal Salmonella. Antimicrob Agents Chemother. 2016;60: 5515–5520.

28. Walker TM, Miotto P, Köser CU, Fowler PW, Knaggs J, Iqbal Z, et al. The 2021 WHO catalogue of Mycobacterium tuberculosis complex mutations associated with drug resistance: A genotypic analysis. Lancet Microbe. 2022;3: e265–e273.

29. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015;15: 1193–1202.

30. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7: 51.

31. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, et al. Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. J Clin Microbiol. 2014;52: 1182–1191.

32. Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, et al. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J Antimicrob Chemother. 2013;68: 2234–2244.

33. Xavier BB, Das AJ, Cochrane G, De Ganck S, Kumar-Singh S, Aarestrup FM, et al. Consolidating and Exploring Antibiotic Resistance Gene Data Resources. J Clin Microbiol. 2016;54: 851–859.

34. Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. Using Genomics to Track Global Antimicrobial Resistance. Front Public Health. 2019;7: 242.

35. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics. 2016;17: 754.

36. Drouin A, Letarte G, Raymond F, Marchand M, Corbeil J, Laviolette F. Interpretable genotype-to-phenotype classifiers with performance guarantees. Sci Rep. 2019;9: 4071.

37. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial Resistance Prediction in PATRIC and RAST. Sci Rep. 2016;6: 27930.

38. Aun E, Brauer A, Kisand V, Tenson T, Remm M. A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. PLoS Comput Biol. 2018;14: e1006434.

39. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, et al. Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. Sci Rep. 2018;8: 421.

40. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, et al. Using Machine Learning To Predict Antimicrobial MICs and Associated Genomic Features for Nontyphoidal Salmonella. J Clin Microbiol. 2019;57. doi:10.1128/JCM.01260-18

41. Khaledi A, Weimann A, Schniederjans M, Asgari E, Kuo T-H, Oliver A, et al. Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics. EMBO Mol Med. 2020;12: e10264.

42. Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L. Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS Comput Biol. 2018;14: e1006258.

43. Her H-L, Wu Y-W. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains. Bioinformatics. 2018;34: i89–i95.

44. Hyun JC, Kavvas ES, Monk JM, Palsson BO. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. PLoS Comput Biol. 2020;16: e1007608.

45. Ren Y, Chakraborty T, Doijad S, Falgenhauer L, Falgenhauer J, Goesmann A, et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics. 2021;38: 325–334.

46. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. EBioMedicine. 2019;43: 356–369.

47. Aytan-Aktug D, Nguyen M, Clausen PTLC, Stevens RL, Aarestrup FM, Lund O, et al. Predicting Antimicrobial Resistance Using Partial Genome Alignments. mSystems. 2021;6: e0018521.

48. Pataki BÁ, Matamoros S, van der Putten BCL, Remondini D, Giampieri E, Aytan-Aktug D, et al. Understanding and predicting ciprofloxacin minimum inhibitory concentration in Escherichia coli with machine learning. Sci Rep. 2020;10: 15026.

49. Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, et al. Machine Learning Predicts Accurately Mycobacterium tuberculosis Drug Resistance From Whole Genome Sequencing Data. Front Genet. 2019;10: 922.

50. Zabeti H, Dexter N, Safari AH, Sedaghat N, Libbrecht M, Chindelevitch L. INGOT-DR: an interpretable classifier for predicting drug resistance in M. tuberculosis. Algorithms Mol Biol. 2021;16: 17.

51. Van Camp P-J, Haslam DB, Porollo A. Prediction of Antimicrobial Resistance in Gram-Negative Bacteria From Whole-Genome Sequencing Data. Front Microbiol. 2020;11: 1013.

52. Kim J, Greenberg DE, Pifer R, Jiang S, Xiao G, Shelburne SA, et al. VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. PLoS Comput Biol. 2020;16: e1007511.

53. Aytan-Aktug D, Clausen PTLC, Bortolaia V, Aarestrup FM, Lund O. Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. mSystems. 2020;5. doi:10.1128/mSystems.00774-19

54. Eyre DW, De Silva D, Cole K, Peters J, Cole MJ, Grad YH, et al. WGS to predict antibiotic MICs for Neisseria gonorrhoeae. J Antimicrob Chemother. 2017;72: 1937–1947.

55. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE Jr, Walker H, et al. Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting β-Lactam Resistance Levels in Streptococcus pneumoniae. MBio. 2016;7. doi:10.1128/mBio.00756-16

56. Rishishwar L, Petit RA 3rd, Kraft CS, Jordan IK. Genome sequence-based discriminator for vancomycin-intermediate Staphylococcus aureus. J Bacteriol. 2014;196: 940–948.

57. Green AG, Yoon CH, Chen ML, Ektefaie Y, Fina M, Freschi L, et al. A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis. Nat Commun. 2022;13: 3817.

58. Gröschel MI, Owens M, Freschi L, Vargas R Jr, Marin MG, Phelan J, et al. GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. Genome Med. 2021;13: 138.

59. Kouchaki S, Yang Y, Walker TM, Sarah Walker A, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics. 2019;35: 2276–2282.

60. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics. 2018;34: 1666–1671.

61. Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, Crook DW, et al. DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis. Bioinformatics. 2019;35: 3240–3249.

62. Yang Y, Walker TM, Kouchaki S, Wang C, Peto TEA, Crook DW, et al. An end-to-end heterogeneous graph attention network for Mycobacterium tuberculosis drug-resistance prediction. Brief Bioinform. 2021;22. doi:10.1093/bib/bbab299

63. Kouchaki S, Yang Y, Lachapelle A, Walker TM, Walker AS, CRyPTIC Consortium, et al. Multi-Label Random Forest Model for Tuberculosis Drug Resistance Classification and Mutation Ranking. Front Microbiol. 2020;11: 667.

64. Marini S, Oliva M, Slizovskiy IB, Das RA, Noyes NR, Kahveci T, et al. AMR-meta: a k-mer and metafeature approach to classify antimicrobial resistance from high-throughput short-read metagenomics data. Gigascience. 2022;11. doi:10.1093/gigascience/giac029

65. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 2018;6: 23.

66. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, et al. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. Nat Commun. 2018;9: 4306.

67. Chowdhury AS, Call DR, Broschat SL. Antimicrobial Resistance Prediction for Gram-Negative Bacteria via Game Theory-Based Feature Evaluation. Sci Rep. 2019;9: 14487.

68. Chowdhury AS, Call DR, Broschat SL. PARGT: a software tool for predicting antimicrobial resistance in bacteria. Sci Rep. 2020;10: 11033.

69. Ruppé E, Ghozlane A, Tap J, Pons N, Alvarez A-S, Maziers N, et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. Nat Microbiol. 2019;4: 112–123.

70. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE Jr, Walker H, et al. Validation of β-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences. BMC Genomics. 2017;18: 621.

71. Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S, Burnham C-AD, 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. doi:10.3389/fmicb.2016.01887

72. Nguyen M, Olson R, Shukla M, VanOeffelen M, Davis JJ. Predicting antimicrobial resistance using conserved genes. PLoS Comput Biol. 2020;16: e1008319.

73. Liu Z, Deng D, Lu H, Sun J, Lv L, Li S, et al. Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of Actinobacillus pleuropneumoniae From Whole Genome Sequences. Frontiers in microbiology. 2020;11: 48.

74. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7: 91.

75. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 2017;45: D535–D542.

76. Antonopoulos DA, Assaf R, Aziz RK, Brettin T, Bun C, Conrad N, et al. PATRIC as a unique resource for studying antimicrobial resistance. Brief Bioinform. 2019;20: 1094–1102.

77. Conesa A, Garofolo G, Di Pasquale A, Cammà C. Monitoring AMR in Campylobacter jejuni from Italy in the last 10 years (2011-2021): Microbiological and WGS data risk assessment. EFSA J. 2022;20: e200406.

78. Qin X, Wang X, Shen Z. The rise of antibiotic resistance in Campylobacter. Curr Opin Gastroenterol. 2023;39: 9–15.

79. Blair JMA, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJV. Molecular mechanisms of antibiotic resistance. Nat Rev Microbiol. 2015;13: 42–51.

80. Darby EM, Trampari E, Siasat P, Gaya MS, Alav I, Webber MA, et al. Molecular mechanisms of antibiotic resistance revisited. Nat Rev Microbiol. 2023;21: 280–295.

81. Wang H, Jia C, Li H, Yin R, Chen J, Li Y, et al. Paving the way for precise diagnostics of antimicrobial resistant bacteria. Front Mol Biosci. 2022;9: 976705.

82. Alekshun MN, Levy SB. Molecular mechanisms of antibacterial multidrug resistance. Cell. 2007;128: 1037–1050.

83. Courvalin P. Transfer of antibiotic resistance genes between gram-positive and gram-negative bacteria. Antimicrob Agents Chemother. 1994;38: 1447–1451.

84. Martinez JL. General principles of antibiotic resistance in bacteria. Drug Discov Today Technol. 2014;11: 33–39.

85. Cummins EA, Snaith AE, McNally A, Hall RJ. The role of potentiating mutations in the evolution of pandemic Escherichia coli clones. Eur J Clin Microbiol Infect Dis. 2021. doi:10.1007/s10096-021-04359-3

86. Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE, Sebra R, et al. Evolutionary History of the Global Emergence of the Escherichia coli Epidemic Clone ST131. MBio. 2016;7: e02162.

87. Marsh JW, Mustapha MM, Griffith MP, Evans DR, Ezeonwuka C, Pasculle AW, et al. Evolution of

Outbreak-Causing Carbapenem-Resistant Klebsiella pneumoniae ST258 at a Tertiary Care Hospital over 8 Years. MBio. 2019;10. doi:10.1128/mBio.01945-19

88. Blount ZD, Lenski RE, Losos JB. Contingency and determinism in evolution: Replaying life's tape. Science. 2018;362. doi:10.1126/science.aam5979

89. Beavan A, Domingo-Sananes MR, McInerney JO. Contingency, repeatability, and predictability in the evolution of a prokaryotic pangenome. Proc Natl Acad Sci U S A. 2024;121: e2304934120.

90. Břinda K, Callendrello A, Ma KC, MacFadden DR, Charalampous T, Lee RS, et al. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. Nat Microbiol. 2020;5: 455–464.

91. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet. 2018;50: 307–316.

92. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015. doi:10.1101/gr.186072.114

93. Chlenski P. A machine learning-based service for estimating quality of genomes using PATRIC. BMC Bioinformatics. 2019. doi:10.1186/s12859-019-3068-y

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

95. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31: 3691–3693.