From sequence to signature: Uncovering multiscale AMR features across bacterial pathogens with supervised machine learning

Abhirupa Ghosh[1,#], Evan P Brenner[1,#], Charmie K Vang[1,#], Ethan P Wolfe[1,#], Joseph T Burke[2], Raymond L Lesiyon[1], Keenan R Manpearl[1], Vignesh Sridhar[2], Jacob D Krol[1], Emily Boyer[1], Jill MR Bilodeaux[1], Karn Jongnarangsin[2], Elliot J Majlessi[2], Janani Ravi[1,*].

[1]Department of Biomedical Informatics, Center for Health Artificial Intelligence, University of Colorado Anschutz Medical Campus, Aurora, CO, USA; [2]Department of Pathobiology and Diagnostic Investigation, Michigan State University, East Lansing, MI, USA. [#]Co-primary, contributed equally. *Corresponding author: janani.ravi@cuanschutz.edu

Keywords: Comparative pathogenomics; Antimicrobial resistance; AMR prediction; Supervised machine learning; Bacterial pathogens

Abstract

Since the clinical introduction of antibiotics in the 1940s, antimicrobial resistance (AMR) has become an increasingly dire threat to global public health. Pathogens acquire AMR much faster than we discover new drugs, warranting new methods to better understand the molecular underpinnings of AMR. Traditional approaches for detecting AMR in novel bacterial strains are time-consuming and labor-intensive. However, advances in sequencing technology offer a plethora of bacterial genome data, and computational approaches like machine learning (ML) provide an optimistic scope for *in silico* AMR prediction. Here, we introduce a comprehensive multiscale ML approach to predict AMR phenotypes and identify AMR molecular features associated with a single drug or drug family, stratified by time and geographical locations. As a case study, we focus on a subset of the World Health Organization's Bacterial Priority Pathogens, frequently drug-resistant and nosocomial bacteria: *Enterococcus faecium*, *Staphylococcus aureus*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and two *Enterobacter* species. We started with sequenced genomes with lab-derived AMR phenotypes, constructed pangenomes, clustered protein sequences, and extracted protein domains to generate pangenomic features across molecular scales. To uncover the molecular mechanisms behind drug-/drug family-specific resistance, we trained logistic regression ML models on our datasets. These yielded ranked lists of AMR-associated genes, proteins, and domains. In addition to recapitulating known AMR features, our models identified novel candidates for experimental validation. They performed well across molecular scales, data types (binary vs. counts), and drugs. The models achieved a median normalized Matthews correlation coefficient of 0.89, suggesting robust performance across molecular scales with counts or presence/absence of features. Model performance showed resilience even when evaluated on geographical and temporal holdouts. Our holistic approach promises reliable prediction of existing and developing resistance in newly sequenced pathogen genomes, along with the mechanistic molecular contributors of AMR.

Introduction

Antimicrobial resistance (AMR) is a growing global health crisis. The rise of resistant bacterial pathogens is steadily compromising the effectiveness of antibiotics, long regarded as a cornerstone of modern medicine since their clinical integration in the 1940s[1–3]. In the absence of effective therapies, even routine infections and common medical interventions could result in outcomes reminiscent of the pre-antibiotic era (e.g., amputation or quarantine)[1]. Bacterial AMR can be directly attributable to more than one million deaths annually and associated with nearly five million, with the highest burden falling on

low-resource settings[4]. The UK Review on Antimicrobial Resistance warns that the direct death toll could reach ten million per year by 2050[5]. Further, the rise of multidrug-resistant (MDR) bacteria, defined as pathogens resistant to three or more classes of antibiotics, has dramatically narrowed viable treatment options[6]. Compounding this threat, resistance to "last-resort" antibiotics such as carbapenems and polymyxins is becoming more common[7,8]. These pan-resistant strains leave clinicians with no effective treatment options, leading to higher mortality rates and impairing outbreak containment. At the same time, there has been a marked decline in the successful development and clinical deployment of new antibiotics[9]. This stagnation not only lags behind the pace of resistance evolution but also is fueled by economic disincentives and regulatory hurdles that have eroded industry investment in antibiotic innovation[10]. Diagnostic limitations further compound the problem, as many healthcare settings lack the capacity to identify resistance phenotypes rapidly and accurately, delaying appropriate treatment and allowing resistant strains to spread unchecked. A group of pathogens designated by the World Health Organization (WHO) as Bacterial Priority Pathogens are among the leading contributors of antibiotic-resistant infections and AMR-associated mortality[11]. The persistence and virulence of these pathogens stem not only from spontaneous mutations and horizontal gene transfer (HGT), but also from their arsenal of intrinsic resistance mechanisms (e.g., biofilm formation) that make them inherently difficult to treat[12]. In clinical settings, this biological advantage is aggravated by human behavior, where up to half of antibiotic prescriptions involve unnecessary or incorrect antibiotics[13,14]. This creates sustained selective pressures that propagate AMR emergence and harm patient outcomes.

Although AMR continues to rise, the molecular mechanisms underlying drug resistance remain poorly understood[15–17]. Most drug susceptibility testing only provides a phenotypic readout (i.e., a growth defect in the presence of an antibiotic) but does not reveal *why* resistance occurs[13]. While the genetic determinants have been identified for select drug-organism pairs through rigorous laboratory testing[7,17], much of the resistome remains unexplored, especially for MDR strains and non-model pathogens. Whole genome sequencing (WGS), when paired with modern technologies like next-generation sequencing[18], offers unprecedented access to complete bacterial genomes. Over the past decade, the rapid expansion of sequencing capacity and plummeting costs have generated vast genomic datasets[19]. Yet, the integration of WGS into routine AMR surveillance or diagnostics has been limited[13].

Machine learning (ML) has been applied in recent years to leverage omics datasets for the identification of biomarkers of human disease, such as cancer and heart disease[20]. Researchers have begun exploring the prediction of AMR with ML and deep learning (DL) as well[18,19]. For instance, DeepARG uses a DL framework trained on curated antibiotic resistance gene (ARG) databases to classify short sequencing reads into resistance categories with high precision[21]. Other efforts like AMRLearn use single nucleotide polymorphism (SNP) genotypes in various ML classifiers to predict AMR[22]. Despite their promise, these models face significant limitations. Most rely on known resistance genes catalogued in curated databases (e.g., Comprehensive Antibiotic Resistance Database (CARD[23]), ResFinder[24]), which biases predictions toward previously characterized mechanisms and limits the discovery of novel or rare variants.

Herein, we demonstrate the use of interpretable logistic regression (LR) models to predict AMR across multiple molecular feature scales and species using public data from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC)[25]. By incorporating gene-, protein-, and domain-level features, our models capture distinct layers of biological function, enabling broader generalization and deeper mechanistic insight into resistance determinants. We demonstrate the utility of this workflow by exploring AMR in the context of a selection of the WHO's Priority Pathogens. Complete and draft (WGS) genomes with available lab-derived AMR phenotypic data were chosen, and feature matrices were created using the presence/absence (binary) or counts of genes, proteins, domains, and gene neighborhood rearrangements from taxon-specific pangenomes. Models were trained and tested on drugs across multiple classes per species. These models perform well across pathogens, drugs, and feature scales by standard ML performance metrics, provide a testable output of features predicted to be associated with AMR, and are computationally fast and lightweight. An inherent strength of ML is that it does not rely on *a priori* loci annotated to be involved.

Instead, our models identify associations between features and genome resistance/susceptibility labels, permitting detection of previously unknown AMR-associated molecular mechanisms. Such approaches are essential as resistance develops and spreads, and new pathogens emerge. This work offers multiscale ML as a holistic method of AMR detection and attribution.

Our framework for elucidating previously obscure AMR mechanisms enables the development of nuanced questions. Will one type of molecular feature prove better than another at predicting resistance for a given combination of pathogen and antibiotic? If so, why might that be? How does the format of input data, such as binary (presence/absence of features) vs. counts, impact model performance and interpretability? Which features contribute to resistance across species and drugs? Finally, we leverage temporal and geographic metadata to test the robustness of our models with various holdouts and assess limitations. This allows us to validate the generalizability of the models to genomes from different continents and future samples. Alongside, we identify features that are unique or shared contributing factors across holdouts.

Methods

Data collection, preprocessing, and metadata curation

To predict whether a bacterial isolate is resistant or susceptible to a given drug based on its genome and to identify the genomic features that drive these predictions, we developed an ML pipeline, as illustrated in **Figure 1**. We downloaded assembled genome sequences, encoded protein sequences, and corresponding isolate metadata, including laboratory-derived drug susceptibility phenotypes per genome, from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC[25]; accessed June 20, 2025) using their BV-BRC command-line interface P3-scripts (`v5.3`) and custom R functions. We focused on a subset of the WHO Bacterial Priority Pathogens List: *Enterococcus faecium*, *Staphylococcus aureus*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter cloacae* and *hormaechei*[26]. Query parameters included both taxon IDs and genome names to enable species-specific targeting of divergent and taxonomically complex clades. Genomes were filtered to include only “Good” quality, “Complete” and “Draft/WGS” assemblies, with plasmid sequences when available. Retrieved genome metadata were parsed and stored in local DuckDB[27] databases in Apache Parquet format to support fast querying, smaller file sizes, and reproducible updates. AMR-associated metadata were downloaded and merged with genome metadata using genome IDs. Reported AMR phenotypes for each genome-drug pair were mapped to binary resistant or susceptible labels. Intermediate phenotypes were excluded due to their low frequency and unclear biological significance in distinguishing true resistance from susceptibility. Drug names were harmonized and mapped to antibiotic classes based on WHO AWaRe guidelines[28] and AntibioticsDB[29]. Metadata curation also included standardizing isolate identifiers and ensuring consistency across genome-, protein-, and domain-level matrices.

Generating pan-feature matrices

Genes. We used Panaroo[30] (`v1.5.2`) to cluster homologous genes and construct bacterial pangenomes, producing both genome-by-gene count matrices (`gene_count`) and presence/absence (binary) matrices (`gene_binary`) with AMR phenotype labels per genome. Genes were grouped based on a 95% sequence identity threshold, allowing up to 5% length difference between sequences to account for minor structural variation. Gene families were defined using a 50% identity threshold. Gene triplet forks (i.e., points in the pangenome graph where gene neighborhoods structurally vary between genomes) were also represented in binary format (`struct`).

Proteins. Annotated protein fasta files per isolate were clustered into panproteomes using CD-HIT[31] (`v4.8.1`) with default parameters, except that description length limits were disabled (-d 0). Protein-level matrices were generated as both counts (`protein_counts`) and binary (`protein_binary`).

Domains. Protein domain analysis was conducted using InterPro[32] (`v5.74-105.0`), focusing on Pfam/Xfam annotations. Domains were similarly converted to count and binary matrices (`domain_counts`, `domain_binary`).

All features were compressed as columnar sparse matrices in Parquet format to minimize file size, then linked through DuckDB views for efficient SQL querying of feature subsets.

Machine learning model set-up

To identify encoded features predictive of AMR phenotypes, we implemented a supervised ML framework using LR. In addition to their strong empirical performance in recovering known AMR genes, LR classifiers[33] were selected for their simplicity, interpretability, strong baseline performance, and reasonable compute times. All models were built using the tidymodels ecosystem in R[33] and applied across multiple feature scales.

For each drug or drug class with sufficient data in each species, we subsetted the feature spaces to those genomes tested against that compound using DuckDB. Datasets were used with 5-fold cross-validation using dynamic stratified sampling to preserve resistant/susceptible class ratios, ensure each fold has sufficient observations, and enforce a minimum of 40 samples per model.

Logistic regression

LR models were trained using *glmnet*[34], with hyperparameters tuned over a predefined grid of penalty (λ) and mixing (α) values. The mixing hyperparameter ranged from 0 (ridge regression, L2) to 1 (lasso, L1), with intermediate values representing elastic net models in increments of 0.2. Models were trained using a grid search over this hyperparameter space, and the best model was selected based on Matthews correlation coefficient (MCC).

Final performance metrics, particularly normalized MCCs (nMCC), were computed on held-out testing sets for comparison across feature types and drug classes. The **MCC** accounts for all four elements of a binary confusion matrix: true/false positives and negatives in a single metric, which we have normalized to a 0–1 range for ease of interpretation. In this framework:

- An **nMCC of 1.0** indicates perfect prediction,
- An **nMCC of 0.5** reflects random performance, and
- An **nMCC of 0.0** represents complete disagreement with the true labels.

Together, they provide a balanced, interpretable framework for assessing predictive performance across drugs and molecular feature scales.

Following model training and evaluation, we extracted feature importance scores using the `vip::vi()` function[35]. To prioritize biologically informative predictors, we ranked features by their absolute variable importance and retained the top *n* features that together capture ≥50% of the overall importance described in the model.

For each drug-specific model, we save performance metrics, beta-coefficients, and the selected top features along with their corresponding importance values and signs. The latter are interpreted as the strength of feature associations with resistance (if negative) or susceptibility (if positive).

Label shuffling

To test model robustness, we performed random shuffling of AMR phenotype labels and retrained the models. Their performances with and without label shuffling were compared to evaluate their ability to discern biological significance in the real data.

Fisher's exact test baseline

As a baseline comparison against ML, we performed feature-wise two-sided Fisher's exact tests on binary genomic matrices representing feature presence or absence to identify individual molecular features associated with AMR. We selected a two-sided alternative hypothesis to capture both positively and negatively associated features. To account for multiple hypothesis testing, we applied Benjamini-Hochberg (BH) correction to control the false discovery rate (FDR). Adjusted *p*-values and binary significance calls (based on an FDR threshold, default $Q = 0.05$) are reported in the *Results* section.

Annotations of the top features

Using HMMER v3.4[36], top features were mapped to NCBI's Reference HMM Catalog of ARGs[37] to identify known AMR-associated mechanisms. Features were also mapped to HMMs for clusters of orthologous groups (COGs[38]) for cross-species comparisons.

Dashboard

We developed an interactive dashboard to support exploration of AMR patterns by integrating manifold metadata with ML outputs. The dashboard, its design inspired by the Pavian metagenomics Shiny app[39], is organized into two primary sections: AMR data and AMR prediction, each tailored to enable researchers to dynamically interrogate the data.

The metadata section of the dashboard enables users to filter, query, and visualize AMR surveillance data by phenotype, drug, specific agents, country of isolation, host, and isolation source. These visualizations reveal both the diversity of resistance data and the heterogeneity of sampling contexts and suggest that resistance trends must be interpreted within a broader ecological and geographical framework. The dashboard allows users to regenerate such plots based on filtered queries and fosters hypothesis generation and data-driven prioritization.

The modeling section of the dashboard allows users to explore pre-computed ML models that predict resistance phenotypes from genomic data. For each bacterial species and drug, users can inspect model performance metrics such as nMCC. Users may also view ranked lists of predictive genomic features, including gene clusters or protein families associated with resistance or susceptibility, thus enabling comparison of models across species, drugs, and data types. Additionally, users can investigate cross model performance, and contributing important features across geographical regions and time.

We offer this dashboard not just for passive viewing but as a starting point for further investigation, guiding users toward the most informative genomic markers and most robust prediction targets across molecular scales and data types.

Results

Compiling a comprehensive (meta)dataset

To systematically predict AMR using a multiscale ML approach, we first assembled a large curated dataset spanning select WHO priority pathogens, a group of clinically significant bacteria known for their multidrug resistance and global health impact. Across the species analyzed, despite 100s of thousands of sequenced genomes, the number of available isolates fell sharply when filtered to those with lab-confirmed AMR phenotype data but was sufficient for most modeling (**Figure 2**). Genomic data and metadata were acquired for *Enterococcus faecium* (n = 1106), *Staphylococcus aureus* (n = 3612), *Acinetobacter baumannii* (n = 1710), *Pseudomonas aeruginosa* (n = 1323), and two *Enterobacter* species (n = 499). Constructing this dataset presented both an opportunity and a challenge: although all data were sourced from BV-BRC, assembling an ML-ready framework required extensive harmonization of genomic sequences, antimicrobial

susceptibility profiles, and molecular annotations. We extracted and integrated multiple layers of information, including sample metadata (e.g., geographic origin and isolate collection data) and multiscale molecular features such as gene, protein, domain, and structural-level count and binary representations. This unified dataset enabled downstream within-species and cross-species analyses of resistance at multiple biological scales.

In total, the data spanned 47 drugs (single or in combination) and 23 drug classes across selected pathogens. AMR metadata are available for visualization through our dashboard (JRaviLab.org/amr, `AMR data` tab), showing patterns across resistance phenotypes, drug classes, geographic origin, host types, and clinical specimen sources. Resistance profiles varied substantially, reflecting both pathogen-specific biology and selective antibiotic pressures due to clinical history. For instance, *A. baumannii* exhibited extensive resistance across β-lactams, fluoroquinolones, aminoglycosides, and tetracyclines (**Figure 2**). Notably, resistance to carbapenems such as meropenem (273 resistant vs. 111 susceptible) was widespread and consistent with *A. baumannii*'s classification as a WHO critical priority pathogen. Resistance to older antibiotics like chloramphenicol and streptomycin also persisted, reflecting multidrug resistance that has become fixed in the population even after discontinuation of the clinical routine use of these older drugs. For instance, metadata analysis of *A. baumannii* also revealed that isolates were predominantly sourced from human hosts, with the majority originating from blood, respiratory, wound, and tissue samples, and was dominated by entries from the US (n = 1107), with some sampled from Tanzania, Thailand, China, and Germany. In *S. aureus*, high levels of resistance were observed against fluoroquinolones (e.g., 1130 resistant to ciprofloxacin), macrolides (e.g., 984 to erythromycin), and lincosamides (505 to clindamycin), reflecting a broad, population-level resistance burden. Geographically, the *S. aureus* dataset was skewed toward isolates from the US and UK (n = 642, 1115 respectively), with additional representation from Asia, Europe, and South America. These patterns likely reflect sampling emphasis and sequencing capacity rather than the true geographic burden of clinical AMR.

This structured metadata landscape reveals distinct, lineage- and region-specific resistance trends across pathogens and highlights critical reservoirs of resistance both within and beyond clinical settings. The integrative view supports targeted modeling efforts and enhances our understanding of AMR dissemination dynamics across ecological and geographic boundaries.

Our models are highly performant across molecular scales and data types

To evaluate the predictive utility of AMR at different molecular resolutions, we trained per-drug LR models using gene, protein, and domain-scale features for binary and count feature matrices independently. The class balance of AMR phenotypes for isolates with data against individual drugs in the original BV-BRC dataset ranged from 1% of *P. aeruginosa* genomes being colistin resistant to 96% of *A. baumannii* genomes being cefotaxime resistant. Such extreme examples with limiting training data for ML were filtered out. To account for variable class balance, we reported nMCC values, which encompass true/false positives and negatives in a single metric. Across species, scales, and drugs, most showed excellent predictive performance (median nMCC = 0.89), far surpassing random classifiers (nMCC ~ 0.5) (**Figure 3A;** JRaviLab.org/amr, `AMR prediction` tab; **Figure S1**), lending confidence to our models' ability to uncover biological significance from multiscale data paired with AMR phenotypes. In fact, no single molecular scale consistently yielded superior performance across the board, likely reflecting the biological contribution of each scale to AMR. In *A. baumannii*, protein-level models were the best performers across drugs and data types, but in *S. aureus*, gene-level models performed better, albeit marginally. *E. faecium* showed uniformly high performance regardless of feature scale. Models by count and binary data types showed comparable nMCC. For most pathogens, nMCCs differed only slightly between binary and count formats, suggesting that the choice of data type has a limited impact on predictive performance. In *E. faecium*, *S. aureus*, and *A. baumannii*, both data types consistently yielded high nMCC values (typically >0.9), indicating that both binary and count-based feature encodings are viable for AMR prediction.

Focusing on drug classes, we investigated whether top predictors for AMR can be generalized across species. We focused on drug classes because antibiotics within a class often share mechanisms of action and resistance, enabling models to capture generalizable biological patterns. This approach also improves statistical power by aggregating data across related drugs, which is especially important when individual compounds have limited representation. Models were also successful in predicting resistance to classes of antibiotics (**Figure 3B**), but performance varied considerably from species to species. While the penicillin class was well-modeled in only two species with sufficient data, the aminoglycoside class performance in tested pathogens ranged from stellar for *E. faecium* and *S. aureus* (nMCC ~ 1.0) to modest for *A. baumannii* and *Enterobacter* spp. (nMCC ~ 0.7–0.8). In our dataset, the aminoglycoside class consisted of amikacin, gentamicin, and tobramycin (individual drug model performance shown in **Figure S2**). Our LR domain binary models achieved high predictive accuracy (nMCC > 0.8) across most species.

Evaluating a case study of AMR in *Staphylococcus aureus* and *Acinetobacter baumannii*

Our models yield ranked AMR feature lists for each combination of pathogen and drug. Here, we highlight *S. aureus* and *A. baumannii*, two evolutionarily distinct opportunistic pathogens that differ in morphology and physiology. *S. aureus* and *A. baumannii* had sufficient data for ~20 antibiotics. A comparison of models across gene, protein, and domain scales shows good performance across a subset of drugs used in both species (**Figure 4A**), and for species-unique antibiotics (**Figure 4B**). In *S. aureus*, the top features associated with gentamicin resistance included well-established aminoglycoside-resistance mechanisms across all scales (Domains: PF13523, acetyltransferase domain; genes: acetyltransferase, aacA-aphD; proteins: acetyltransferase, AAC(6')-Ib/AAC(6')-II). Tetracycline resistance was predicted by tetracycline-resistance proteins and genes such *tetK* and *tetM*, and also returned a plasmid replication initiation factor domain (PF02486) and replication initiation protein that implicate small plasmids, known as the dominant staphylococcal tetracycline resistance mechanism[40]. These findings validate the ability of our model to recover known resistance determinants from genome-wide feature sets, as well as highlighting the importance of mobile genetic elements (MGEs) in AMR. In *A. baumannii*, top predictors for fluoroquinolone resistance included plasmid replication proteins, transposases, and prophage-associated hypothetical proteins, and for aminoglycosides, phage integrase domains in addition to resistance-conferring *aadB* genes. In erythromycin-resistant *S. aureus*, our models recovered multiple genes and proteins associated with resistance, including ErmC, MsrA, *ermA*, and *ant(9)-I*, which are often embedded within genomic islands or clusters with hypothetical proteins.

Many high-importance features lacked functional annotations and were frequently labeled as "hypothetical" or "domain of unknown function." Notably, these features often appeared in close genomic proximity to known resistance genes or MGEs, suggesting potential co-selection or regulatory linkage. This was particularly striking for *A. baumannii*, where resistance profiles appeared to rely more heavily on poorly annotated genomic content, compared to the more known-ARG-centric patterns in *S. aureus*.

We also observed both shared and drug-specific resistance markers. Some genes, such as *ant(9)-I* and mobile element proteins, were selected across drug models, indicating their potential roles in multidrug resistance.

Overall, the median nMCC across *S. aureus* and *A. baumannii* models was high (0.94 and 0.90, respectively). Some models yielded questionable perfect nMCC performance in some extremely imbalanced datasets – for example, vancomycin models for *S. aureus* nearly all returned nMCC values of 1. While vancomycin resistance is sufficiently conferred by *van* operons, numerous model predictors evenly share top feature importance values while the resistance-determining features *vanA/B* do not stand out; this result may arise from very small numbers of clonal resistant isolates yielding an array of mobile genetic elements (transposases, phage integrases, etc.) and genomic features potentially unrelated to vancomycin AMR (e.g., thymidylate synthase, dihydrofolate reductase, DegV family protein, etc.), but that models detect as

associated. Such models will not be generalizable to other, phylogenetically unrelated outbreak settings and should be treated with caution.

In *A. baumannii*, *S. aureus*, and our other analyzed pathogens, our approach yields many high-performing models backed by interpretable features across multiple scales that recapitulate known resistance mechanisms, emphasize the critical roles of MGEs in AMR, and expose underexplored features that may contribute to the adaptive landscape of AMR. These findings offer both mechanistic insight and avenues for hypothesis generation regarding novel resistance determinants. Performance results for all models are available in **Table S1** and our dashboard (JRaviLab.org/amr).

Testing the robustness of our model signatures

Robustness is an essential quality for ML models, and refers broadly to whether a model can still perform predictably while withstanding perturbations in input data types. A model trained on historic samples from a single hospital's patient population may perform exceptionally well when tested on samples from the same population, but it may be overfit to detecting only a narrow slice of biologically possible features observed in that setting. A robust model can be generalized to other settings and is resilient to input data changes. For our AMR models to be useful for isolates from around the world and into the future, we asked whether the features associated with resistance to the same drug or class vary with geographical location or temporal bins. Will our models be useful for isolates from newly sampled countries, or on samples collected five years from today? We trained models specific to countries and five-year bins to compare the context-specific models across drugs or classes. Further, to assess generalizability, we trained models on isolates from one stratified group and tested them on others. As an example, we compare the country-specific models to predict lincosamide resistance in *S. aureus*. In that species, lincosamide data were available for 4 countries: Ireland, Singapore, Thailand, and the US. The lincosamide model with all countries taken together in *S. aureus* with gene count features is a performant model with nMCC of 0.95. When the data was stratified, the performance increased for Ireland to nMCC = 1; however, the model specific to Singapore genomes fails to predict Susceptible samples, whereas the US and Thailand-specific models achieved 0.9 nMCC. All these country-specific models were further subjected to a blind test using genomes from the remaining countries. **Figure 6A** shows the performance of each country-specific model tested on others. Although all models have performed poorly while testing on the US dataset, the US model was successful in predicting resistance phenotypes across genomes from other countries.

Similar to the geographical stratification, temporal holdouts and cross-testing on held-out time periods were performed. The collection years of the genomes were binned into 5-year intervals, and each dataset specific to a species and antibiotic combination was stratified based on these bins. An example would be the aminoglycosides drug class dataset in *S. aureus*, with AMR data from 2000 to 2019, stratified into 4 bins, and, therefore, 4 trained models. The model of the aminoglycosides class that considers entire time frame in *S. aureus* was almost perfectly predictive with nMCC = 0.99; the stratified temporal models of aminoglycosides performed relatively well, too, with nMCC > 0.9, with the exception of the 2010–2014 group, whose performance dropped to ~0.7. Each of the temporal models were further tested on other temporal groups to evaluate the robustness of the models. It is interesting to observe from the nMCC comparisons in **Figure 6B,** all the models performed poorly when tested on the 2015–2019 dataset. In the 2000–2004 model, aminoglycoside acetyltransferases—a known mechanism of aminoglycoside resistance—were two of hundreds of ranked features with approximately equal and low feature importance, indicating that while they were predictive, their contribution to AMR did not stand out in this time period. Conversely, the top two features in the 2010–2014 model were the same aminoglycoside acetyltransferases, but now in a much shorter list of top features and with ~1.3x and ~2.0x greater importance scores than the next two closest top ranked features on the list, respectively, suggesting increasing prominence of these genes in AMR. Finally, in the 2015–2019 model, these two same features now showed ~8x greater feature importance than the next more important genes, reflecting their dominant role in aminoglycoside resistance. These genes are prevalent

in mobile genetic elements and plasmids and are often horizontally transferred, and this shifting importance across time reflects the establishment and fixation of this mechanism of resistance. The performance of the geographical and temporal models are available in **Table S2 and S3** and it can be visualized along with important features in the dashboard (JRaviLab.org/amr).

To evaluate how well our ML models identify key resistance-associated genes, we compared their feature importance scores to results from Fisher's exact test, a standard statistical baseline for detecting gene-level associations given AMR phenotypes per isolate. As shown in **Figure S3A**, many features were statistically significant after multiple testing correction (Benjamini-Hochberg), yet most exhibited low ML importance scores, indicating that statistical enrichment alone does not equate to predictive power. Conversely, several features with high ML importance did not pass the adjusted *p*-value threshold, suggesting that ML can detect multivariate or context-dependent signals that univariate tests may miss. This divergence highlights a key limitation of traditional statistical approaches: they fail to account for interactions and feature dependencies present in complex biological systems. **Figure S3B** ranks the top 20 features by ML importance, revealing that while many overlap with statistically significant features, several (e.g., PF13550, PF05159) would have been overlooked by Fisher's test alone. These features may play predictive roles only when considered alongside others, an insight that classical methods cannot uncover. Together, these results demonstrate the necessity of incorporating ML to uncover nuanced, biologically relevant patterns that go beyond what is statistically enriched. By combining both approaches, we gain a more complete and functional understanding of the features associated with resistance.

Characterizing multispecies AMR patterns against a single drug

We asked whether the same drug selects for similar resistance mechanisms across phylogenetically diverse bacterial species. By assessing conserved patterns of resistance, we uncover shared evolutionary responses to antimicrobial pressure, which could inform surveillance strategies, predict resistance emergence, and guide cross-species stewardship efforts. Gentamicin was uniquely suited for this analysis, since AMR data were available across all pathogens (**Figure 5A**). The proportion of resistant genomes ranged from 0.12 in *S. aureus* to 0.87 in *A. baumannii*. Nonetheless, using nMCC, which remains robust under imbalanced classes, we compared data types and molecular feature scales for predicting gentamicin resistance across bacterial species. All models showed strong performance with nMCC > 0.80, with the exception of the *P. aeruginosa* gene structure model. In particular, models using domain-level or protein-level binary encodings generally achieved the highest median nMCC, followed closely by domain and protein count data. Gene-level models, whether with binary values or counts, also performed well but showed slightly lower performance and greater variability across species. Structural variant features underperformed, which suggests that added feature complexity may not necessarily yield better predictions. Species-level differences persisted, with *S. aureus* and *Enterobacter* spp. achieving high performance above nMCC > 0.9, while *E. faecium*, *A. baumannii*, and *P. aeruginosa* showed more variability. It is important to note, however, that although nMCC in *S. aureus* was near perfect, such performance can be a warning sign of overfitting or extreme class imbalance. However, in *S. aureus*, these suspiciously high values are well-supported by high feature importance scores for top predictive features that are known causative mechanisms of aminoglycoside resistance (i.e., acetyltransferase domains, genes, and proteins).

E. faecium, *P. aeruginosa*, and *A. baumannii* returned more moderate performance metrics, but also showed more imbalanced resistance distributions (e.g., *A. baumannii* res_prop ≈ 0.87). These data remind us that class imbalance can inflate performance metrics, and it is critical to consider resistance prevalence when interpreting model accuracy. Additionally, the number of observations varied significantly by species, with *S. aureus* having the largest dataset, potentially contributing to model stability. These findings underscore the need for balanced and sufficiently large datasets to ensure fair cross-species comparisons of predictive performance.

At the feature level, we mapped the top features to existing ARGs from NDARO (**Figure 5B**). Among domain binary models, ANT(9) was identified as a top feature in *E. faecium*, *P. aeruginosa,* and *E. species,* indicating shared resistance mechanisms against gentamicin. Other ARG hits included *mecI*, the transcriptional repressor for *mecA* and cause of methicillin resistance, and qac SMR, the small multidrug resistance antibiotic efflux pump. While distinctly different mechanisms, they may serve as markers for multidrug resistance and help effectively predict gentamicin resistance as well. Among COGs (**Figure 5C**), we identified a nucleotidyltransferase in the MJ0604 family, as well as numerous phage, plasmid, transpose/insertion sequence elements across multiple species. These COGs have not been implicated in AMR directly, but most are linked to HGT, which underlies typical aminoglycoside resistance.

Questions enabled by our approaches

Our data dashboard will aid navigation through these vast curated AMR datasets across multiple species, starting with the WHO pathogens highlighted here. The dashboard includes two primary sections allowing users to explore AMR isolate data and our AMR prediction models. AMR data illustrates the distribution of genomes across species, geographical location, and years. AMR Prediction allows comparison of multiple drug or drug class models in one species, as well as comparison of one drug/drug class across multiple species. The model performance and important variables can be visualised with both options. Along with these, geographical and temporal holdouts and cross-testing model performances and important variables can be compared. Our analysis, available in an interactive easy-to-use manner via the dashboard (JRaviLab.org/amr), enables us to answer questions such as which molecular scale performs better than others in each species? What are the features conserved across molecular scales for a particular drug or drug class resistance? How do the features vary among drugs of the same drug class in a species? What are the shared or unique features across bacterial pathogens for the same drug or drug class? Is a model trained on a particular country or time point able to predict on data from other countries or newer time frames? Do the important features change over time or vary with geography?

Discussion

Our investigation yields significant insights into the genetic underpinnings of AMR in WHO priority pathogens, demonstrating the predictive capabilities of ML across diverse bacterial species, molecular scales, and data types. Multiscale molecular features for resistance prediction consistently achieved high predictive performance across pathogens.

Feature scale performance: Across bugs and drugs, we observed roughly equivalent performance for different feature scales (**Figure 3B**). Structural variants tended to slightly underperform by nMCC metrics compared to other scales, but also took more compute time, yielded larger, more diluted lists of predictive features. On the other hand, domain models ran ~50x faster than struct and ~43x faster than the protein features that the domains were originally extracted from **(Figure S4**). This comparison does not factor in the extensive time to cluster the protein space through CD-HIT or to extract domains from InterPro, but it does illustrate how efficient domain-based models are without sacrificing performance in the tested species. It is important to note that not all proteins have annotated domains, and if such proteins yielded resistance, they would escape detection by a domain-only approach. Additionally, feature binary and count models typically performed equivalently in the tested species, and when excluding binary-only struct models, their model runtimes were nearly identical. Given the potential that specific forms of AMR could be modulated by gene copy number, we may give an edge to count models for their ability to detect such differences, but it yielded few advantages in our testing.

Top AMR features: We find that for most models, known ARGs constituted only a minor fraction of the top predictive features across our studied species, despite their undisputed clinical relevance. This challenges the conventional reliance on canonical ARG databases as the sole source for accurate resistance prediction. Instead, our models also find predictive value in other genomic elements described below, which could impact resistance through mechanisms such as gene regulation, epistasis, or co-selection. For example, regulatory changes can influence AMR by increasing the expression of efflux pumps, helping bacteria evade antibiotic effects even without changes to the underlying genes. Epistasis may affect AMR when a compensatory gain or loss of a gene or mutation restores bacterial fitness while preserving resistance. Co-selection occurs when genes linked to AMR are maintained in a population due to selective pressures from unrelated agents like metals, promoting resistance even in the absence of antibiotic exposure.

Hypothetical proteins, genes, and domains of unknown function were common contributors to accurate resistance predictions, sometimes comprising over half of the top predictive genes. The prevalence of uncharacterized loci highlights a substantial knowledge gap in our understanding of AMR and presents a compelling opportunity for novel discovery in their AMR roles.

Features associated with HGT and mobile genetic elements (MGEs) were highly prevalent among predictive features as well. This aligns with the well-documented key role of MGEs in the acquisition and dissemination of resistance in taxa such as *S. aureus* and *A. baumannii*. When analyzing the top features driving AMR, we must consider the critical roles that gene flow within and across diverse taxonomic groups plays. While Pfam domains can be applied readily for cross-species analysis, comparing wildly variable gene or protein names can introduce problems even within a single species. To work around the roadblocks to understanding which features are common drivers of AMR across pathogens, we leveraged the power of hidden Markov models (HMMs) for distant homology detection through HMMER and an HMM database of COGs[38]. By assigning gene or protein clusters to COGs, we bridge phylogenetic distance and find readily comparable features (**Figure 5C**).

Nonetheless, our models frequently successfully identified established AMR mechanisms. A direct and compelling validation for known AMR mechanisms was observed in *S. aureus*, where we consistently identified acetyltransferases as the top two predictive features for gentamicin resistance across both gene and protein count and binary models. Furthermore, the GNAT acetyltransferase[41] domains contained within proteins were also the most predictive for domain count and binary models. As such, our models directly align with established mechanisms of aminoglycoside resistance, where aminoglycoside acetyltransferases are known to chemically modify the drug to prevent its ribosomal binding and confer resistance[41,42].

Advances in the AMR field: ML approaches are rapidly growing as pivotal tools in addressing the critical global health threat of AMR by offering novel avenues for prediction and intervention. These sophisticated models harness a diverse array of data from clinical patient information and detailed genomic sequences to microbiome insights and epidemiological trends[43,44]. Significant successes have been achieved by prior ML approaches in predicting AMR. Notably, ML algorithms such as XGBoost have predicted carbapenem-resistant *Klebsiella pneumoniae* infections in intensive care units with an average accuracy of 83%[45]. Other ML models have also shown competitive performance in predicting AMR in *Acinetobacter baumannii* directly from *k*-mer features extracted from WGS data, achieving essential agreement and category agreement exceeding 90% and 95% respectively[46].

Despite these compelling successes, prior ML approaches for AMR prediction face several significant limitations. A primary concern is data availability and specificity. Many models are developed using exclusive data from a single hospital population, thereby lacking diverse, representative training data. More often than not, single-species models perform poorly even on closely related species or fail to generalize across different drug classes. This necessitates rigorous external and prospective validation to ensure their adaptability and generalizability to other clinical settings. Moreover, current AMR prediction models are often constrained by their reliance on predefined resistance genes, limited feature types (e.g., only SNPs or ARGs), or narrow

contexts, stalling the discovery of novel AMR mechanisms and posing risks of misclassification when using rule-based methods to assign AMR in resistant populations that do not carry the majority AMR genotype. Our approach addresses this by integrating multiscale molecular features (genes, proteins, domains) derived from bacterial pangenomes and captures both canonical and noncanonical resistance signals. By moving beyond curated ARG databases and incorporating signals from mobile elements, hypothetical proteins, and structural genomic motifs, this work expands the feature space and allows the identification of emergent resistance determinants that would be missed by rule-based or single-scale methods. Our workflow is generalizable across available bacterial data and permits granular AMR insight and discovery.

Limitations and considerations: Although our models deliver strong, insightful results across pathogens, drugs, and molecular scales, we must consider some inherent limitations. First, our metadata are skewed towards a handful of high-resource countries, limited context studies such as sampling efforts within a single hospital system or outbreak, and primarily human hosts, leaving a substantial portion of the AMR story untold. **Figure 6A** demonstrates how models trained on isolates from one country, then tested on isolates from other countries can yield dramatic performance shortfalls. These stark differences are essential to consider given the disparity in public data availability with broad geographic distribution: regions like South Asia, sub-Saharan Africa, and parts of Latin America exhibit some of the highest AMR-related mortality globally, as reported by a 2024 systematic analysis[4], but often have no available genomic data with paired AMR phenotypes. A smaller but notable subset of isolates was sourced from animals, including pigs, dogs, and poultry. A One Health-aware surveillance framework that incorporates such data is necessary, especially considering the zoonotic potential of resistant *S. aureus*[47] and *P. aeruginosa*[48,49] strains. Further, given the even greater volume of antibiotics used in livestock than in humans, often for non-medical purposes[50], future analysis of AMR patterns requires much more extensive sequencing and phenotyping of isolates from animal hosts.

Our feature scales do not capture single nucleotide-level variations. Such variants are critical to certain types of AMR, and we must caution that our models are not designed to identify every possible form of AMR that may arise. However, even for drug resistance profiles that are dominated by SNPs such as in fluoroquinolone drugs (where resistance is typically caused by target site mutations[51]), we still observe some highly accurate models with predictive performance well above baseline (**Figure S5**). This suggests that even without incorporating SNP-level information, we can observe higher genomic level changes that are predictive of these underlying mutations and their resulting phenotypes.

Scarcity of phenotypic data and severe class imbalances for select antibiotics (and classes) in some bacterial backgrounds also complicated model predictions. In such cases, performance fluctuated between unreasonably strong (nMCC = 1) and baseline performance (nMCC = 0.5), with many combinations lacking sufficient data for ML modeling in the first place. This harmed our ability to make meaningful comparisons across molecular scales. Cross-species consideration is important but was challenging in our analyses due to inconsistent cluster annotations between pangenomes. As a work-around, we have added COG annotations for cross-species comparisons. Given the scarcity of a hypothetical pan-species pan-feature matrix and the computational burden that would be involved in constructing it, innovative methods are needed to bridge the cross-species gap in the future. The vast space of hypothetical proteins with no functional assignment represented an additional limitation for the depth and breadth of our analyses. Finally, future methods are warranted for training models to make predictions on newly sequenced genomes without the need to generate entire pan-feature matrices afresh.

Impacts: All considered, this work demonstrates the utility of interpretable, multiscale ML models in uncovering novel resistance-associated features. It opens the door to mechanistically informed AMR prediction, where domain and protein-level insights point toward candidate resistance genes and genomic contexts. This framework supports a new paradigm for AMR research: one in which ML is not merely a black

box classifier but a hypothesis-generating engine that directs experimental validation and surveillance strategies. We bolster this approach through our interactive data dashboard, allowing others to accessibly browse and interpret data from hundreds of models and thousands of isolates. In the future, we plan to integrate transcriptomic or proteomic data to refine predictions and use transfer learning across species to identify conserved or convergent resistance mechanisms. Transcriptomic or proteomic data could be incorporated to capture condition-specific resistance responses and post-transcriptional regulation, which are often missed by genome-only approaches. Furthermore, the use of transfer learning may reveal conserved or convergent resistance mechanisms across phylogenetic distances. This is especially relevant for identifying pan-species biomarkers of resistance and for preempting the spread of resistance traits in newly emerging threats. Finally, coupling this computational pipeline with real-time genomic surveillance efforts such as those run by public health laboratories could provide actionable insights into emerging resistance patterns. As more high-resolution phenotype-genotype datasets become available, we anticipate a virtuous cycle where ML-derived hypotheses drive experimental discovery, which in turn, enhances the predictive power and interpretability of future models.

Conclusion

Our multiscale machine learning framework offers a powerful and interpretable solution to one of the most urgent challenges in modern medicine: the rapid and reliable prediction of AMR. Using publicly available data to construct pangenomes for various bacterial pathogens and applying established ML methods, we present a robust set of models that, using only resistant/susceptible labels for individual genomes, predict genes associated with drug resistance *de novo*. We share our data, metadata, and results through an interactive data dashboard available at JRaviLab.org/amr. nMCC values support the use of these models in correctly predicting susceptibility to different antibiotic classes from complete genome and WGS data. Many of our top feature hits have plentiful literature support (e.g., *S. aureus mecA* for methicillin resistance, *ermC* for macrolide resistance), while others are associated with pathogenicity islands and other mobile genetic elements that include resistance mechanisms (e.g., *S. aureus ugpQ* in the SCC*mec* locus). We also identify many features that are potentially novel AMR contributors. As genomic data continues to expand, our framework serves as a timely and adaptable tool for improving AMR surveillance, guiding therapeutic decisions, and accelerating the development of new diagnostics and interventions.

Figures

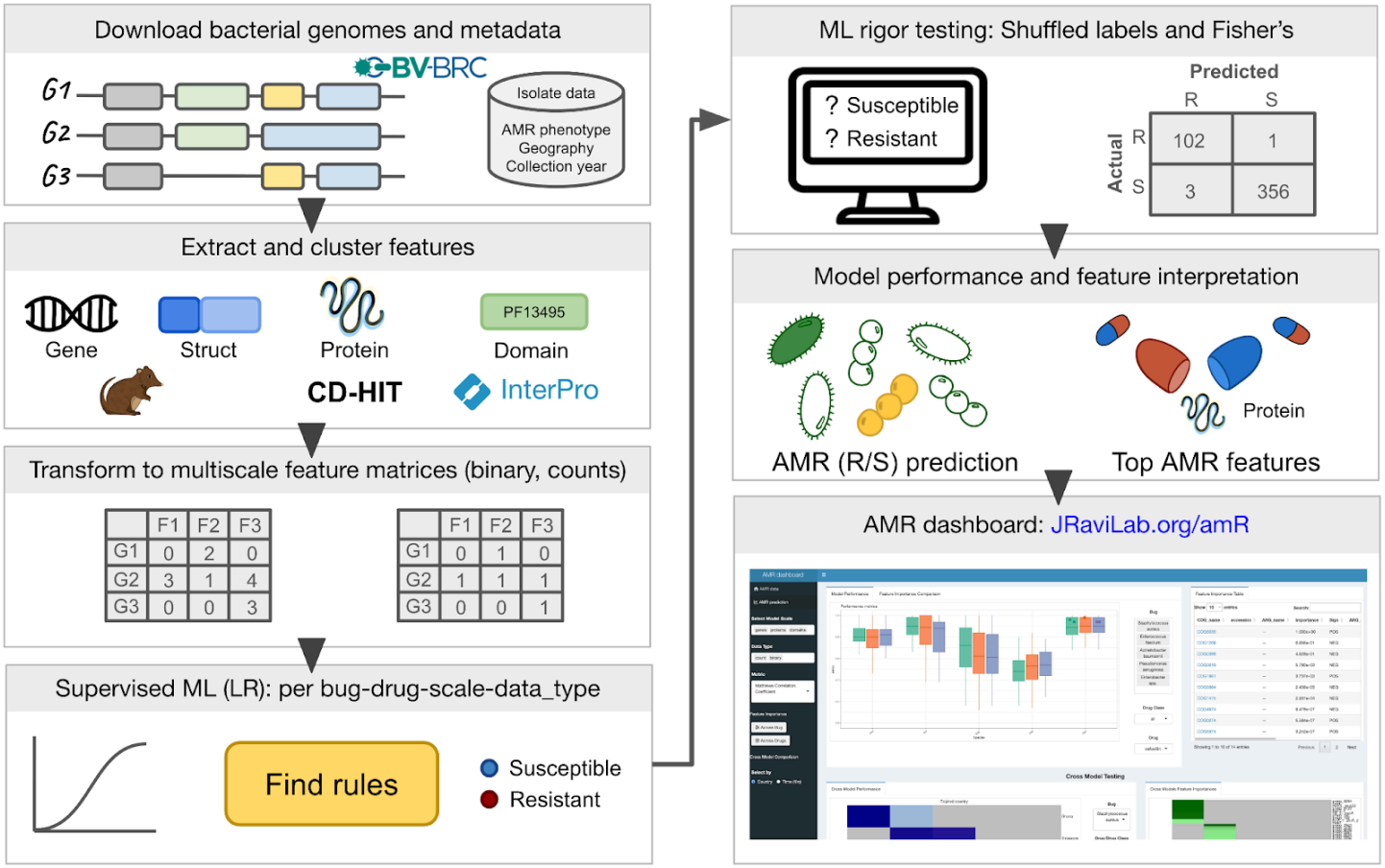


Figure 1. Overview of AMR prediction workflow. We developed a bacterial species-agnostic workflow using public data and supervised machine learning (ML) methods to predict bacterial isolate antimicrobial resistance (AMR) phenotypes against specific drugs. In brief, the pipeline starts with 1. using the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) to acquire genomic sequences (.fna) and annotated protein sequences (.faa) with paired AMR phenotype labels and metadata from chosen bacterial species. 2. Features are clustered using Panaroo (genes and gene structure) and CD-HIT (proteins), and domains extracted from the proteins using InterPro. 3. Per genome feature presence/absence and count data type matrices are created for ML input. 4. Supervised logistic regression (LR) ML models are trained per bug/drug/feature scale/data type on drug resistance/susceptibility labels. 5. As a baseline, labels are shuffled per matrix and ML models trained to create a control without biological signals, and non-ML analysis is performed by Fisher's exact testing of features. 6. Model performance is compared across species, drugs, molecular scales, and data types and the top predictive molecular features are analyzed for their relevance to AMR. 7. We share our results in an interactive data dashboard for exploration, visualization, and hypothesis generation.

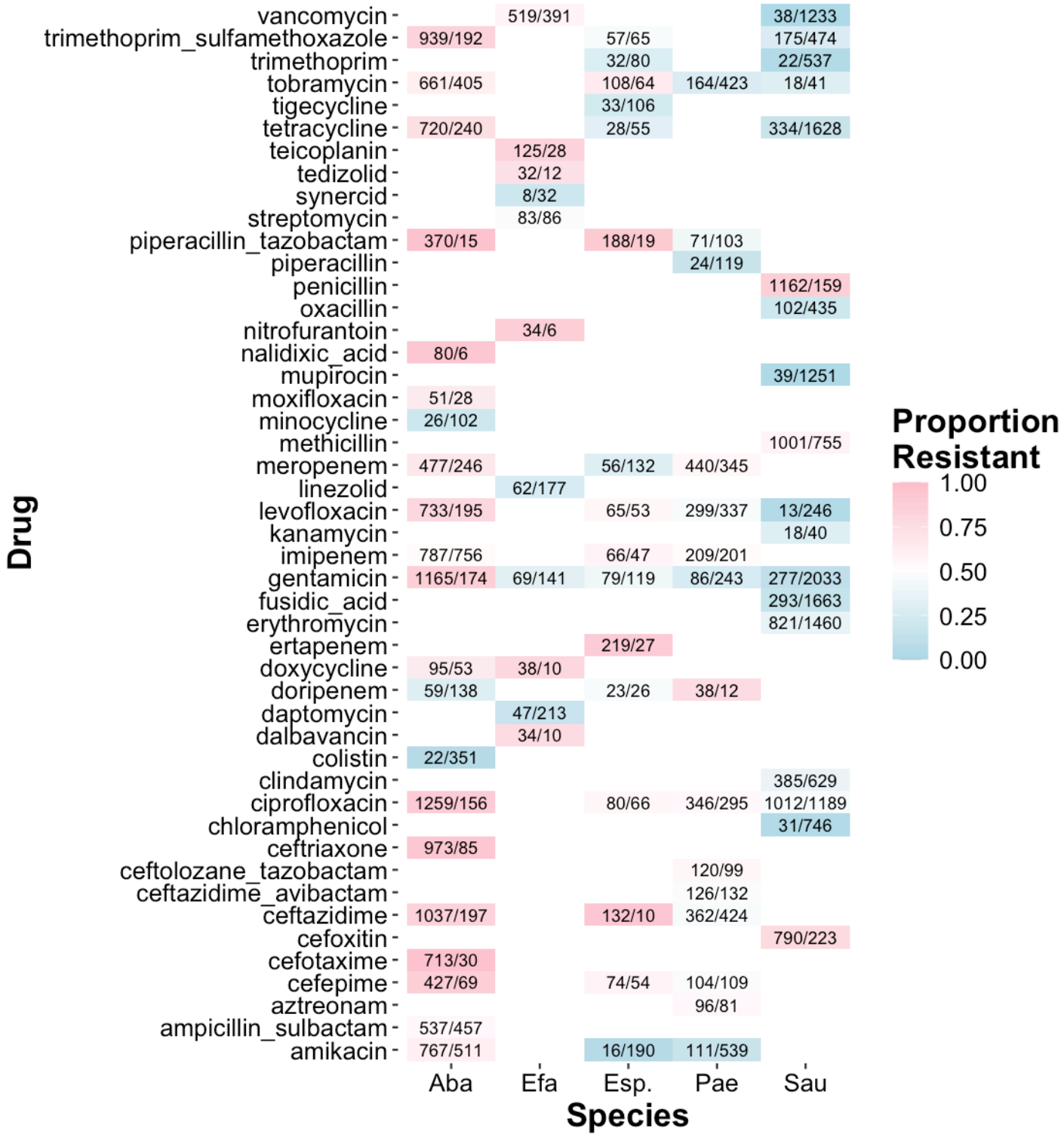


Figure 2. Data availability and AMR prevalence across pathogens. Fractions displayed in the tiles report resistant (numerator) over susceptible (denominator) isolates for a given drug with available whole genome sequences from BV-BRC; redder tiles indicate more frequently identified resistance, while bluer tiles indicate susceptibility. Only combinations with sufficient samples and class balance for our initial ML filtering are reported here (e.g., species-drug combinations with fewer than 40 samples may be available through BV-BRC, but are not shown). *A. baumannii* presents the most frequent overall resistance across the greater number of drugs. Gentamicin is the only drug in the set with data across all tested species.

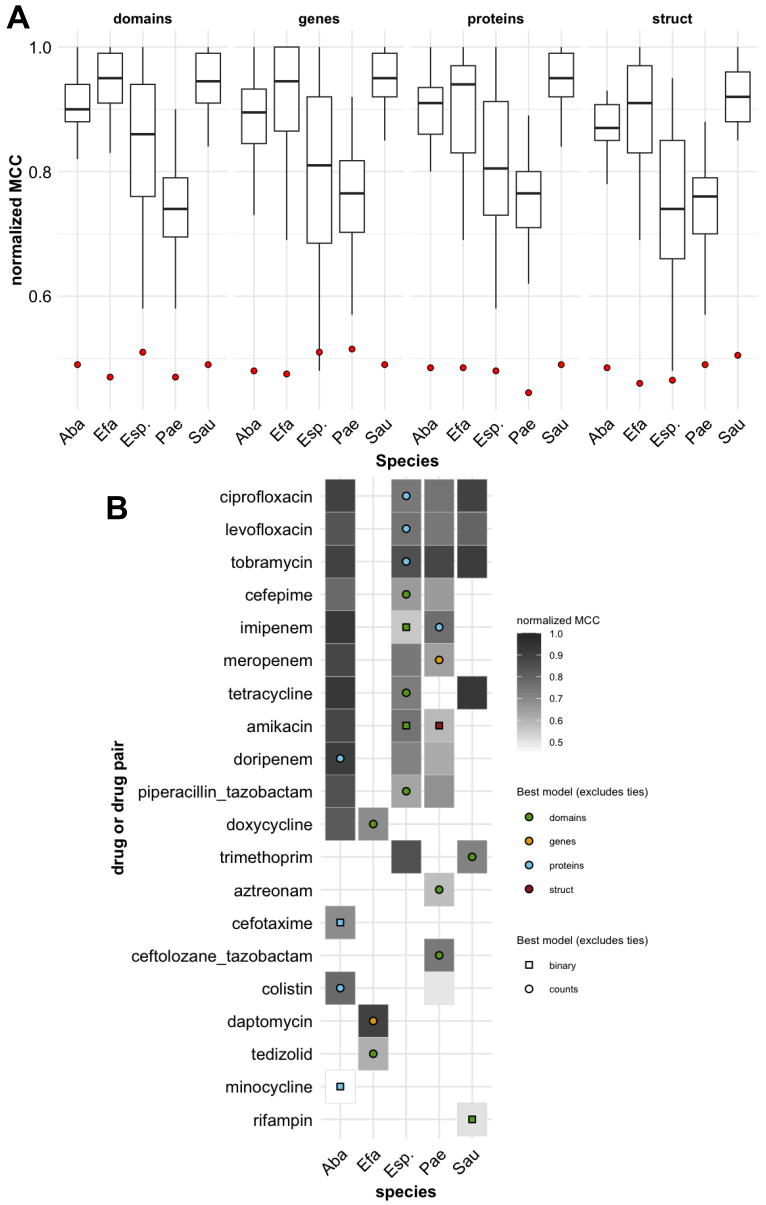


Figure 3. Summarized model performance across species and drugs. A. Model performances against label-shuffled baselines. Each box in the plot represents the performance range of all drug and drug class models based on different feature types of a molecular scale (gene/protein/domain/struct) for one species. The red point is the median nMCC for the same combination of models but where AMR phenotype labels were randomly shuffled as a control to eliminate detectable biological signal between features and labels. These label-shuffled models represent random performance (nMCC ~ 0.5) and emphasize the prediction capabilities of the true models. **B. Best performing feature scales across tested drug classes per species.** Grayscale colored tiles represent normalized Matthews correlation coefficients for models, where black (1.0) is perfect predictive performance, and white (0.5) is random performance. Colored squares (feature presence/absence) or squares (feature counts) inside the tiles represent cases where one feature scale outperformed others. Across modeled species, aminoglycosides showed the highest prediction metrics. Feature scales were comparable in performance, and overall model performance high across most drug classes and species.

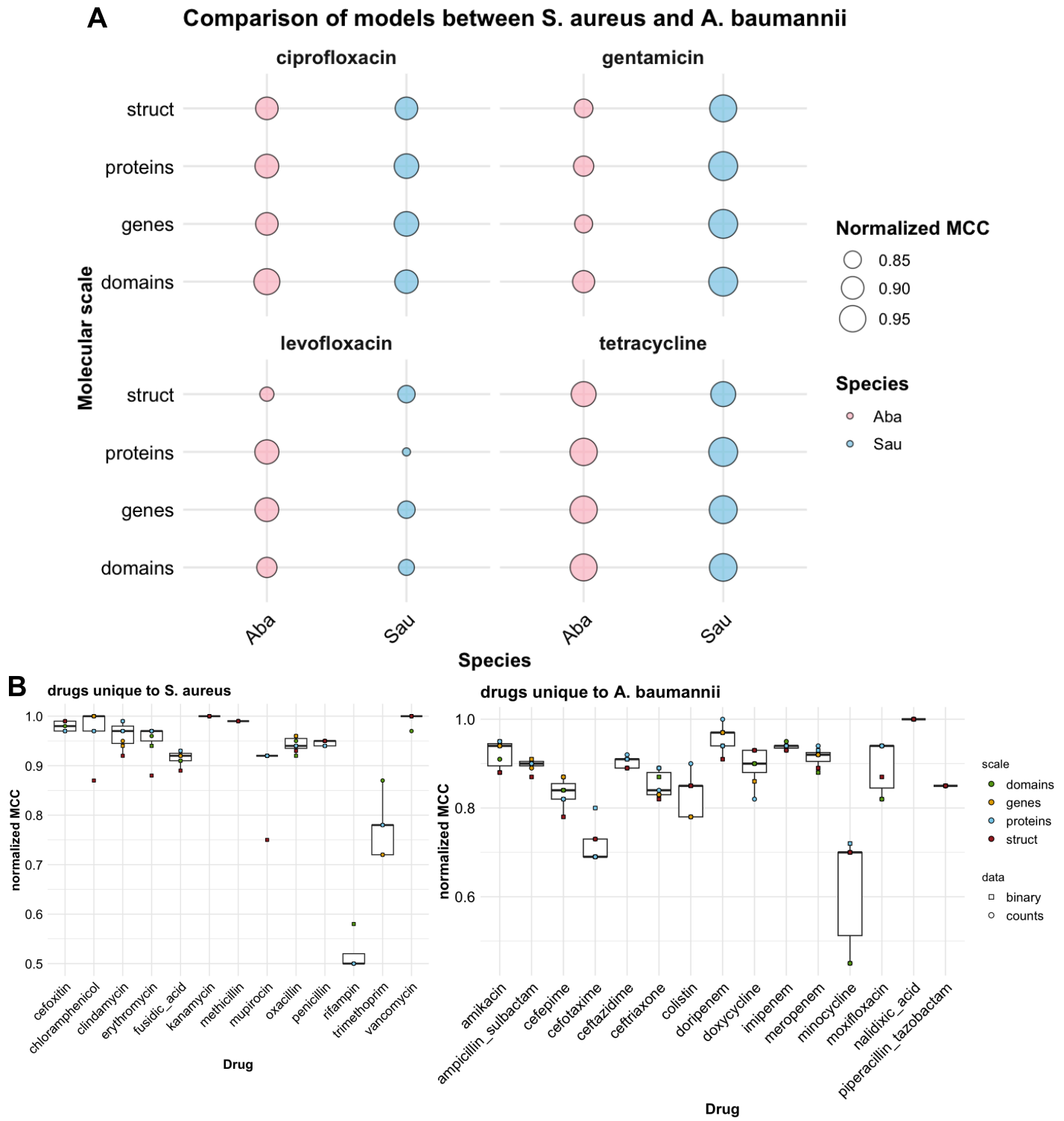
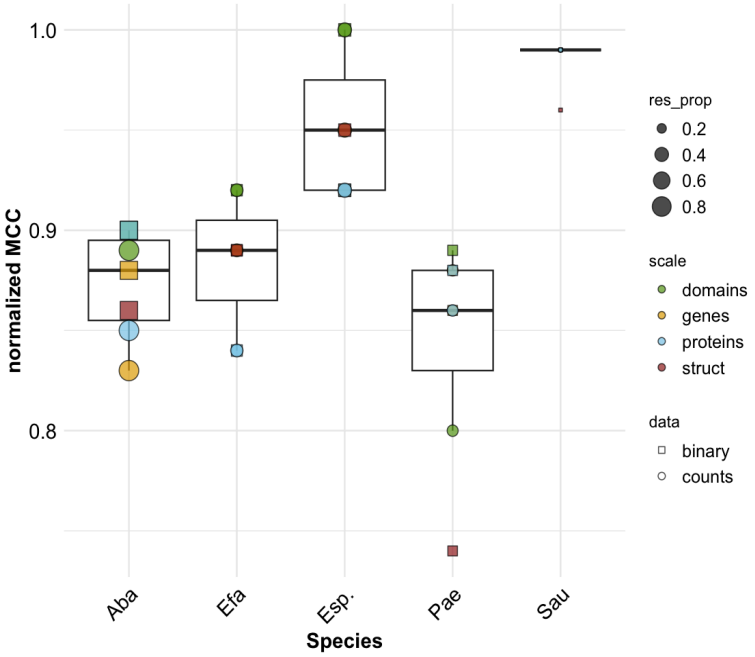


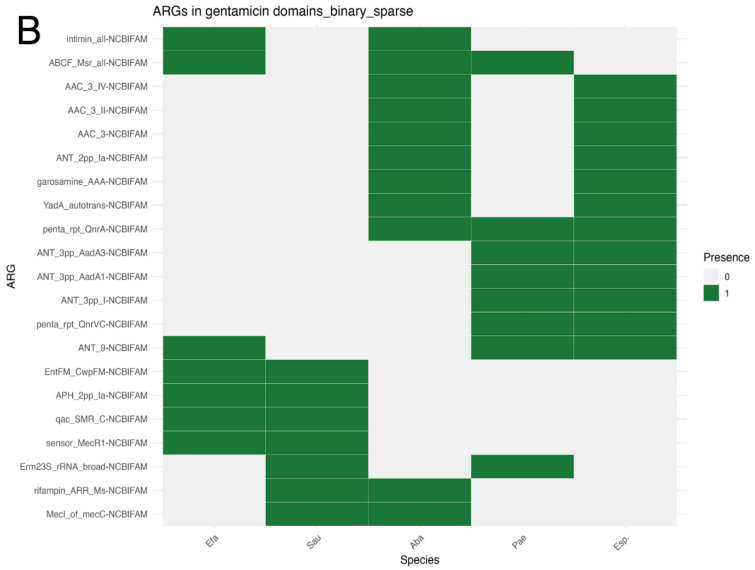
Figure 4. Model performance case study for drugs used against *S. aureus* and *A. baumannii*. A: Comparison of model performance for drugs used in both *S. aureus* and *A. baumannii*. Points indicate model nMCC values per drug (facets) for *S. aureus* (blue) and *A. baumannii* (pink). For drugs used in both species, performance metrics are comparable, and no single feature scale dominates in prediction. Presence/absence and counts models are averaged for this plot, but are also comparable. **B: Per-drug model performance for antibiotics not shared between the two species**. *S. aureus* (left) performance is robust by nMCC values for nearly all models, with random baseline performance for rifampin, and variable but still good

performance for trimethoprim. Rifampin-resistant *S. aureus* occurs by SNPs in *rpoB*[52], while trimethoprim resistance can occur either by dihydrofolate reductase mutations, or acquisition of variant dihydrofolate reductase genes through horizontal gene transfer[53]. *A. baumannii* (right) performance shows more variation in performance, including baseline performance for minocycline resistance by domain models. The dominant mechanism for minocycline efflux in *A. baumannii* is through TetB, which contains PF07690, a sprawling secondary transporter superfamily across domains of life[54], likely obscuring signal to domain models and explaining why other feature scales perform dramatically better for this one drug. Our models generally perform well, with exceptions for small numbers of observations, extreme class imbalance, and AMR mediated primarily by SNPs that are not captured by our molecular feature scales.

A


Gentamicin performances across species
normalized MCC
1.0
0.9
0.8
Aba
Efa
Esp.
Pae
Sau
Species
res_prop
0.2
0.4
0.6
0.8
scale
domains
genes
proteins
struct
data
binary
counts

B


ARGs in gentamicin domains_binary_sparse
ARG
intimin_all-NCBIFAM
ABCF_Msr_all-NCBIFAM
AAC_3_IV-NCBIFAM
AAC_3_II-NCBIFAM
AAC_3-NCBIFAM
ANT_2pp_Ia-NCBIFAM
garosamine_AAA-NCBIFAM
YadA_autotrans-NCBIFAM
penta_rpt_QnrA-NCBIFAM
ANT_3pp_AadA3-NCBIFAM
ANT_3pp_AadA1-NCBIFAM
ANT_3pp_I-NCBIFAM
penta_rpt_QnrVC-NCBIFAM
ANT_9-NCBIFAM
EntFM_CwpFM-NCBIFAM
APH_2pp_Ia-NCBIFAM
qac_SMR_C-NCBIFAM
sensor_MecR1-NCBIFAM
Erm23S_rRNA_broad-NCBIFAM
rifampin_ARR_Ms-NCBIFAM
MecI_of_mecC-NCBIFAM
Efa
Sau
Aba
Pae
Esp.
Species
Presence
0
1

C

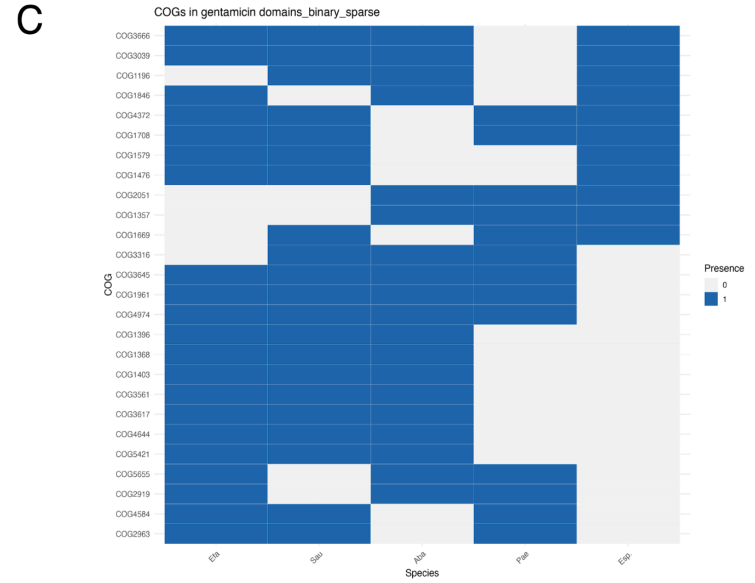

COGs in gentamicin domains_binary_sparse
COG
COG3666
COG3039
COG1196
COG1846
COG4372
COG1708
COG1579
COG1476
COG2051
COG1357
COG1669
COG3316
COG3645
COG1961
COG4974
COG1396
COG1368
COG1403
COG3561
COG3617
COG4644
COG5421
COG5655
COG2919
COG4584
COG2963
Efa
Sau
Aba
Pae
Esp.
Species
Presence
0
1

Figure 5. Identifying common mechanisms of resistance across species for a single drug. A. Gentamicin model performance across species. The aminoglycoside gentamicin is used across the entire set of species analyzed herein, yielding an opportunity to compare the performance of the workflow for the same antibiotic in phylogenetically distinct taxa. The size of each colored shape represents the proportion of resistant isolates (larger = more resistance), colors represent feature scale, and squares (presence/absence) vs. circle (counts) represent data type. Overall model performance is high, with only the struct model in *P. aeruginosa* falling below nMCC = 0.80. **B. Common features mapping to known ARG HMMs across species in gentamicin models.** We mapped top features in each species' gentamicin resistance models using HMMER[36] with databases built from NCBI antimicrobial resistance gene (ARG) family HMMs. Numerous known aminoglycoside-associated ARGs were observed across species, with an apparent divide between Gram-positive *E. faecium* and *S. aureus* and the Gram-negative *A. baumannii*, *P. aeruginosa*, and *Enterobacter* spp.; Gram-negative pathogens showed a diverse repertoire of aminoglycoside ARGs, likely reflecting gentamicin's primary use against these pathogens and the subsequent evolutionary pressure towards these varied mechanisms. **C. Common features mapping to known COGs across species in gentamicin models.** We also mapped top features to clusters of orthologous genes (COGs) HMMs[38] as another means of comparing effectively across species. When analyzing COGs mapped to top domains, a majority were associated with mobile genetic elements (e.g., plasmid replication initiation proteins, prophage elements, transposases, etc.), underscoring the critical role of horizontal gene transfer events in the exchange of aminoglycoside modifying enzymes in particular and AMR in general.

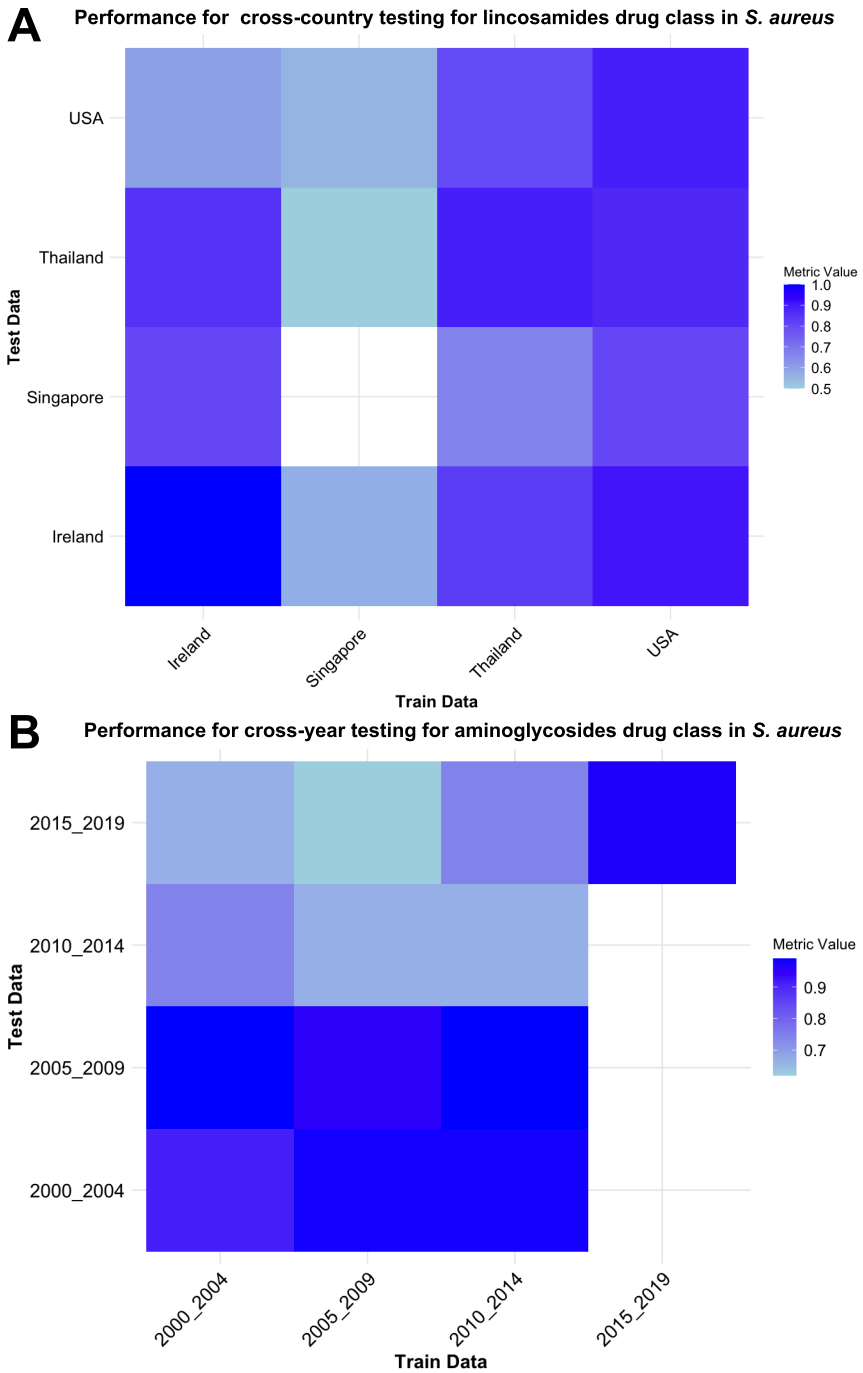


Figure 6. Evaluating effects of metadata-based holdouts on model performance. A. Geographic holdout performance for *S. aureus* lincosamide drug class models. Others have established that AMR mechanisms and prevalence meaningfully vary based on their geographic origins. To assess how our predictive performance is influenced by country of origin, we demonstrate an example in *S. aureus* and the lincosamide drug class. We all-vs.-all stratified our training and testing data based on their metadata for country of origin, trained models on one country, and tested them against data from other countries. Models trained on US data perform well on Irish, Thai, and Singaporean data, but other comparisons yielded distinct performance deficits. While Irish training data performed similarly in Singapore and Thailand as US training data, the Irish model showed a sharp drop when tested on US data. Limited Singaporean data likely contributed to poor overall performance against other nations' data. While geographic patterns of AMR do exist

and are important to consider, these differences may also be explained by sampling bias – genomic surveillance paired with antimicrobial susceptibility testing data is more widespread in the US, allowing more diverse mechanisms of resistance to be captured in a larger dataset. Distinguishing truly unique regional patterns of AMR will require greater emphasis on sampling, sequencing, and data sharing through global public health collaborations**. B. Temporal holdout testing for aminoglycoside drug class models in *S. aureus*.** As in A, we sought to discern the effects of sampling date on predictive performance and performed metadata-based holdouts of training and testing data. To demonstrate, we binned 4 intervals from the year 2000 until 2020, trained models on single bins, and tested them against other bins and show models using pre-2010 training data have deficits in performance against post-2010 testing data.

References

1. Michael, C. A., Dominey-Howes, D. & Labbate, M. The Antimicrobial Resistance Crisis: Causes, Consequences, and Management. *Front. Public Health* **2**, (2014).
2. Poku, E., Cooper, K., Cantrell, A., Harnan, S., Sin, M. A., Zanuzdana, A., *et al.* Systematic review of time lag between antibiotic use and rise of resistant pathogens among hospitalized adults in Europe. *JAC-Antimicrob. Resist.* **5**, dlad001 (2023).
3. Ho, C. S., Wong, C. T. H., Aung, T. T., Lakshminarayanan, R., Mehta, J. S., Rauz, S., *et al.* Antimicrobial resistance: a concise update. *Lancet Microbe* **6**, (2025).
4. Murray, C. J. L., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629–655 (2022).
5. Review on Antimicrobial Resistance. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations. (2014).
6. Kaur, J. N., Singh, N., Smith, N. M., Klem, J. F., Cha, R., Lang, Y., *et al.* Next generation antibiotic combinations to combat pan-drug resistant Klebsiella pneumoniae. *Sci. Rep.* **14**, 3148 (2024).
7. Stockholm: European Centre for Disease Prevention and Control and World Health Organization. Antimicrobial resistance surveillance in Europe 2023 - 2021 data. (2023).
8. Li, W., Yang, Z., Hu, J., Wang, B., Rong, H., Li, Z., *et al.* Evaluation of culturable 'last-resort' antibiotic resistant pathogens in hospital wastewater and implications on the risks of nosocomial antimicrobial resistance prevalence. *J. Hazard. Mater.* **438**, 129477 (2022).
9. The antibiotic resistance crisis and the development of new antibiotics - Brüssow - 2024 - Microbial Biotechnology - Wiley Online Library. https://enviromicro-journals.onlinelibrary.wiley.com/doi/full/10.1111/1751-7915.14510.
10. Cook, M. A. & Wright, G. D. The past, present, and future of antibiotics. *Sci. Transl. Med.* **14**, eabo7793 (2022).
11. Sati, H., Carrara, E., Savoldi, A., Hansen, P., Garlasco, J., Campagnaro, E., *et al.* The WHO Bacterial Priority Pathogens List 2024: a prioritisation study to guide research, development, and public health strategies against antimicrobial resistance. *Lancet Infect. Dis.* **0**, (2025).
12. Mancuso, G., Midiri, A., Gerace, E. & Biondo, C. Bacterial Antibiotic Resistance: The Most Critical Pathogens. *Pathogens* **10**, 1310 (2021).
13. Vasala, A., Hytönen, V. P. & Laitinen, O. H. Modern Tools for Rapid Diagnostics of Antimicrobial Resistance. *Front. Cell. Infect. Microbiol.* **10**, 308 (2020).
14. Ventola, C. L. The antibiotic resistance crisis: part 1: causes and threats. *P T Peer-Rev. J. Formul. Manag.* **40**, 277–283 (2015).
15. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* **13**, 42–51 (2015).
16. Lin, X., Poeta, P. & Peng, B. Editorial: The Molecular Mechanisms of Antibiotic Resistance in Aquatic Pathogens. *Front. Cell. Infect. Microbiol.* **10**, 586460 (2020).
17. Munita, J. M. & Arias, C. A. Mechanisms of Antibiotic Resistance. *Microbiol. Spectr.* **4**, 4.2.15 (2016).
18. Anahtar, M. N., Yang, J. H. & Kanjilal, S. Applications of Machine Learning to the Problem of

Antimicrobial Resistance: an Emerging Model for Translational Research. *J. Clin. Microbiol.* **59**, e01260-20 (2021).
19. Ghosh, A., Vang, C. K., Brenner, E. P. & Ravi, J. Unlocking antimicrobial resistance with multiomics and machine learning. *Trends Microbiol.* S0966842X25001465 (2025) doi:10.1016/j.tim.2025.04.017.
20. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* **49**, 107739 (2021).
21. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P. & Zhang, L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).
22. Zhang, X., Hu, Y., Cheng, Z. & Archibald, J. M. AMRLearn: Protocol for a machine learning pipeline for characterization of antimicrobial resistance determinants in microbial genomic data. *STAR Protoc.* **6**, 103733 (2025).
23. Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–D699 (2023).
24. Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **75**, 3491–3500 (2020).
25. Olson, R. D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, J. J., *et al.* Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* **51**, D678–D689 (2023).
26. *WHO Bacterial Priority Pathogens List 2024: Bacterial Pathogens of Public Health Importance, to Guide Research, Development, and Strategies to Prevent and Control Antimicrobial Resistance*. (World Health Organization, Geneva, 2024).
27. Raasveldt, M. & Mühleisen, H. DuckDB: an Embeddable Analytical Database. in *Proceedings of the 2019 International Conference on Management of Data* 1981–1984 (ACM, Amsterdam Netherlands, 2019). doi:10.1145/3299869.3320212.
28. *WHO AWaRe (Access, Watch, Reserve) Classification of Antibiotics for Evaluation and Monitoring of Use.* (2023).
29. Farrell, L. J., Lo, R., Wanford, J. J., Jenkins, A., Maxwell, A. & Piddock, L. J. V. Revitalizing the drug pipeline: AntibioticDB, an open access database to aid antibacterial research and development. *J. Antimicrob. Chemother.* **73**, 2284–2297 (2018).
30. Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
31. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma. Oxf. Engl.* **28**, 3150–3152 (2012).
32. Blum, M., Andreeva, A., Florentino, L. C., Chuguransky, S. R., Grego, T., Hobbs, E., *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res.* **53**, D444–D456 (2025).
33. Kuhn, M. & Wickham, H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. (2020).
34. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
35. Greenwell, B. M. & Boehmke, B. C. Variable Importance Plots—An Introduction to the vip Package. *R J.* **12**, 343–366 (2020).
36. Sean R., E. HMMER: biosequence analysis using profile hidden Markov models. (2023).
37. Reference HMM Catalog - Pathogen Detection - NCBI. https://www.ncbi.nlm.nih.gov/pathogens/hmm/.
38. Dibrova, D. V., Konovalov, K. A., Perekhvatov, V. V., Skulachev, K. V. & Mulkidjanian, A. Y. COGcollator: a web server for analysis of distant relationships between homologous protein families. *Biol. Direct* **12**, 29 (2017).

39. Breitwieser, F. P. & Salzberg, S. L. Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics* **36**, 1303–1304 (2020).
40. Schwarz, S., Shen, J., Wendlandt, S., Feßler, A. T., Wang, Y., Kadlec, K., *et al.* Plasmid-Mediated Antimicrobial Resistance in Staphylococci and Other *Firmicutes*. *Microbiol. Spectr.* **2**, 2.6.13 (2014).
41. Favrot, L., Blanchard, J. S. & Vergnolle, O. Bacterial GCN5-Related *N* -Acetyltransferases: From Resistance to Regulation. *Biochemistry* **55**, 989–1002 (2016).
42. Ramirez, M. S. & Tolmasky, M. E. Aminoglycoside modifying enzymes. *Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother.* **13**, 151–171 (2010).
43. Karnwal, A., Jassim, A. Y., Mohammed, A. A., Al-Tawaha, A. R. M. S., Selvaraj, M. & Malik, T. Addressing the global challenge of bacterial drug resistance: insights, strategies, and future directions. *Front. Microbiol.* **16**, (2025).
44. Sherry, N. L., Lee, J. Y. H., Giulieri, S. G., Connor, C. H., Horan, K., Lacey, J. A., *et al.* Genomics for antimicrobial resistance—progress and future directions. *Antimicrob. Agents Chemother.* **69**, e01082-24 (2025).
45. Alparslan, V., Güler, Ö., İnner, B., Düzgün, A., Baykara, N. & Kuş, A. A novel approach to antimicrobial resistance: Machine learning predictions for carbapenem-resistant Klebsiella in intensive care units. *Int. J. Med. Inf.* **195**, 105751 (2025).
46. Gao, Y., Li, H., Zhao, C., Li, S., Yin, G. & Wang, H. Machine learning and feature extraction for rapid antimicrobial resistance prediction of Acinetobacter baumannii from whole-genome sequencing data. *Front. Microbiol.* **14**, (2024).
47. Haag, A. F., Fitzgerald, J. R. & Penadés, J. R. *Staphylococcus aureus* in Animals. *Microbiol. Spectr.* **7**, 7.3.11 (2019).
48. EFSA Panel on Animal Health and Welfare (AHAW), Nielsen, S. S., Bicout, D. J., Calistri, P., Canali, E., Drewe, J. A., *et al.* Assessment of listing and categorisation of animal diseases within the framework of the Animal Health Law (Regulation (EU) No 2016/429): antimicrobial-resistant Pseudomonas aeruginosa in dogs and cats. *EFSA J.* **20**, (2022).
49. Abd El-Ghany, W. A. Pseudomonas aeruginosa infection of avian origin: Zoonosis and one health implications. *Vet. World* 2155–2159 (2021) doi:10.14202/vetworld.2021.2155-2159.
50. Antibiotics: useless medicines | Heinrich Böll Stiftung | Brussels office - European Union. https://eu.boell.org/en/2021/09/07/antibiotics-useless-medicines.
51. Redgrave, L. S., Sutton, S. B., Webber, M. A. & Piddock, L. J. V. Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. *Trends Microbiol.* **22**, 438–445 (2014).
52. Zhou, W., Shan, W., Ma, X., Chang, W., Zhou, X., Lu, H., *et al.* Molecular characterization of rifampicin-resistant Staphylococcus aureus isolates in a Chinese teaching hospital from Anhui, China. *BMC Microbiol.* **12**, 240 (2012).
53. Nurjadi, D., Olalekan, A. O., Layer, F., Shittu, A. O., Alabi, A., Ghebremedhin, B., *et al.* Emergence of trimethoprim resistance gene dfrG in Staphylococcus aureus causing human infection and colonization in sub-Saharan Africa and its import to Europe. *J. Antimicrob. Chemother.* **69**, 2361–2368 (2014).
54. Wang, S. C., Davejan, P., Hendargo, K. J., Javadi-Razaz, I., Chou, A., Yee, D. C., *et al.* Expansion of the Major Facilitator Superfamily (MFS) to include novel transporters as well as transmembrane-acting enzymes. *Biochim. Biophys. Acta Biomembr.* **1862**, 183277 (2020).