Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond the Black Box: A Calibrated, Gene-Centric Pipeline for High-Precision Ciprofloxacin Resistance Prediction in Salmonella enterica

Hevar N. Barznji

hevar.ness7@gmail.com

Komar University of Science and Technology https://orcid.org/0009-0001-3783-5998

Research Article

Keywords: Salmonella enterica, Antimicrobial Resistance (AMR), Machine Learning, Genomic Epidemiology, Model Calibration, SHAP(SHapley Additive exPlanations), Phylogenetic Leakage, Ciprofloxacin

Posted Date: March 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9230875/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Beyond the Black Box: A Calibrated, Gene-Centric Pipeline for High-Precision Ciprofloxacin Resistance Prediction in *Salmonella enterica*

Hevar N. Barznji[1]

[1]MLS Department, Komar university of science and technology, Sulaymaniyah, Iraq

Corresponding Author: Hevar N. Barznji

Email: hevar.ness7@gmail.com – **Phone**:+9647701994221

Abstract

The transition from traditional methods to genomic-based antimicrobial susceptibility testing (AST) for *Salmonella* requires models that are biologically interpretable, clinically calibrated, and discriminative. This study analyzed 8,759 *Salmonella* isolates from the NCBI Pathogen Detection database, to predict ciprofloxacin resistance through two distinct computational architectures: an alignment-free, agnostic k-mer hashing model and a feature-engineered gene-based model (Logistic Regression, Random Forest, and XGBoost). Methodological auditing revealed a theoretical collision probability of 100% when 10 million unique k-mers are mapped into a 16.7 million (2^{24}) feature space via MurmurHash3, leading to a complete loss of biological signal. In comparison, the gene-based models demonstrated superior performance, with XGBoost achieving a Receiver Operating Characteristic Area Under the Curve (ROC AUC) of 0.989. Furthermore, to ensure clinical reliability, post-hoc Isotonic Regression was applied, refining probability estimates to an average Brier Score of 0.0063, significantly outperforming current clinical benchmarks. Explainability analysis using SHAP identified the *gyrA_D87Y* mutation as the primary indicator of resistance with an importance of 0.88, while an inverse relationship with *aph(3")-Ib* suggested potential collateral sensitivity trade-offs. However, LOSO validation revealed a significant performance decay in clonal, human-restricted lineages such as *S. Typhi*, showcasing Phylogenetic Leakage as the primary barrier to universal generalization. These findings demonstrate that while gene-based models provide high-fidelity AMR prediction, future frameworks must integrate efflux regulation and lineage-robust audits to move beyond static genomic anchors toward real-time, personalized antimicrobial stewardship.

Keywords: *Salmonella enterica*, Antimicrobial Resistance (AMR), Machine Learning, Genomic Epidemiology, Model Calibration, SHAP(SHapley Additive exPlanations), Phylogenetic Leakage, Ciprofloxacin

1 Introduction

The global threat of antimicrobial resistance (AMR) presents a significant challenge for modern medicine, and for the management of infectious diseases (WHO, 2025). *Salmonella* species particularly exemplify this threat; ciprofloxacin, historically served as the cornerstone therapy for managing invasive salmonellosis, but as of recently is losing efficacy (WHO, 2025). Recent surveillance data from Global Antimicrobial Resistance and Use Surveillance System (GLASS) developed by World Health Organization (WHO) indicates that in 2023, approximately one in six laboratory confirmed pathogens exhibited resistance, with 40% increase of resistance among the monitored pathogens between 2018 and 2023 (WHO, 2025). In the United States alone the percentage of non-susceptible *Salmonella* had increased by 10% in 2017 (CDC, 2025). This rapid increase in resistance is driven by the expansion of multi-drug resistant lineages, such as *S. Infantis*, which was responsible for 25% of all *Infantis* infections in the US by 2018 (CDC, 2025). Another example is *S. Dublin* where 50% of the clinical isolates of this strain now show resistance to seven or more classes of antibiotics (CDC, 2024).

The clinical efficacy of Ciprofloxacin depends on the drug's ability to target intracellular pathogens, but due to the rise of strains with reduced fluoroquinolone susceptibility (RFS) its clinical impact has been significantly reduced (Slinger et al., 2004). These RFS strains are categorized by high resistance rate ranging between 0.12 and 0.5 µg/mL, and delayed signs of improvement often with persistent fever lasting anywhere from 9 to 10 days despite clinical interventions (Slinger et al., 2004). Consequently, the Clinical and Laboratory Standard Institute has revised the clinical breakpoint of ciprofloxacin and now a strain with MIC value ≥ 0.12 µg/mL is considered resistant (Angelo et al., 2016). The primary drivers of this level of resistance are chromosomal mutations in the Quinolone Resistance Determining Region (QRDR) of DNA-gyrase and topoisomerase IV (Shariati et al., 2022). It has been shown that with sequential mutations in *gyrA* at codons 83 and 87 (e.g., S83Y, D87Y) serve as the primary indicators of resistance (Ferrari et al., 2013). The foundation of these chromosomal mutations are often augmented by Plasmid-Mediated Quinolone Resistance (PMQR) genes, such as *qnr* variants, *aac(6')-Ib-cr*, and the *oqxAB* efflux pump, which can be easily transferred between strains via horizontal gene transfer (Wong et al., 2014; Zhang et al., 2025).

As the diagnostic tools for detecting resistance are shifting toward genomic Antimicrobial Susceptibility Testing (gAST), machine learning algorithms have continuously encountered a diagnostic ceiling known as the "Plateau of 95%" (Ahvar et al., 2025). This plateau occurs as a result of models relying on the presence of high-prevalence genomic markers and genome-wide Linkage Disequilibrium (LD) rather than learning the complex resistance patterns (Ahvar et al., 2025). Consequently, ML models often exploit "proxy" markers correlated with resistance but lacking a causal relationship (Ross et al., 2026). The root cause of this limitation is the concept of "Phylogenetic Leakage", wherein the model memorizes the underlying structure (lineage) of that particular strain instead of the functional biology of resistance (Ross et al., 2026). While it is true that these models often achieve high accuracy on validation data that mimic the training set, they lack generalizability, which makes them dangerous to clinical environments where resistance is often transferred to new lineages (Ross et al., 2026).

Furthermore, the probability given by "black box" models are inherently not trustworthy for clinical decisions without rigorous calibration (Kelly, 2026). Calibration is essential in ensuring that the predicted probabilities are in line with the true phenotypic resistance (Kelly, 2026). Based on current

literature, deep learning models are notorious for being overconfident, where their prediction do not achieve expected accuracy (Qureshi et al., 2026). To overcome these challenges, the adoptions of metrics like Expected Calibration Error (ECE) and Isotonic Regression is necessary to ensure that gAST tools are "clinically honest" and safe for medical deployment (Tolan et al., 2025).

To address these issues, we present a "forensic" ML pipeline designed to decouple biological resistance from phylogenetic ancestry. The approach in this study involves a distillation process, auditing 16.7 million agnostic k-mers against a curated list of universal genomic features to find stable genetic markers for resistance prediction. By using Isotonic Regression, the clinical reliability of the model was checked to ensure medical safety. Finally, the generalizability of the model have been check using Leave-One-Serovar-Out (LOSO) testing method to identify and analyze diagnostic blind spots across the *Salmonella* pangenome.

2 Methods

2.1 Dataset Collection and Phenotype Extraction

Metadata and antimicrobial susceptibility testing (AST) data were gathered from NCBI Pathogen Detection database (NCBI, 2026). Genomic integrity was supported by selecting assemblies that passed the NCBI Pathogen Detection quality control pipeline. Using Python (V 3.10.19) and the Pandas library (V 2.3.3) the data was processed for phenotypic extraction. The AST phenotypes were used to retrieve ciprofloxacin susceptibility category. Resistant and susceptible phenotypes were processed to retain only "S" and "R" string; any field with intermediate, unknown, and invalid genomic assembly identifiers were discarded to mitigate potential label noise arising from historical shifts in clinical breakpoints. Then the string values were binary encoded to 1 (resistant) and 0 (susceptible). The final curated dataset consisted of 8,759 isolates, each linking a genome assembly accession to a binary ciprofloxacin resistance phenotype.

2.2 Gene Feature Engineering and Lineage Mitigation

Gene-level genomic features for the curated dataset were retrieved from the "AMR phenotypes" columns. Their presence and absence were converted into a binary matrix. Following that, the Serovar names were standardized using regex-based cleaning protocol, the names were removed from any S, serovar prefixes and monophasic variants were consolidated into single monophasic variant. To prevent lineage-driven confounding, features were filtered using a custom Phylogenetic Leakage Audit (Collins and Didelot, 2018; Saber and Shapiro, 2020):

- **Variance Filtering:** Removal of genes with variance < 0.01.
- **Correlation-Adjusted Lineage Filtering:** Genes that were present in $> 95\%$ of any specific serovar were flagged as a potential lineage markers. These flagged genes were then checked for their Pearson correlation coefficient and values < 0.1 were used to discard ancestry-informative noise, ensuring the retention of homoplastic mutations.

2.3 Whole Genome k-mer Feature Extraction and Collision Audit

To capture possible resistance patterns without annotated genes, k-mers (k=31) were extracted from the whole genome using KMC3 (5,7). Given the known issue with high dimensionality of k-mer spacing, the sequences were transformed into a fixed-length vector of 2^{24} (16.7 million) features using MurmurHash3 (mmh3) (V 5.2.0) (Senuma, 2025; Caragea et al., 2012). The inherent characteristic of this approach is loss of information; to quantify that loss, a Hashing Collision Audit was conducted by modeling the probability $P(n, D)$ of at least one collision in a feature space of size D after hashing n unique items:

$$P(n, D) \approx 1 - \exp\left(-\frac{n(n-1)}{2D}\right)$$

Using the "Birthday Problem" logic to a pangenome scale where unique k-mers (N) exceeded 10^7, the theoretical collision probability of 100% was calculated. This provided a mathematical baseline to explain the subsequent failure of agnostic hashing to capture functional biological signals. While it is true expanding the bit space to 2^{64} would prevent the 100% collision audit, doing so introduces a significant algorithmic overhead and memory mapping complexity which is not compatible with standard machine learning frameworks. By staying at 2^{24} space, the study highlights the resource-performance trade-off. Gene-centric models provide superior, interpretable results without the necessity for the massive, sparse feature architectures required by agnostic k-mer approaches.

2.4 ML Development and Hyperparameter Optimization

The machine learning models were developed using stratified 80/20 train-test splits where the test set was constructed to include at least one isolate from every serovar. The models included Logistic Regression, Random Forest , and XGBoost (Liang et al., 2021; Pedregosa et al., 2018). To ensure best model performance, RandomizedSearchCV was used to tune the hyperparameter protocol. For the XGBoost model max_depth (range 3–12), learning_rate (0.01–0.2), subsample (0.5–1.0), and gamma (0–5.0) over 20 iterations using 3-fold cross-validation were evaluated. Due to the imbalance of the data, scale_pos_weight was dynamically calculated to address that imbalance.

2.5 Model Calibration and Evaluation

ML models trained on biological data are notorious for being overconfident; to address this, probabilistic outputs were refined using isotonic regression calibration via CalibratedClassifierCV and using five-fold internal cross-validation (da Silva, 2024). This non-parametric approach identifies a non-decreasing function by minimizing the weighted least-squares objective:

$$\min \sum_{i=1}^{n} (y_i - f(\hat{p}_i))^2$$

This non-parametric calibration method was used to correct the non-linear miscalibration and S-shaped reliability curves. Brier Score analysis was then used to quantify the success of the calibration method, ensuring that the predicted probabilities were mathematically aligned with empirical outcomes. The Brier Score (BS) represent the mean squared difference between the predicted probabilities p_i and the actual phenotypic outcomes o_i for n isolates:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$$

Furthermore, to account for the low outcome prevalence and avoid the naive model fallacy, the Brier Skill Score (BSS) was calculated as: $1 - (BS_{model}/BS_{naive})$, where BS_{naive} represented the variance of the sample labels. Following this Decision Curve Analysis (DCA) utility test was applied to calculate the Net Benefit of model-guided therapy across a range of risk thresholds (0.01 to 0.99), ensuring the pipeline's actionability compared to "Treat All" and "Treat None" strategies.

2.6 Cross-Serovar Generalization and Lineage Error Audit

The model's lineage-robustness was evaluated using Leave-One-Serovar-Out (LOSO) validation (Geroldinger et al., 2023). To address lineage-specific bias, a post-hoc lineage error audit was used. For the high-interest serovars identified in SHAP analysis (*Enteritidis*, *Uganda*, *Worthington*, *Heidelberg*, *Montevideo*, *Hadar*) recall, precision, and error rate were calculated independently. Furthermore, to evaluate the rigorousness of the test a genetic independence audit was conducted, by calculating the mean Jaccard distance between the held-out isolates and their nearest neighbors in the training set. This ensured that predictive failures were not clustered within a specific ancestral branch of the *Salmonella* phylogeny.

2.7 Feature Stability Analysis

To identify robust predictors, Stability Selection was performed using 100 iterations of subsampling at a 50% fraction (Haftorn et al., 2023).

- **K-mer Tiered Audit:** For the high-dimensional k-mer data first χ^2(Chi-square) univariate filter was used as a first-tier filter to select the top 50,000 features.
- **Regularized Probe:** Subsamples were fitted using an Elastic-Net regularized SGDClassifier (for k-mers) and L1-regularized Logistic Regression (for genes). A stability score was calculated as the selection frequency across iterations; features with a frequency ≥ 0.5 were considered stable biological predictors.

2.8 Model Explainability and Forensic Failure Analysis

For the model explainability, SHAP values extracted from the XGBoost model via pred_contribs were used to quantify feature contributions. The biological findings of this method were verified by cross-referencing high-impact SHAP features with established antimicrobial resistance (AMR) determinants (e.g., *gyrA*). For Forensic Failure Analysis, misclassified isolates were audited to identify genomic drivers of error. Misclassification profiles, including error type (FP/FN) and serovar-specific SHAP distributions, were visualized using clustered heatmaps and OncoPrints (Gu et al., 2014; Gu, 2022; Gu et al., 2016).

3 Results

3.1 Dataset Composition and Lineage Distribution

The final curated datasets used for model training consisted of 8,759 *Salmonella* isolates, representing a diverse pangenome as shown in Figure 1. However the serovar distributions were dominated by several primary lineages with *S. Kentucky* (*n*=1501) and *S. Typhimurium* (*n*=1128) occupying the largest sample volumes. These were followed by *S. Heidelberg* (*n*=721), *S. Enteritidis* (*n*=646), and *S. Infantis* (*n*=536), making the dataset suitable for evaluating resistance across the major *Salmonella* serovars.

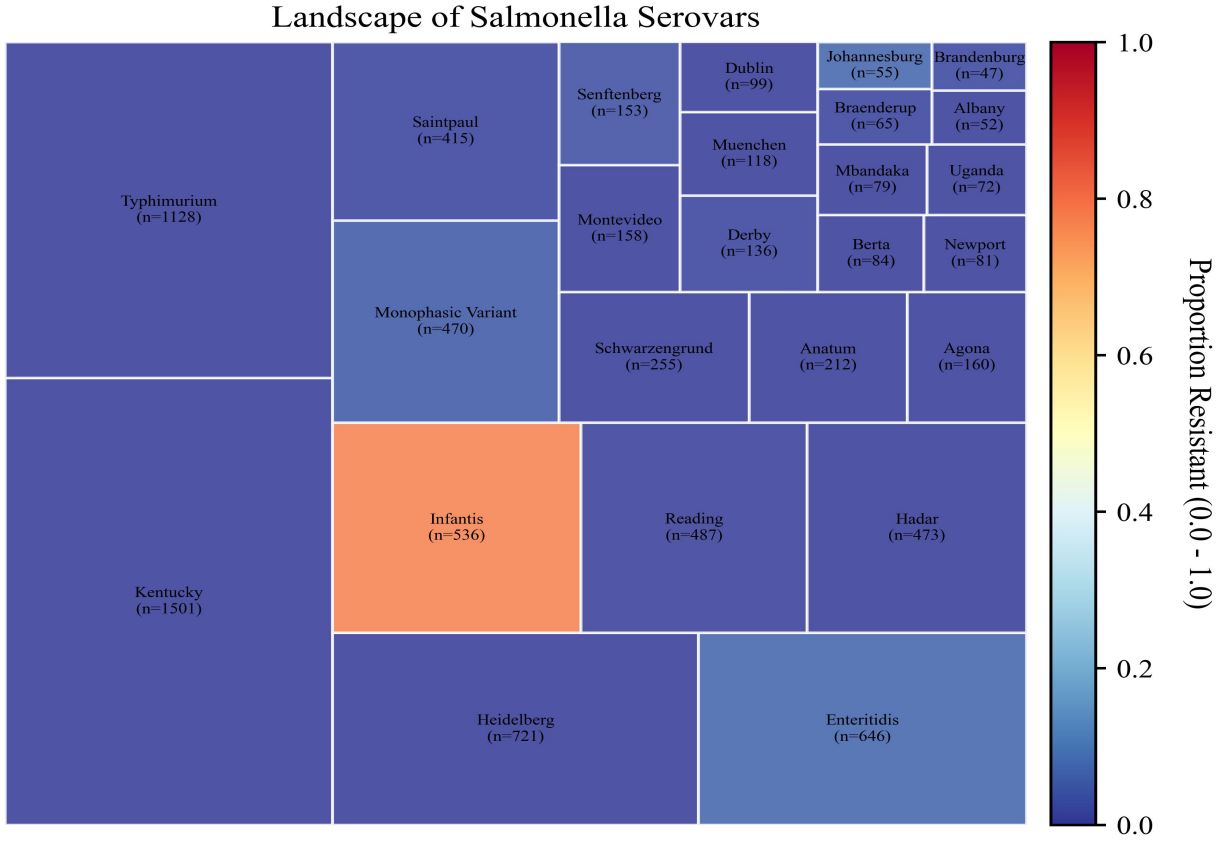


Figure 1: Landscape of Salmonella Serovars. Treemap where rectangle area indicates sample size (*n*) and color represents the proportion of ciprofloxacin resistance (0.0 to 1.0).

Further phenotypic analysis revealed a sharp, lineage-specific divide in ciprofloxacin resistance. As shown in Table 1, the majority of the serovar classes including large clades like *S. Typhimurium*, *S. Heidelberg*, and *S. Hadar*—exhibited 0.0% resistance, which is visually represented as dark blue in Figure 1. While *S. Kentucky* was the most prevalent serovar (*n*=1501), only 0.1% (*n*=2) of its isolates were resistant. This low resistance rate in the dataset is likely a representation of isolates from U.S. food and environmental surveillance, rather than the global superbug ST198 lineage. On the other hand *S. Infantis* with 536 isolates served as distinct outlier with high resistance rate of 77.2% (*n*=414), appearing as a distinct orange block in the treemap.

Table 1: Isolate demographic and resistances prevalence. Summary of the top fifteen serovar, with total number of isolates (N) and resistance/susceptibility counts.

Serovar	**Total** N	**Resistant** n **(%)**	**Susceptible** n **(%)**
Kentucky	1501	2 (0.1%)	1499 (99.9%)
Typhimurium	1128	0 (0.0%)	1128 (100.0%)
Heidelberg	721	0 (0.0%)	721 (100.0%)
Enteritidis	646	45 (7.0%)	601 (93.0%)
Infantis	536	414 (77.2%)	122 (22.8%)
Reading	487	2 (0.4%)	485 (99.6%)
Hadar	473	0 (0.0%)	473 (100.0%)
Monophasic Variant	470	23 (4.9%)	447 (95.1%)
Saintpaul	415	5 (1.2%)	410 (98.8%)
Schwarzengrund	255	0 (0.0%)	255 (100.0%)
Anatum	212	0 (0.0%)	212 (100.0%)
Agona	160	0 (0.0%)	160 (100.0%)
Montevideo	158	0 (0.0%)	158 (100.0%)
Senftenberg	153	5 (3.3%)	148 (96.7%)
Derby	136	2 (1.5%)	134 (98.5%)

As illustrated in Figure 1, the top right corner showcases the minority group but with noticeable resistance levels, including the Monophasic Variant (4.9%, n=23), *S. Enteritidis*(7.0%, n=45), and *S. Senftenberg* (3.3%, n=5). Other minority groups including *Montevideo* (n=158) and *S. Newport* (n=81), remained entirely sensitive. This heavy imbalance in the data presents a significant robustness challenge for training ML models and assessing whether it can distinguish biological predictors independent of ancestral noise.

3.2 Comparative Model Performance: Genes vs. K-mers

The predictive performance of the k-mer agnostic model was evaluated against the gene-based model. The agnostic k-mer model with k=31 and 2^{24} features resulted in complete methodological failure, as the theoretical hashing collision audit predicted collision probability of 100.0% when 10 million unique k-mers are mapped into a 16.7 million (2^{24}) feature space. Because of this high collision, no stable features were selected, resulting in a complete loss of biological signal. Therefore, the model was unable to proceed with training and evaluation, establishing a "diagnostic ceiling" where agnostic hashing failed to capture the genomic determinants of resistance.

In comparison to the agnostic model, the gene-based models showed exceptional predictive performance. As highlighted in Figure 2A, all three genes-based models produced Receiver Operating Characteristic (ROC) curves that prioritized high True Positive (TP) Rates with minimal False Positives (FP). XGBoost among the three performed the best with an ROC value of 0.989, followed by Random Forest (0.986) and Logistic Regression (0.984).

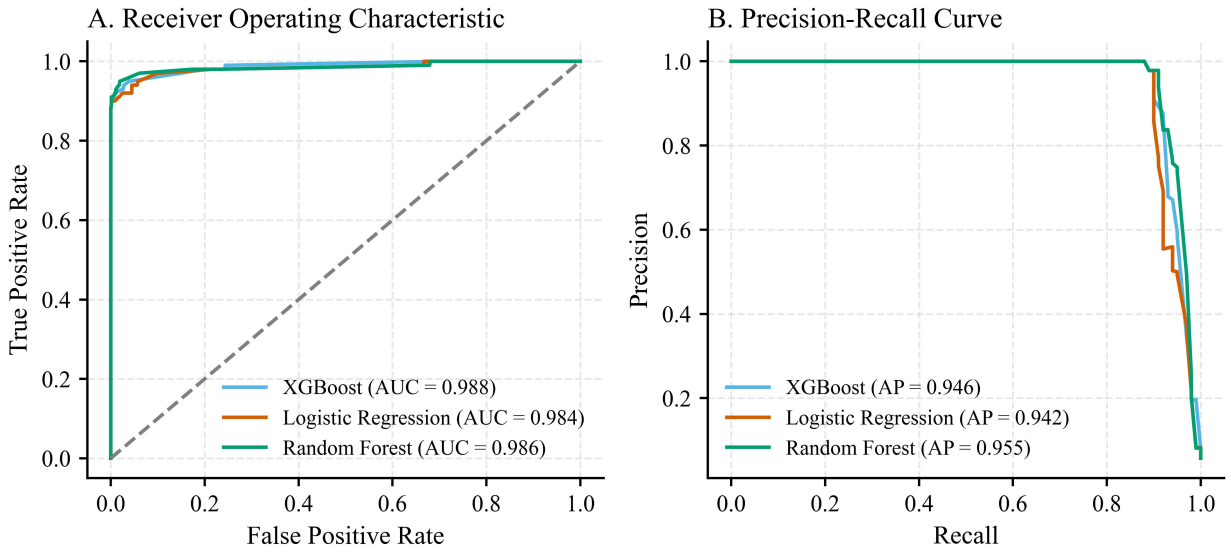


Figure 2: Performance Curves for Gene-Based Models. (A) ROC curves showing XGBoost (orange), Random Forest (green), and Logistic Regression (blue) performance. (B) Precision-Recall curves for all three models in same color formatting.

Figure 2B which shows the Precision-Recall curves for all three models demonstrates the ability to maintain perfect precision (1.0) until reaching approximately 0.9 recall. Random Forest model among the three managed to maintain the highest precision for the longest duration with an average precision (AP) of 0.955, this was closely followed by XGBoost (0.946) and Logistic Regression (0.942). Overall, based on the models' metrics Random Forest had the highest F1-score of 0.943, while Logistic Regression showed lower performance with an F1-score of 0.923. The full performance metrics for these uncalibrated models are summarized in Table 2.

Table 2: Performance metrics for uncalibrated gene-based models. The table showcases accuracy, precision, recall, F1-Score, ROC AUC, brier score, VME and ME for all three tested models.

	Model Performance Metrics									
Model	**Acc**	**Prec**	**Rec**	**F1**	**AUC**	**Brier**	**BSS**	**Rel**	**VME**	**ME**
RF Uncalibrated	0.994	0.978	0.91	0.943	0.986	0.00799	0.852	0.0788	9	0.121
LR Uncalibrated	0.991	0.947	0.9	0.923	0.984	0.00993	0.816	0.117	10	0.303
XGB Uncalibrated	0.993	0.978	0.89	0.932	0.989	0.0471	0.124	0.0602	11	0.121

3.3 Calibration and Probability Refinement

The output of the uncalibrated models showed significant probability distortion, making it critical to perform post-hoc calibration to ensure clinical reliability. As shown in Figure 3A, the raw output of the uncalibrated models showed erratic behavior with their curve deviating frequently from the diagonal line (y=x). Upon closer inspection, all three uncalibrated models showed signs of

complete broken behavior in the mid-probability range (0.4-0.6); for example, Logistic Regression predicted zero fraction of positives before sharply jumping to 100% at the 62% probability mark. Furthermore, both Random Forest and the XGBoost model showed marked signs of over and under confidence throughout.

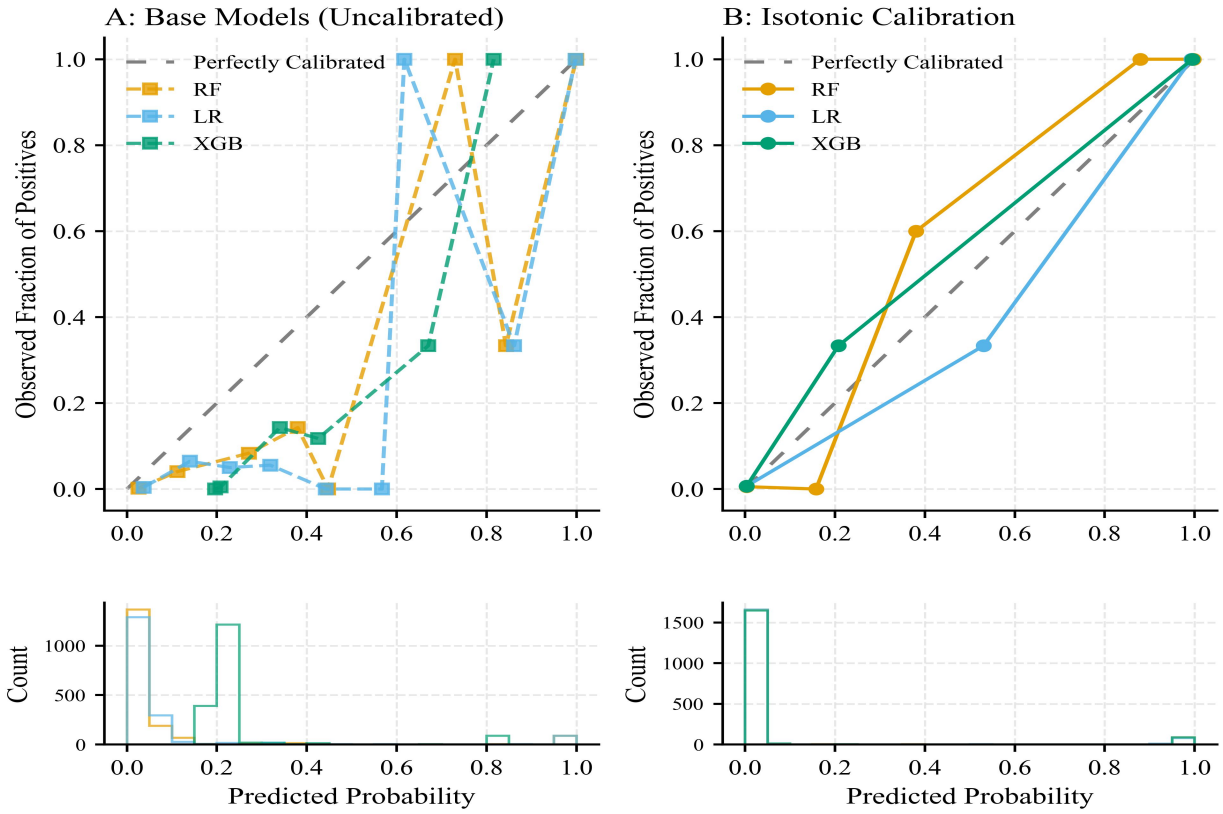


Figure 3: Reliability Diagrams and Probability Distributions. Panel (A) shows the jagged, non-monotonic reliability curves of the uncalibrated models showcasing signs of over and under confidence for all models. Panel (B) shows the reliability curves for the models following isotonic regression, which shows significant improvement compared to the uncalibrated models. Bottom histograms illustrate the frequency of predicted probabilities.

As stated before, to correct these misalignments of the models, isotonic regression was applied. This calibration method resulted in non-decreasing reliability curves as illustrated in Figure 3B. This refinement made the predicted probability much closer to the observed fraction of positives, even though all the models maintained "push-to-extremes" behavior. The accompanying histogram plots in Figure 3 revealed that nearly all isolates after calibration were concentrated in the 0.0 probability bin (>1500 count) with only a small subset localized at 1.0, effectively binarizing the model's confidence. Furthermore, the impact of the isotonic regression has been quantified and shown in Table 3. Although accuracy and ROC AUC for all three tested models remained largely stable, the Brier Score, a direct measure of calibration error showed substantial improvement after applying calibration. Among the models XGBoost showed great signs of improvement, where the Brier Score dropped from 0.0471 (the worst of the base models) to 0.0064; while the Random Forest model overall showed the best reliability with the lowest Brier Score of 0.0059. These results

showed that while alone the genetic noise of the data was enough for discriminative power, the post-hoc calibration is essential for transforming the models raw output into biologically meaningful probability estimates. Critically, while the calibration method managed to bring down the Brier Score it also introduced a marginal clinical tax in the form of increased Very Major Error (VME) rates. For instance the VME score of the XGBoost model increased from 11 to 12, and Random Forest from 9 to 12. However, this was balanced by a complete elimination of Major Errors (ME=0) in both calibrated version of the XGBoost and Random Forest models, ensuring a highly specific diagnostic profile that meets clinical standards for ciprofloxacin AST (VME< 3%). All the performance metric related to the clibration of the models were evaluated against FDA 510(k) and ISO 20776-2:2007 regulatory criteria for automated AST systems, which mandate a Major Error (ME) rate of $\leq 3\%$ and a Very Major Error (VME) rate of $\leq 3\%$.

Table 3: Impact of Isotonic Calibration on model reliability. The table summarizes performance metrics for all three models after calibration, it shows Accuracy, Precision, Recall, F1-Score, ROC AUC, Brier Score, VME and ME with significant improvement compared to the uncalibrated models (Table 2).

	Model Performance Metrics									
Model	**Acc**	**Prec**	**Rec**	**F1**	**AUC**	**Brier**	**BSS**	**Rel**	**VME**	**ME**
RF Calibrated	0.993	1	0.88	0.936	0.985	0.00586	0.891	0.0176	12	0
LR Calibrated	0.993	0.978	0.89	0.932	0.982	0.00665	0.877	0.013	11	0.121
XGB Calibrated	0.993	1	0.88	0.936	0.988	0.00645	0.88	0.00527	12	0

To further evaluate the clinical utility of these refined probabilities Decision Curve Analysis (DCA) was applied. As shown in Figure 4, all three calibrated models showed a significant net benefit over the treat all and treat none baseline across a broad range of risk thresholds (0.05 to 0.50). This demonstrates that the model guided approach to ciprofloxacin administration effectively reduces unnecessary antibiotic use, and in the same time maintaining high sensitivity for resistant cases.

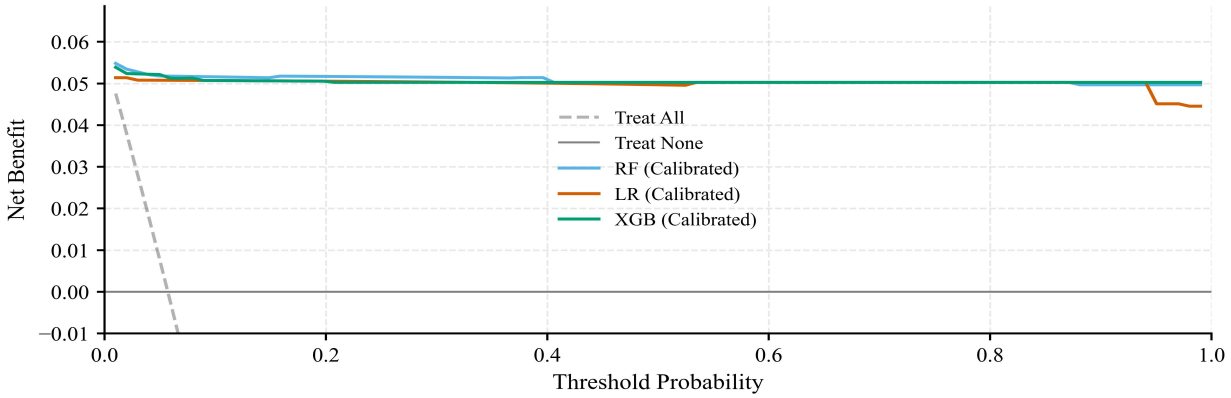


Figure 4: Decision Curve Analysis (DCA). The plot demonstrates the clinical Net Benefit of the calibrated gene-based models (RF, LR, and XGB) compared to standard empirical strategies across varying risk thresholds.

3.4 Feature Stability and Explainability

The difference between the gene-based and agnostic k-mer model is most evident in how they respond to feature stability selection and their biological interpretability. As illustrated in Figure 5, the k-mer model starting with 1.8×10^7 features, experienced near-instantaneous collapse in stability. None of the features even survived the 0.05 stability threshold, indicating that the high-dimensional feature selection was entirely dependent on specific training subsets and lacked reproducible signal. In comparison, the gene-based models remained mostly stable with 21 features remaining at the 0.4 threshold and 4 features maintaining a stability score of 1.0 across all iterations.

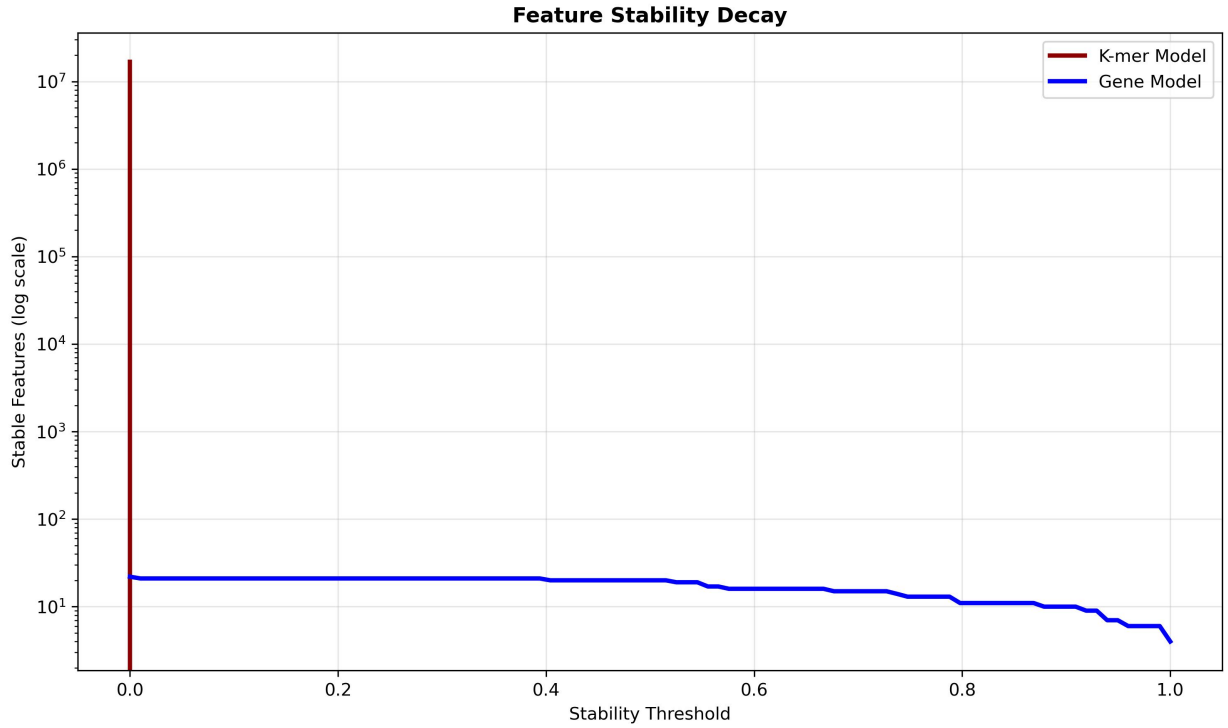


Figure 5: Feature Stability Decay. It shows a comparison of feature reproduction between the gene-based models (Blue line) and the agnostic k-mer model (Red line). The X-axis shows the total model iteration in which the feature must be selected for it to be considered stable, the Y-axis is the logarithmic scale.

Furthermore, cross-validation (CV) stability further confirmed the reliability of the gene-based models, as shown in Figure 6. The ROC AUC metrics for XGBoost (0.945) and Random Forest (0.94) remained stable across the folds; while it is true the calibration method introduced some variance especially in the case of Logistic Regression, where an outlier dropped to 0.39. The overall discriminative power of the XGBoost and Random Forest model remained tightly packed with most being clustered between 0.92 and 0.99, while the Logistic Regression model exhibited a significant failure in one fold (0.39). Therefore, the data gathered from the CV plot highlights that the identified features are not artifacts of specific data splits but represent genuine genomic determinants.

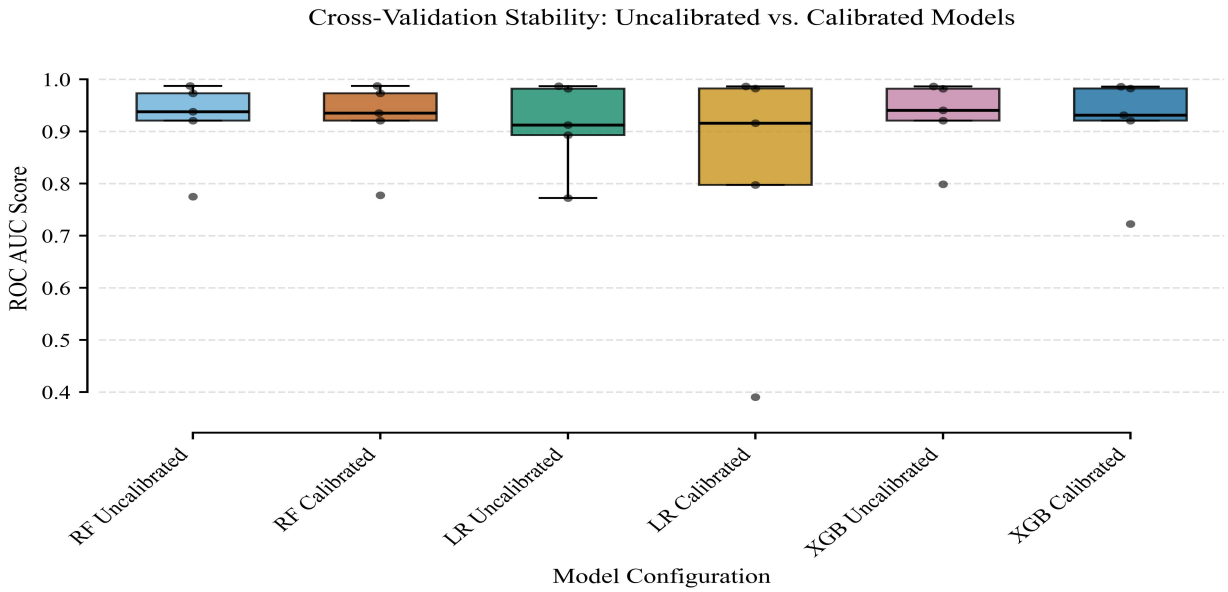


Figure 6: Cross-Validation Stability. The boxplot overlaid with jitter points shows the ROC AUC metrics for both the calibrated and uncalibrated gene-based models over the five-fold cross-validation test. Across the folds the model maintained high metrics tightly between 0.92 and 0.99 ROC AUC.

To understand the biological drivers of these predictions, SHAP (SHapley Additive exPlanations) values were extracted. The analysis revealed the *gyrA_D87Y* mutation as the strongest predictor of resistance with an importance factor of 0.88 driving most of the model's decision. As shown in the beeswarm plot (Figure 7), the absence of *gyrA_D87Y* (blue points) strongly pushed the model toward susceptibility, while its presence (red points) significantly increased the resistance probability.

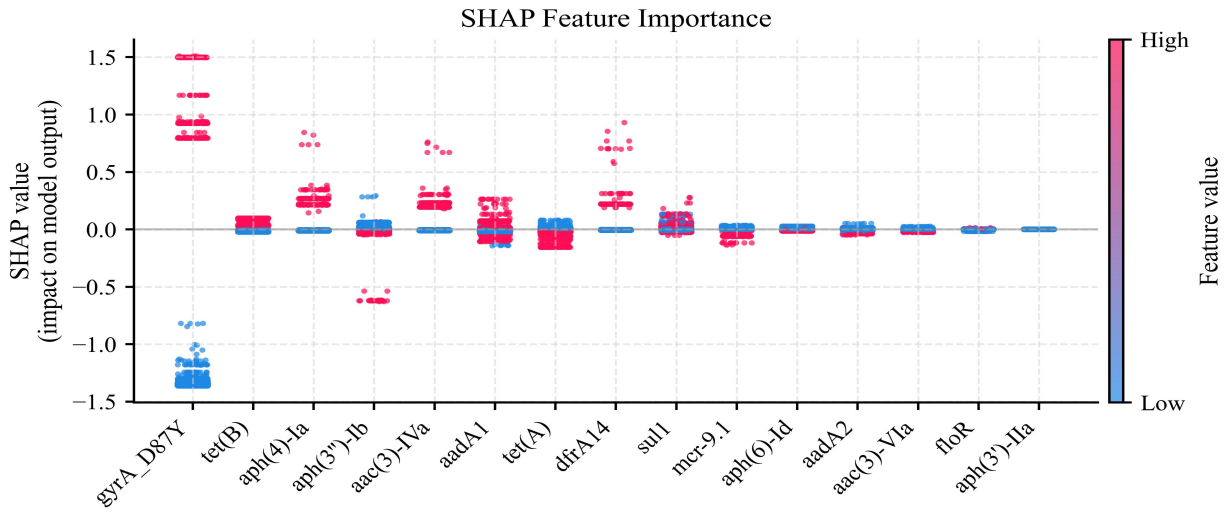


Figure 7: SHAP Feature Importance. A Beeswarm plot that shows the importance of the top 20 genes in driving the model's output. Red points indicate presence of the genes pushing the model toward resistance, while blue points indicate absence/low value targets.

Secondary contributors included *aph(4)-Ia*, *aac(3)-IVa*, and *dfrA14*, which showed moderate positive impacts on resistance predictions. Interestingly, *aph(3")-Ib* demonstrated an inverse relationship, where its presence pushed the model toward the susceptibility region. Despite the initial selection of 22 features, the bottom 12 genes—including *tet(C)*, *blaCTX-M-65*, and *blaTEM*—showed negligible SHAP values and importance scores near zero, suggesting that the model effectively prioritized a concentrated subset of primary AMR determinants.

3.5 Cross-Serovar Generalization (LOSO)

To evaluate the model's lineage-robustness, a Leave-One-Serovar-Out (LOSO) validation was performed, and to quantify the rigor of this test a genetic independence audit was conducted. This validation test was done to see if the model could identify universal genomic markers for resistance or it would just memorize lineage-specific patterns. As shown in Figure 8, the model demonstrated robust performance with high density serovars, but exhibited significant performance decay in lineages with sparse resistant representation, possibly due too low number of resistance cases on those specific serovars.

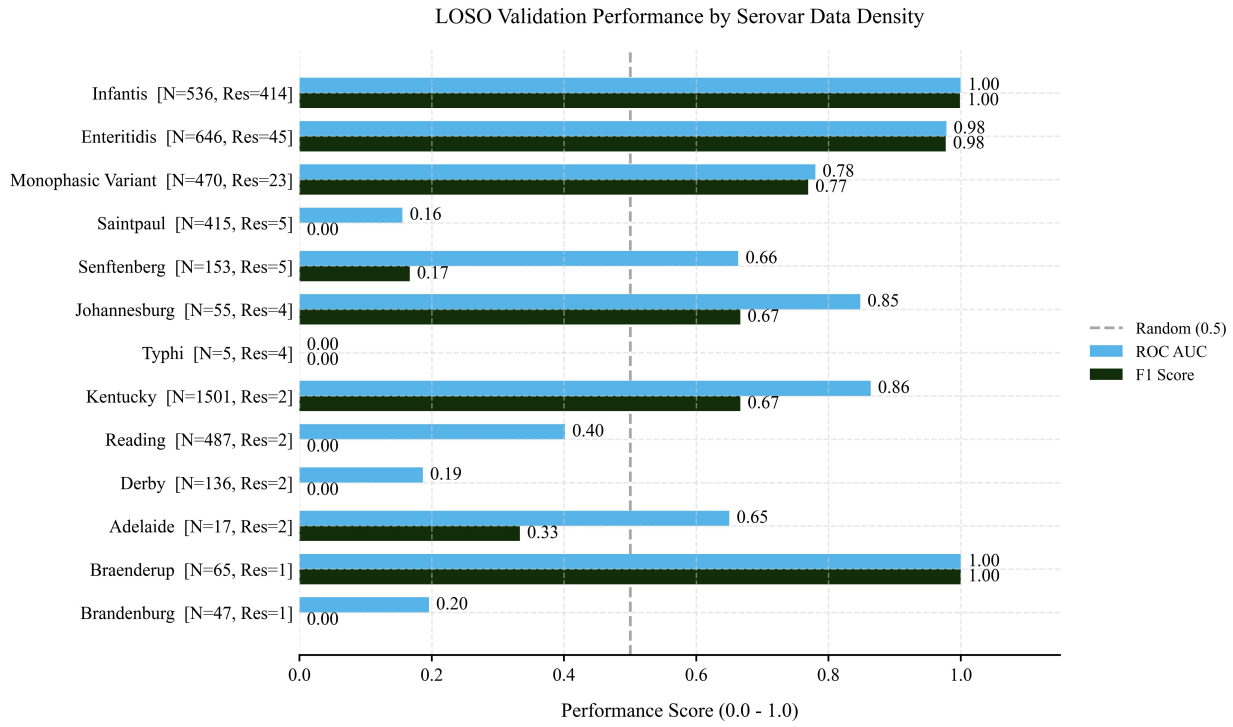


Figure 8: LOSO Validation Performance by Serovar Data Density. A grouped bar chart showing ROC AUC (blue) and F1 Score (green) highlighting the performance of the model on specific serovars. The dashed red line in the middle indicates random guessing.

Overall, serovars with high resistance cases ($Res \geq 23$) made the generalizability of the model nearly perfect. For instance *S. Infantis* ($Res = 414$) and *S. Enteritidis* ($Res = 45$) yielded ROC AUC scores of 1 and 0.978, respectively. This showed that the model can accurately identify resistance within these large clades, even when they are entirely excluded from the training set. However, when it comes to low-density groups, the performance of the model becomes highly variable. While the model successfully generalized to *S. Johannesburg* ($Res = 4, AUC = 0.85$) and *S. Kentucky* ($Res = 2, AUC = 0.86$), it failed completely on *S. Typhi* ($Res = 4$) and *S. Saintpaul* ($Res = 5$), where ROC AUC fell below the 0.5 threshold. This collapse in performance highlights a 'Trust Horizon'; when the genomic background is 15% divergent from the training set. In these cases the model can no longer rely on learned pangenomic anchors. Conversely the success on *S. Infantis* (AUC=1.0) at a distance of 0.062, proved that the model can generalize to distinct lineages, provided the core resistance signatures are preserved within that pangenomic distance.

Table 4: LOSO performance metrics across diverse Salmonella lineages. Summary of accuracy, ROC AUC, F1-score, and error rates (VME/ME) for held-out serovars.

Left Out Serovar	*N*	R Count	Acc	AUC	F1	VME	ME	Gene Dist
Adelaide	17	2	0.765	0.65	0.333	50	20	0
Braenderup	65	1	1	1	1	0	0	0
Brandenburg	47	1	0.979	0.196	0	100	0	0.00426
Derby	136	2	0.684	0.187	0	100	30.6	0.00725
Enteritidis	646	45	0.997	0.978	0.977	4.44	0	0
Infantis	536	414	0.998	1	0.999	0	0.82	0.0616
Johannesburg	55	4	0.964	0.848	0.667	50	0	0.00455
Kentucky	1501	2	0.999	0.864	0.667	50	0	0.00176
Monophasic Variant	470	23	0.981	0.78	0.769	34.8	0.224	0.00247
Reading	487	2	0.996	0.402	0	100	0	0.00302
Saintpaul	415	5	0.988	0.156	0	100	0	0.019
Senftenberg	153	5	0.935	0.664	0.167	80	4.05	0.0239
Typhi	5	4	0	0	0	100	100	0.15

The error analysis shown in Table 4 highlights that the model's performance across the serovars largely depends on data density. Serovars with failing ROC AUC values across the board have Very Major Error (VME) rates of 100%, indicating a total failure to detect resistance when the underlying genomic background was unfamiliar to the model. A notable outlier was *S. Braenderup*, which achieved perfect metrics (1.00) despite having only a single resistant case. A noteworthy finding is that on several serovars with perfect metrics, isolates exhibited a genetic distance of 0.0 suggest that the genetic difference between the training and test set was identical. This suggests that performance in these cases represents clonal recognition rather than true mechanistic generalization. These results show that while the model can perform well on large-density clades that represents resistance mechanisms well, its predictive power remains constrained by the diversity of the training pangenome.

3.6 Forensic Failure Analysis

The final evaluation of the model's performances involved a deep analysis of the model's predictive error to understand if misclassifications were stochastic or driven by specific genomic blind spots. As shown in the OncoPrint (Figure 9) the model, despite achieving high global accuracy, it tend to rely on very specific set of narrow features. Primarily the *gyrA_D87Y* mutation served as the primary anchor for predicting resistance; this was specifically shown within *S. Infantis* (Purple) cluster where phenotype (Orange) and genotype (Grey) show perfect synchronization. In contrast to *gyrA_D87Y* mutation, genes like *tet(A)* were found in approximately 50% of the isolates, but they provided low specificity for ciprofloxacin resistance prediction.

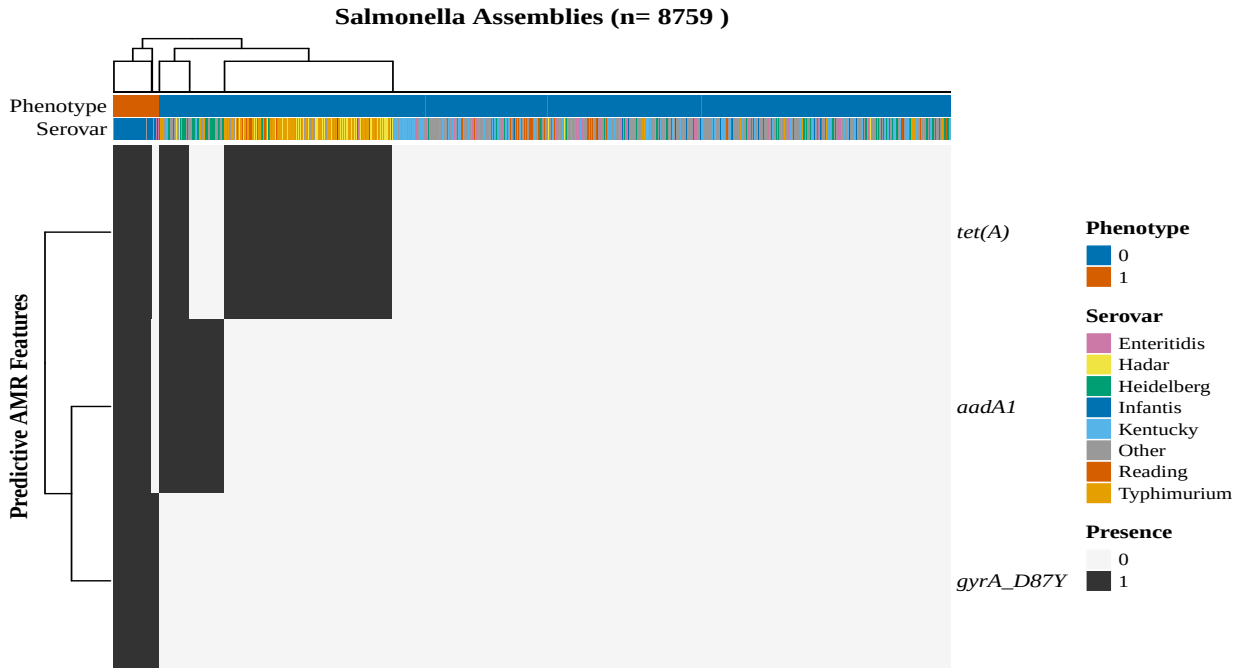


Figure 9: Genomic Landscape of Salmonella (n=8,759). An OncoPrint showcasing the distribution of resistance genes (*tet(A)*, *aadA1*, *gyrA_D87Y*) across the *Salmonella* serovars. Grey boxes indicate presence of the genes, with orange phenotypes indicating positive presence of the AMR feature in that species serovars.

Further failure analysis revealed a critical vulnerability of the model: Feature Over-reliance. As demonstrated in the failure heatmap (Figure 10), all the missed cases are false negative assemblies due to the absence of *gyrA_D87Y* mutation, represented by consistent teal blocks in the SHAP contribution column. In these specific cases the absence of *gyrA_D87Y* drove such a strong signal that pushed the model toward susceptible, even when secondary resistance genes like *aph(4)-Ia* and *aac(3)-IVa* were present and contributing positive SHAP values. This data highlights alternative mechanisms for resistance in which the model has not yet weighted heavily enough to override the *gyrA* signal.

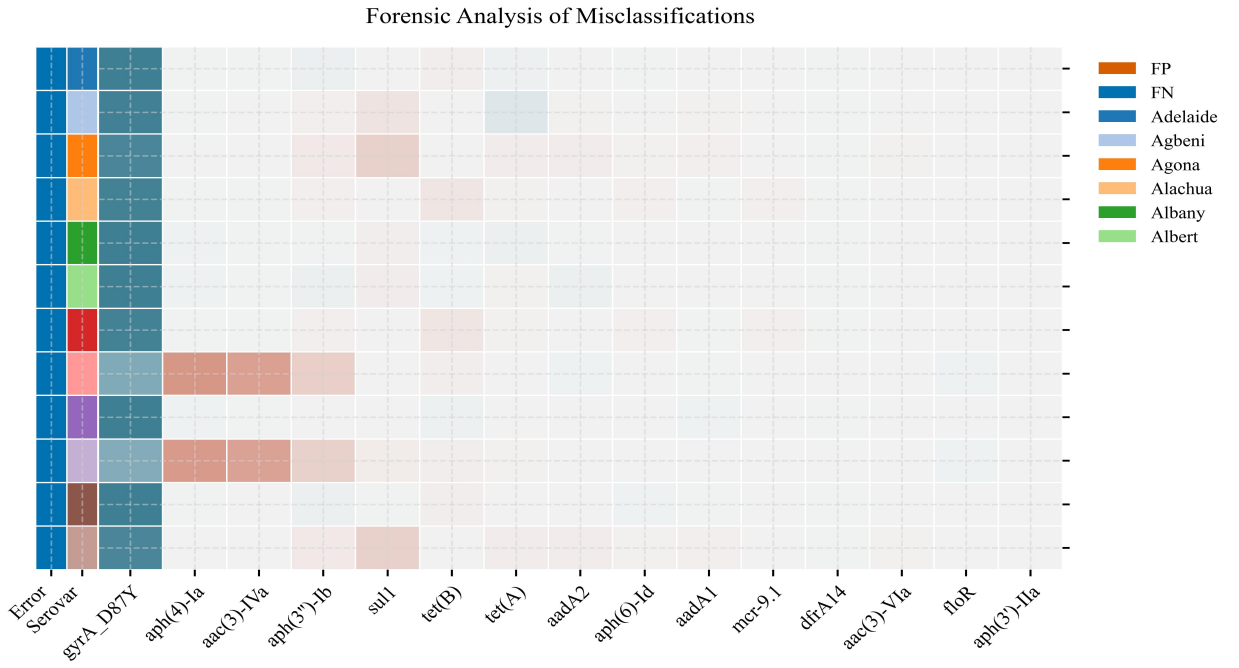


Figure 10: Forensic Analysis of Misclassifications. Clustered heatmap of SHAP values for 12 False Negative assemblies across multiple lineages.

The performance consistence of the model across the high-interest serovar further validates the model's reliability on specific resistances signals. This is particularly shown with *S. Enteritidis* (n=111) in Table 5, the model managed to maintain perfect precision and recall. However, the 0.0% error rate observed for some of the serovars must be interpreted with caution, particularly for the *S. Uganda* and *S. Hadar*, because these clades entirely consisted of susceptible isolates, and model's ability to detect resistance in these specific genomic backgrounds remains unproven. This forensic audit confirms that while the model is highly sensitive to dominant resistance genotypes—primarily *gyrA*-mediated resistance in *S. Infantis*—its clinical utility for rare or emerging mechanisms is currently constrained by the "genomic anchors".

Table 5: Serovar-Specific Performance Metrics. The table summarizes the performance metrics for the specific serovars, consisting of Recall, Precision, Lineage-Specific error rate, and their Briar Score.

Serovar	Total Samples	Lineage-Specific Error Rate (%)	Recall	Precision	Calibration Error
Enteritidis	111	0	1	1	6.44e-06
Uganda	18	0	0	0	9.84e-06
Worthington	7	0	0	0	9.36e-06
Heidelberg	112	0	0	0	1.73e-05
Montevideo	29	0	0	0	0.000383
Hadar	84	0	0	0	5.01e-06

4 Discussion

The transition from laboratory techniques to computational frameworks to detect Antimicrobial Resistance (AMR) presents a significant leap forward, but it also necessitates a rigorous framework for how genomic data are being represented and interpreted. Our results showed that frameworks designed around alignment-free, agnostic k-mers often face a "diagnostic ceiling". The 100% theoretical collision audit observed with the k-mer model is a basic limitation of using probabilistic data structures like Compact Hash Tables (CHT) (Wood et al., 2019). The inability of the k-mer model to select stable features is due to "soft collisions" when 10 million features are mapped into a 2^{24} numerical space using functions like MurmurHash3, leading to genomic features overlapping and the loss of biological signal (Wood et al., 2019). While it is true that k-mer-based models offer significant computational velocity and memory optimization, they often struggle with low-abundance genes (common in AMR) and they lack the resolution to tell the difference between core polymorphisms and accessory genes within closely related *Salmonella* lineages (Wood et al., 2019; Moore et al., 2024; Wu et al., 2018). Consequently, this shows the importance of alignment-based models as they preserve gene order and the details necessary for AMR prediction (Moore et al., 2024).

The dominance of the *gyrA_D87Y* mutation as the anchor for predicting resistance, and with an importance factor of 0.88, is an accurate reflection of the hierarchical nature of fluoroquinolone resistance. It has been shown in *S. Infantis* and *S. Enteritidis* that the primary resistance mechanisms toward fluoroquinolone resistance are through chromosomal target-site mutation in Quinolone Resistance-Determining Regions (QRDR) (Ferrari et al., 2013; Chang et al., 2021; Acheampong et al., 2019). Mutations, particularly at codons *Ser83* and *Asp87* of the *gyrA* subunit, have been shown to greatly affect the affinity of the drug binding to the DNA-enzyme complex (Ferrari et al., 2013). While it is true that Plasmid-Mediated Quinolone Resistance genes were present in the dataset alongside the *gyrA_D87Y* gene, such as *qnrS1*, *aac(6')-Ib-cr*, and *oqxAB*, they only provide low-level resistance toward ciprofloxacin (Ferrari et al., 2013; Rushdy et al., 2013). Therefore, they function as evolutionary genes to provide a "survival window" under low antibiotic pressure, giving the strain enough time to facilitate and develop high resistance through chromosomal mutations (Chang et al., 2021). This shows that the model's over-reliance on *gyrA_D87Y* as the primary indicator of resistance is not a sign of technical over-fitting, but a verification that the model has identified the primary driver of the resistant phenotype (Zhang et al., 2025).

The failure analysis revealed 12 false negative assemblies, showcasing the biological boundaries of a gene-based model's predictive power. Based on literature, phenotypic resistance in the absence of *gyrA_D87Y* mutation stems from a complex interplay between alternative mechanisms, including non-*gyrA* QRDR mutations in *parC* or *gyrB* (Guerra et al., 2003; Acheampong et al., 2019). Alternatively, the overexpression of some efflux pump systems like the ones associated with *AcrAB-TolC* Resistance-Nodulation-Division (RND) superfamily, can increase resistance by preventing high level of intracellular drug levels (Zhang et al., 2025). On top of these alternative mechanisms, the presence of mobile genetic elements flanked by IS26 insertion sequences can lead to horizontal gene transfer of complex resistance profiles that are not fully captured by a static 22-gene feature set (Chen et al., 2020). These datapoints suggest that while the model can accurately detect resistance within common genotypes, its reliability for uncommon or rare resistance pathways requires a broader diagnostic lens that incorporates efflux regulation and porin modulation (Zhang et al., 2025; Chen et al., 2020). Finally, one inherent limitation of short-read

WGS-based models is their inability to detect unstable heteroresistance, where a small fraction of the population is resistant but they may fall below the bioinformatic detection threshold while still conferring a resistant phenotype in clinical assays (Zhai et al., 2024).

The reality of applying any predictive models in clinical settings hinges on the distinction between discrimination and calibration. While the models' raw outputs had high AUROC scores, they exhibited significant probability distortion. This could lead to clinical harm in predicting patient risk through overestimation or underestimation (Jiang et al., 2011). This is particularly concerning in AST models where test results might lead to inappropriate use of broad-spectrum antibiotics, or even inadequate therapy for that specific infection (Jiang et al., 2011). By applying isotonic regression, the raw outputs of the models were transformed into refined probability estimates, achieving an average Brier Score of 0.0063. This score outperforms established benchmarks used in other machine learning applications in clinical settings. For instance, a Brier Score of 0.022 is used as a benchmark in dialysis studies, while 0.21 is used for hospitalization risk models (Chang et al., 2026). Such high-fidelity calibration is essential for individualized treatment decisions, as it ensures that a predicted probability of resistance corresponds accurately to empirical frequency (Jiang et al., 2011; Zhao et al., 2025). Furthermore achieving a Brier Skill Score of 0.88 confirmed that the model low calibration error a result of high-resolution genomic learning rather than statistical artifacts of an imbalanced dataset.

The performance decay observed in the Leave-One-Serovar-Out (LOSO) validation test, particularly for *S. Typhi*, illustrates one of the core challenges in many genomic ML tasks: "Phylogenetic Leakage". Often models trained on genomic data learn to recognize the genomic signature of a specific lineage and associate it with resistance, rather than learning the universal mechanisms (Bhandari et al., 2023). The primary explanation for the performance collapse on *S. Typhi* could be due to its highly clonal population structure, which is a challenge for prediction when models are trained on diverse serovars, and which is why performance sharply drops when applied to such human-restricted lineages (Akshay et al., 2023).

To spot and mitigate the risk of Phylogenetic Leakage, strategies must involve a "species-holdout" protocol and interpretability frameworks like SHAP, to ensure models learn biological markers rather than random genetic differences (Akshay et al., 2023; Sikorski et al., 2022). By using this "species-holdout" one can ensure that the model can bypass the noise of natural evolution and prioritize the homoplastic mutations needed for AMR prediction (Hurtado and Pancaldi, 2026; Green et al., 2025). These homoplastic variants like the *gyrA* QRDR mutations provide high-fidelity signals for convergent evolution driven by antibiotic pressure, which occurred independently on at least 94 separate occasions in *S. Typhi* (Hurtado and Pancaldi, 2026; admin, 2026). This rigorous strategy ensures that high performance metric of a model is based on evolutionary "hotspots" that remain stable across diverse genomic backgrounds, effectively mitigating overoptimistic estimates caused by lineage-specific signatures (Green et al., 2025). Furthermore, while the model is effectively calibrated, future iterations will benefit from Out-of-Distribution (OOD) detection to formally define the model's 'Trust Horizon'. Specifically, the 99.9% susceptibility of *S. Kentucky* suggested that the training data is weighted heavily toward the U.S. food-surveillance isolates rather than the global ST198 resistant lineage. This further highlights that the model's 'Trust Horizon' is ecologically constrained and it requires future validation on international 'superbug' clones.

Finally, the inverse relationship of the *aph(3")-Ib* gene with resistance could suggest the presence of collateral sensitivity or metabolic trade-offs within the *Salmonella* pangenome. Aminoglycoside resistance genes are known to reduce the proton motive force (PMF), which can sometimes reduce

the activity of multi-drug efflux pumps and restore sensitivity to fluoroquinolones (Azimi and Rastegar Lari, 2017; Pál et al., 2015). On top of that, carrying excess cargo genes imposes a significant fitness cost on the bacterium (Fuzi, 2025). Often in environmental conditions where antibiotic pressure goes up and down, bacteria must find a balance for expression of resistance genes which can sometimes lead to the loss or suppression of certain markers in successful international clones (Fuzi, 2025). Furthermore, a targeted co-occurrence audit within the database confirmed that *aph(3")-Ib* exhibits negligible linkage with functional PMQR determinants in this population, with co-carriage rates of 0.15% for *qnrS1* and 0.04% for *qnrB1*. While these trade-offs are not always robust in clinical populations, they offer a promising avenue for antibiotic cycling strategies aimed at mitigating the rise of multi-drug resistant lineages (Pál et al., 2015; Vanderwoude et al., 2024).

5 Conclusion

This study demonstrated a diagnostic ceiling for agnostic k-mer hashing in *Salmonella* pangenomics, proving that information loss through hashing collisions can completely obscure resistance signals. In comparison, the gene-based models successfully managed to identify *gyrA_D87Y* as the primary anchor for resistance prediction, achieving a state-of-the-art Brier Score of 0.0059. This level of calibration is essential for clinical settings in ensuring that predicted probabilities correctly reflect clinical indicators and are not merely statistical artifacts. However, the performance decay observed with clonal, human-restricted lineages like *S. Typhi* still presents a significant challenge in genomic surveillance. Future frameworks must also incorporate efflux pump regulation and collateral sensitivity trade-offs, moving beyond static gene-presence models. Ultimately, this work provides a robust, interpretable path for transforming raw genomic data into high-fidelity, calibrated predictions, marking a necessary shift toward real-time, personalized antimicrobial stewardship.

References

1. Godfred Acheampong, Michael Owusu, Alex Owusu-Ofori, Isaac Osei, Nimako Sarpong, Augustina Sylverken, Hung-Jui Kung, Shu-Ting Cho, Chih-Horng Kuo, Se Eun Park, Florian Marks, Yaw Adu-Sarkodie, and Ellis Owusu-Dabo. Chromosomal and plasmid-mediated fluoroquinolone resistance in human Salmonella enterica infection in Ghana. *BMC Infectious Diseases*, 19:898, October 2019. ISSN 1471-2334. doi: 10.1186/s12879-019-4522-1.

2. admin. Antibiotic Resistance Profile of Salmonella Typhi Clinical Isolates in Urban Dhaka, February 2026.

3. Negar Ahvar, Mohsen Mohammadi, Pourya Zafari, and Mehran Shokri. Artificial Intelligence to Combat Antimicrobial Resistance: Comparative Benchmark of Predictive Methods. *InfoScience Trends*, 2(10):63–80, November 2025. ISSN 3041-9875. doi: 10.61882/ist.202502.10.05.

4. Sadanand Dangari Akshay, Srajana Nayak, Vijaya Kumar Deekshit, Anusha Rohit, and Biswajit Maiti. Differential expression of outer membrane proteins and quinolone resistance determining region mutations can lead to ciprofloxacin resistance in Salmonella Typhi. *Archives of Microbiology*, 205(4):136, March 2023. ISSN 1432-072X. doi: 10.1007/s00203-023-03485-0.

5. Kristina M. Angelo, Jared Reynolds, Beth E. Karp, Robert Michael Hoekstra, Christina M. Scheel, and Cindy Friedman. Antimicrobial Resistance Among Nontyphoidal Salmonella Isolated From Blood in the United States, 2003–2013. *The Journal of infectious diseases*, 214 (10):1565–1570, November 2016. ISSN 0022-1899. doi: 10.1093/infdis/jiw415.

6. Leila Azimi and Abdolaziz Rastegar Lari. Collateral sensitivity between aminoglycosides and beta-lactam antibiotics depends on active proton pumps. *Microbial Pathogenesis*, 112: 122–125, November 2017. ISSN 0882-4010. doi: 10.1016/j.micpath.2017.09.049.

7. Menuka Bhandari, Jelmer W. Poelstra, Michael Kauffman, Binta Varghese, Yosra A. Helmy, Joy Scaria, and Gireesh Rajashekara. Genomic Diversity, Antimicrobial Resistance, Plasmidome, and Virulence Profiles of Salmonella Isolated from Small Specialty Crop Farms Revealed by Whole-Genome Sequencing. *Antibiotics*, 12(11):1637, November 2023. ISSN 2079-6382. doi: 10.3390/antibiotics12111637.

8. Cornelia Caragea, Adrian Silvescu, and Prasenjit Mitra. Protein sequence classification using feature hashing. *Proteome Science*, 10(Suppl 1):S14, June 2012. ISSN 1477-5956. doi: 10.1186/1477-5956-10-S1-S14.

9. CDC. NARMS in Action. https://www.cdc.gov/narms/stories/index.html, 2024.

10. CDC. 2019 Antibiotic Resistance Threats Report. https://www.cdc.gov/antimicrobial-resistance/data-research/threats/index.html, April 2025.

11. Cheng-Hao Chang, Mingchih Chen, Ming-Hsien Tsai, Yen-Chun Huang, Hung-Hsiang Liou, Ben-Chang Shia, Chingying Liang, and Yu-Wei Fang. Machine Learning Prediction of Progression to Dialysis in Patients With Polycystic Kidney Disease: Population-Based Retrospective

Cohort Study. *JMIR Medical Informatics*, 14:e80343–e80343, March 2026. ISSN 2291-9694. doi: 10.2196/80343.

12. Man-Xia Chang, Jin-Fei Zhang, Yin-Huan Sun, Rong-Sheng Li, Xiao-Ling Lin, Ling Yang, Mark A. Webber, and Hong-Xia Jiang. Contribution of Different Mechanisms to Ciprofloxacin Resistance in Salmonella spp. *Frontiers in Microbiology*, 12:663731, May 2021. ISSN 1664-302X. doi: 10.3389/fmicb.2021.663731.

13. Kaichao Chen, Chen Yang, Ning Dong, Miaomiao Xie, Lianwei Ye, Edward Wai Chi Chan, and Sheng Chen. Evolution of Ciprofloxacin Resistance-Encoding Genetic Elements in Salmonella. *mSystems*, 5(6):e01234–20, December 2020. ISSN 2379-5077. doi: 10.1128/mSystems.01234-20.

14. Caitlin Collins and Xavier Didelot. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLOS Computational Biology*, 14(2):e1005958, February 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005958.

15. Cainã Max Couto da Silva. Probability Calibration in Machine Learning: Enhancing Model Usability, September 2024.

16. Rafaela Ferrari, Antonio Galiana, Rosa Cremades, Juan Carlos Rodríguez, Marciane Magnani, M.C.B. Tognim, Tereza C.R.M. Oliveira, and Gloria Royo. Plasmid-mediated quinolone resistance (PMQR) and mutations in the topoisomerase genes of Salmonella enterica strains from Brazil. *Brazilian Journal of Microbiology*, 44(2):651–656, October 2013. ISSN 1517-8382. doi: 10.1590/S1517-83822013000200046.

17. Miklos Fuzi. The fitness connection of antibiotic resistance. *Frontiers in Microbiology*, 16, April 2025. ISSN 1664-302X. doi: 10.3389/fmicb.2025.1556656.

18. Angelika Geroldinger, Lara Lusa, Mariana Nold, and Georg Heinze. Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—a simulation study. *Diagnostic and Prognostic Research*, 7:9, May 2023. ISSN 2397-7523. doi: 10.1186/s41512-023-00146-0.

19. Anna G. Green, Mahbuba Tasmin, Roger Vargas, and Maha R. Farhat. The structural context of mutations in proteins predicts their effect on antibiotic resistance, September 2025. ISSN 2692-8205.

20. Zuguang Gu. Complex heatmap visualization. *iMeta*, 1(3):e43, 2022. ISSN 2770-596X. doi: 10.1002/imt2.43.

21. Zuguang Gu, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. Circlize Implements and enhances circular visualization in R. *Bioinformatics*, 30(19):2811–2812, October 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu393.

22. Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, September 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw313.

23. Beatriz Guerra, Burkhard Malorny, Andreas Schroeter, and Reiner Helmuth. Multiple Resistance Mechanisms in Fluoroquinolone-Resistant Salmonella Isolates from Germany. *Antimicrobial Agents and Chemotherapy*, 47(6):2059, June 2003. ISSN 0066-4804. doi: 10.1128/AAC.47.6.2059.2003.

24. Kristine L. Haftorn, Julia Romanowska, Yunsung Lee, Christian M. Page, Per M. Magnus, Siri E. Håberg, Jon Bohlin, Astanand Jugessur, and William R. P. Denault. Stability selection enhances feature selection and enables accurate prediction of gestational age using only five DNA methylation sites. *Clinical Epigenetics*, 15:114, July 2023. ISSN 1868-7075. doi: 10.1186/s13148-023-01528-3.

25. Marcelo Hurtado and Vera Pancaldi. A new pipeline for cross-validation fold-aware machine learning prediction of clinical outcomes addresses hidden data-leakage in omics based 'predictors'., March 2026. ISSN 2692-8205.

26. Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Smooth Isotonic Regression: A New Method to Calibrate Predictive Models. *AMIA Summits on Translational Science Proceedings*, 2011:16–20, March 2011. ISSN 2153-4063.

27. Anthony Kelly. Extrinsic Trust as a Contractual Framework for Accountable AI in Health Care: Viewpoint. *Journal of Medical Internet Research*, 28(1):e83903, March 2026. doi: 10.2196/83903.

28. Heng Liang, Kun Jiang, Tong-An Yan, and Guang-Hui Chen. XGBoost: An Optimal Machine Learning Model with Just Structural Features to Discover MOF Adsorbents of Xe/Kr. *ACS Omega*, 6(13):9066–9076, April 2021. doi: 10.1021/acsomega.1c00100.

29. Matthew P. Moore, Mirjam Laager, Paolo Ribeca, and Xavier Didelot. KmerAperture: Retaining k-mer synteny for alignment-free extraction of core and accessory differences between bacterial genomes. *PLOS Genetics*, 20(4):e1011184, April 2024. ISSN 1553-7390. doi: 10.1371/journal.pgen.1011184.

30. NCBI. Home - Pathogen Detection - NCBI. https://www.ncbi.nlm.nih.gov/pathogens, 2026.

31. Csaba Pál, Balázs Papp, and Viktória Lázár. Collateral sensitivity of antibiotic-resistant microbes. *Trends in microbiology*, 23(7):401–407, July 2015. ISSN 0966-842X. doi: 10.1016/j.tim.2015.02.009.

32. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python, June 2018.

33. Awais Qureshi, Abdul Wahid, Shams Qazi, Muhammad K. Shahzad, and Hashir Moheed Kiani. DynaBiomeX: An Interpretable Dual-Strategy Deep Learning Framework for Architectural Noise Filtration in Sparse Longitudinal Microbiome Data, February 2026. ISSN 2692-8205.

34. Theodor A. Ross, Anna K. Pöntinen, Einar Holsbø, Ørjan Samuelsen, Kristin Hegstad, Michael Kampffmeyer, Jukka Corander, and Rebecca A. Gladstone. Machine learning-based lineage prediction from antimicrobial susceptibility testing phenotypes for Escherichia coli sequence type 131 clade C surveillance across infection types. *Microbial Genomics*, 12(1):001608, January 2026. ISSN 2057-5858. doi: 10.1099/mgen.0.001608.

35. Abeer Ahmed Rushdy, Mona Ibrahim Mabrouk, Ferialla Abdel-Hamid Abu-Sef, Zeinab Hassan Kheiralla, Said Mohamed Abdel -All, and Neveen Mohamed Saleh. Contribution of different mechanisms to the resistance to fluoroquinolones in clinical isolates of *Salmonella enterica*. *The Brazilian Journal of Infectious Diseases*, 17(4):431–437, July 2013. ISSN 1413-8670. doi: 10.1016/j.bjid.2012.11.012.

36. Morteza M. Saber and B. Jesse Shapiro. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microbial Genomics*, 6(3):e000337, February 2020. ISSN 2057-5858. doi: 10.1099/mgen.0.000337.

37. Hajime Senuma. Mmh3: A Python extension for MurmurHash3. *Journal of Open Source Software*, 10(105):6124, January 2025. ISSN 2475-9066. doi: 10.21105/joss.06124.

38. Aref Shariati, Maniya Arshadi, Mohammad Ali Khosrojerdi, Mostafa Abedinzadeh, Mahsa Ganjalishahi, Abbas Maleki, Mohsen Heidary, and Saeed Khoshnood. The resistance mechanisms of bacteria against ciprofloxacin and new approaches for enhancing the efficacy of this antibiotic. *Frontiers in Public Health*, 10:1025633, December 2022. ISSN 2296-2565. doi: 10.3389/fpubh.2022.1025633.

39. Michael J. Sikorski, Tracy H. Hazen, Sachin N. Desai, Susana Nimarota-Brown, Siaosi Tupua, Michelle Sialeipata, Savitra Rambocus, Danielle J. Ingle, Sebastian Duchene, Susan A. Ballard, Mary Valcanis, Sara Zufan, Jianguo Ma, Jason W. Sahl, Mailis Maes, Gordon Dougan, Robert E. Thomsen, Roy M. Robins-Browne, Benjamin P. Howden, Take K. Naseri, Myron M. Levine, and David A. Rasko. Persistence of Rare Salmonella Typhi Genotypes Susceptible to First-Line Antibiotics in the Remote Islands of Samoa. *mBio*, 13(5):e01920–22, September 2022. doi: 10.1128/mbio.01920-22.

40. Robert Slinger, Marc Desjardins, Anne E. McCarthy, Karam Ramotar, Peter Jessamine, Christiane Guibord, and Baldwin Toye. Suboptimal clinical response to ciprofloxacin in patients with enteric fever due to Salmonella spp. with reduced fluoroquinolone susceptibility: A case series. *BMC Infectious Diseases*, 4(1):36, September 2004. ISSN 1471-2334. doi: 10.1186/1471-2334-4-36.

41. Hüseyin Kerem Tolan, İrfan Aydın, Handan Tanyildizi-Kokkulunk, Mehmet Karakuş, Yüksel Akkaya, Osman Kaya, and Ferruh Kemal İşman. Machine Learning Model for Predicting Multidrug Resistance in Clinical Escherichia coli Isolates: A Retrospective General Surgery Study. *Antibiotics*, 14(10):969, October 2025. ISSN 2079-6382. doi: 10.3390/antibiotics14100969.

42. Jelly Vanderwoude, Sheyda Azimi, Timothy D. Read, and Stephen P. Diggle. The role of hypermutation and collateral sensitivity in antimicrobial resistance diversity of Pseudomonas aeruginosa populations in cystic fibrosis lung infection. *mBio*, 15(2):e03109–23, January 2024. doi: 10.1128/mbio.03109-23.

43. WHO. WHO warns of widespread resistance to common antibiotics worldwide. https://www.who.int/news/item/13-10-2025-who-warns-of-widespread-resistance-to-common-antibiotics-worldwide, 2025.

44. Marcus H. Wong, Edward W. Chan, Li Z. Liu, and Sheng Chen. PMQR genes oqxAB and aac(6′)Ib-cr accelerate the development of fluoroquinolone resistance in Salmonella typhimurium. *Frontiers in Microbiology*, 5, October 2014. ISSN 1664-302X. doi: 10.3389/fmicb.2014.00521.

45. Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257, November 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1891-0.

46. Douglas C. Wu, Jun Yao, Kevin S. Ho, Alan M. Lambowitz, and Claus O. Wilke. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, 19:510, July 2018. ISSN 1471-2164. doi: 10.1186/s12864-018-4869-5.

47. Bing Zhai, Chen Liao, Siddharth Jaggavarapu, Yuanyuan Tang, Thierry Rolling, Yating Ning, Tianshu Sun, Sean A. Bergin, Mergim Gjonbalaj, Edwin Miranda, N. Esther Babady, Oliver Bader, Ying Taur, Geraldine Butler, Li Zhang, Joao B. Xavier, David S. Weiss, and Tobias M. Hohl. Antifungal heteroresistance causes prophylaxis failure and facilitates breakthrough Candida parapsilosis infections. *Nature medicine*, 30(11):3163–3172, November 2024. ISSN 1078-8956. doi: 10.1038/s41591-024-03183-4.

48. Zengfeng Zhang, Hang Zhao, and Chunlei Shi. Clonal Spread and Genetic Mechanisms Underpinning Ciprofloxacin Resistance in Salmonella enteritidis. *Foods*, 14(2):289, January 2025. ISSN 2304-8158. doi: 10.3390/foods14020289.

49. Wenting Zhao, Pei Sun, Wei Li, and Linping Shang. Machine Learning-Based Prediction Model for Multidrug-Resistant Organisms Infections: Performance Evaluation and Interpretability Analysis. *Infection and Drug Resistance*, 18:2255–2269, May 2025. ISSN 1178-6973. doi: 10.2147/IDR.S459830.