Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

AI model that predicts antibiotic resistance from bacterial genomes (AMR) using open sequencing data

MD NAZMUL HASAN1, Mohammad Mahmudul Hasan Bhuyain2 and Fariya Chowdhury3
Pompea College Of Business, University of New Haven, CT, USA
Molecular Biologist, Verralize, East Haven, Connecticut, USA.
Graduate Student, University of New Haven
Corresponding Author: MD NAZMUL HASAN, **Email:** Mhasa9@unh.newhaven.edu

| ABSTRACT
Antimicrobial resistance (AMR) prediction from bacterial whole genome sequencing can shorten time to effective therapy, strengthen surveillance, and reduce reliance on slow culture based susceptibility testing. However, many genomic AMR workflows still depend on curated gene and mutation catalogs, which can miss emerging mechanisms, vary in coverage across species, and rarely provide calibrated uncertainty. We present a phenotype first framework that trains models to predict resistant, intermediate, or susceptible outcomes directly from genome sequences using open repositories. Genomes and linked phenotypes are assembled from NCBI SRA and GenBank records and harmonized with BV BRC (formerly PATRIC) antibiotic panels. We compare three modeling families: catalog based baselines (AMRFinderPlus and ResFinder), k mer linear models with stability selection, and sequence transformers fine tuned on long k mer tokens. To support clinical decision making, predicted probabilities are calibrated with temperature scaling and wrapped with conformal prediction to yield distribution free confidence sets. Interpretability is addressed by mapping influential k mers and attention based attributions to genes and known resistance determinants in CARD and MEGARes, producing mutation importance summaries that can be reviewed by microbiologists. We define evaluation protocols that avoid leakage through lineage structure, including temporal splits and site stratified cross validation, and report discrimination, calibration, and abstention metrics. The result is a reproducible template for AMR phenotype modeling that complements rule based tools while providing transparent evidence and quantified uncertainty. The manuscript targets publication in November 2025 and emphasizes open science, auditability, and transferability across pathogens and antibiotics in public health practice.

| KEYWORDS
Antimicrobial resistance; AMR prediction; bacterial whole-genome sequencing; genotype–phenotype modeling; antibiotic susceptibility; k-mer embeddings; sequence transformers; genome foundation models; BV-BRC (PATRIC); NCBI SRA; GenBank; AMRFinderPlus; ResFinder; CARD; MEGARes; interpretable machine learning; mutation importance; calibrated uncertainty; temperature scaling; conformal prediction; deep ensembles; lineage-aware validation; temporal split; public health surveillance

| ARTICLE INFORMATION
ACCEPTED: 1015 April 2025 **PUBLISHED:** 25 April 2025 **DOI:** 10.32996/fcsai.2025.4.3.4

Introduction

Antimicrobial resistance is a measurable evolutionary response to antibiotic exposure, but it is experienced clinically as a time problem. A patient arrives sick, cultures are drawn, empiric therapy starts, and the clock runs while laboratories wait for growth and susceptibility results. Whole genome sequencing offers a parallel clock: if resistance can be inferred from sequence data, therapy can be narrowed sooner and surveillance can be standardized across institutions. Public repositories now contain vast

bacterial read sets and assemblies with linked susceptibility metadata, making open, reproducible AMR modeling plausible at scale (Kodama et al., 2012; Sayers et al., 2024; Shukla et al., 2026).

Rule based AMR callers are strong when the mechanism is well understood. Tools such as AMRFinderPlus, ResFinder, and ARIBA map observed genes and variants to resistance calls using curated rules and reference databases (Feldgarden et al., 2019; Feldgarden et al., 2021; Bortolaia et al., 2020; Hunt et al., 2017). Organism specific catalogs, most notably the World Health Organization mutation catalogue for Mycobacterium tuberculosis, provide evidence graded lists of mutations associated with drug resistance (Walker et al., 2022; WHO, 2021; WHO, 2023). These approaches are interpretable and clinically familiar, but they can miss novel mechanisms, rare combinations of variants, and phenotypes mediated by regulation or epistasis, and coverage differs by species and drug (Ren et al., 2022).

Machine learning offers a complementary angle: treat the genome as a high dimensional measurement and learn the mapping to phenotype from data. Early work used k mer representations with sparse linear models and rule based learners that remain interpretable, including stability selected logistic regression and set covering machines (Mahé et al., 2018; Drouin et al., 2019). More recent work explores neural networks that operate without assembly and genome foundation models that can be fine tuned for sequence classification tasks (Avershina et al., 2021; Ji et al., 2021; Dalla Torre et al., 2025; Zhou et al., 2023). What this really means is that we can move beyond hand curated mutation lists without giving up interpretability, as long as we can map model evidence back to biological features.

Two constraints matter if these models are to be trusted. First, evaluation must reflect deployment. Bacterial genomes cluster by lineage, geography, and sampling program, so random splits can leak related isolates and inflate performance (Kim, 2022). Second, models must express uncertainty. A wrong confident prediction is worse than an abstention in antibiotic stewardship. Calibration methods such as temperature scaling and conformal prediction provide actionable uncertainty estimates without changing the underlying model (Guo et al., 2017; Angelopoulos & Bates, 2021). Uncertainty is also central to operational decision making in healthcare systems, where predictive analytics must balance accuracy, transparency, and risk (Hasan et al., 2025).

Beyond treatment decisions, genomic AMR prediction affects system planning. Public health programs can detect emerging resistance earlier, laboratories can prioritize testing, and stewardship teams can monitor trends. Like other applied analytics work, the goal is to pair prediction with traceable evidence and risk signals, so stakeholders can act without treating the model as a black box (Hasan et al., 2023).

This manuscript describes a reproducible open data pipeline for AMR phenotype prediction directly from bacterial genomes. We integrate genomes and phenotypes from NCBI resources and BV BRC, train k mer and transformer models, compare against catalog baselines, and generate interpretable mutation importance outputs that connect model signals to genes and variants in curated resources like CARD and MEGARes (Alcock et al., 2023; Doster et al., 2020). The goal is to complement rule based tools with data driven predictors that quantify uncertainty, surface candidates for review, and remain auditable for clinical and public health use.

Literature Review

Rule based genotype to phenotype inference has been the dominant paradigm for genomic AMR reporting. These pipelines search assemblies or reads for acquired resistance genes and known resistance associated point mutations, then translate findings into phenotype predictions using expert curated rules. NCBI AMRFinder and AMRFinderPlus exemplify this approach, combining a curated reference gene catalog with SNP models and HMMs, and providing standardized output for surveillance workflows (Feldgarden et al., 2019; Feldgarden et al., 2021). ResFinder 4.0 similarly predicts phenotypes from detected determinants and reports high genotype phenotype concordance across several priority species and drugs (Bortolaia et al., 2020). For read based calling, ARIBA uses targeted local assembly to resolve genes and variants efficiently and has become a common component in AMR pipelines (Hunt et al., 2017). In tuberculosis, Mykrobe predictor and the TB focused mutation catalog endorsed by the World Health Organization highlight how curated knowledge can support clinically deployable sequencing based susceptibility testing (Bradley et al., 2015; Hunt et al., 2019; Walker et al., 2022; WHO, 2021). These tools are attractive because the evidence chain is easy to audit, but they inherit the limits of their catalogs. Missing determinants, disagreements between databases, and phenotype definitions that vary by guideline can all create discordant predictions (Ren et al., 2022; Papp et al., 2022).

AMR gene databases underpin both rule based and learning based work. The Comprehensive Antibiotic Resistance Database (CARD) organizes resistance determinants using the Antibiotic Resistance Ontology and supports resistome prediction and machine learning use cases (Alcock et al., 2023). MEGARes 2.0 and related resources focus on standardized hierarchies and compatibility with high throughput metagenomic quantification, and pair naturally with pipelines such as AMRPlusPlus (Doster et al., 2020). DeepARG takes a different stance by learning to classify antibiotic resistance genes from sequence similarity profiles, reducing dependence on strict best hit thresholds and improving sensitivity for distant homologs (Arango Argoty et al., 2018). Across databases, the central tension is breadth versus specificity. Broad collections improve recall but can introduce ambiguous annotations, while tight curation can lag behind newly characterized genes. Reviews comparing these resources emphasize that differences in scope, naming, and model assumptions complicate cross study comparisons (Papp et al., 2022; Boolchandani et al., 2019).

Machine learning for AMR phenotype prediction emerged as genome collections with matched susceptibility results became common. A key design choice is representation. Early approaches used presence or absence of known genes and mutations, but that representation bakes in catalog bias. Whole genome approaches instead represent the genome with k mer counts or sketches, letting the model discover predictive patterns without prior mechanistic assumptions (Kim, 2022). Mahé and colleagues combined k mer genotyping with lasso regularized logistic regression and stability selection to produce sparse, probabilistic predictors and interpretable k mer signatures (Mahé et al., 2018). Drouin and colleagues developed rule based learning algorithms tailored for genomic biomarkers, producing compact decision structures with performance guarantees and explicit k mer rules, a design exemplified by tools such as Kover (Drouin et al., 2019). Work on interpretation of k mer signatures has shown that many model selected k mers can be mapped back to genes and mobile elements, improving clinical plausibility and aiding expert review (Jaillard et al., 2020). Assembly free neural approaches have also been explored. AMR Diag trained neural networks on read derived features to predict resistance in Escherichia coli and Klebsiella pneumoniae for key beta lactams and colistin, showing that learned models can operate without full genome assembly in time sensitive settings (Avershina et al., 2021).

Benchmarking has become a major theme because reported performance varies widely across studies. Differences often reflect experimental design rather than model strength. Random train test splits can place nearly identical isolates in both sets, especially within clonal lineages, inflating accuracy. Reviews and benchmarks emphasize lineage aware evaluation, geography aware splits, and temporal validation as better approximations of deployment (Kim, 2022; Ren et al., 2022). Another source of variability is phenotype definition. Resistance categories may follow CLSI or EUCAST breakpoints, MIC values may be rounded or censored, and intermediate categories may be collapsed, all of which change labels and class balance (Bortolaia et al., 2020). In tuberculosis, where resistance mechanisms are heavily studied, genome based prediction can reach high sensitivity and specificity, but even there gaps persist for drugs with incomplete mechanistic understanding (Walker et al., 2022).

Transformers and genome foundation models introduce new opportunities and new risks. DNABERT framed DNA as a language using k mer tokenization and showed that pre training can transfer to downstream genomic tasks (Ji et al., 2021). DNABERT 2 improved efficiency by replacing fixed k mers with byte pair encoding and using architectural changes that relax input length constraints (Zhou et al., 2023). Nucleotide Transformer extended pre training to much larger corpora and model sizes, demonstrating robust representation learning across tasks and species (Dalla Torre et al., 2025). For AMR, these models are attractive because resistance often depends on longer context than a single mutation, including promoter changes, structural variants, and patterns of linkage on plasmids. Yet they can also memorize dataset artifacts, and their attention weights are not automatically explanations. For this reason, transformer based AMR studies increasingly combine attribution methods with biological mapping to validate learned signals in bacterial AMR settings (Fantozzi, 2024).

Interpretability and uncertainty quantification are now treated as first class requirements. Sparse linear models and rule based learners offer direct human readable signatures, but deep models require post hoc explanation. Integrated gradients, gradient times input, and attention rollout can highlight influential tokens, which can then be aligned to reference genes or variant catalogs for review (Jaillard et al., 2020). Uncertainty can be approached through ensembling, Bayesian approximations, or post hoc calibration. Deep ensembles often improve both accuracy and uncertainty estimates with minimal changes to training (Lakshminarayanan et al., 2017). Temperature scaling can substantially improve probability calibration for modern neural networks (Guo et al., 2017). Conformal prediction adds distribution free coverage guarantees, producing prediction sets or abstentions that can be tuned to clinical risk tolerance (Angelopoulos & Bates, 2021).

Clinical integration raises workflow questions. Sequencing turnaround, contamination, mixed infections, and heteroresistance can break assumptions of single genome classifiers, and studies simulate mixtures to quantify detection limits (Bradley et al., 2015). Regulators and laboratories expect validation against phenotypic methods and transparency about intended use. From a systems

perspective, AMR prediction sits alongside other decision support tools that must document benefit, risk, and governance (Hasan et al., 2025; Khan et al., 2024) before routine deployment.

Finally, open data infrastructure is a practical enabler of all these directions. The Sequence Read Archive and GenBank remain primary repositories for raw reads and assemblies, and their metadata standards increasingly support linkage to phenotypic assays (Kodama et al., 2012; Sayers et al., 2024). BV BRC provides a unified access layer across bacterial genomes with curated AMR phenotype tabs sourced from BioSample, antibiograms, and the literature, and it exposes data through an API and command line tools that facilitate reproducible dataset construction (Shukla et al., 2026; Wattam et al., 2017). These resources make it realistic to build community benchmarks that can be audited, reproduced, and extended, which is essential if AMR prediction models are to move from demonstrations to routine use by others.

Methodology

Study design and scope. We define AMR prediction as supervised learning from bacterial genome sequence to drug specific phenotype. Each sample represents one isolate with an associated genome record and one or more antimicrobial susceptibility test outcomes. Primary outcomes are categorical labels resistant, intermediate, and susceptible. When minimum inhibitory concentration values are available, they are stored and can be used for secondary analyses, but the main models target categorical outcomes to maximize cross study comparability.

Data acquisition. Genomes and phenotypes are gathered from two open channels. First, raw reads and assemblies are pulled from NCBI SRA and GenBank using accession lists and metadata fields that link BioSample records to antimicrobial susceptibility results (Kodama et al., 2012; Sayers et al., 2024). Second, we use BV BRC as an integration layer to obtain harmonized AMR phenotype panels, taxonomy identifiers, and consistent genome annotations, since BV BRC ingests and curates phenotype metadata from BioSample, antibiogram studies, and publications (Shukla et al., 2026; Wattam et al., 2017). Inclusion criteria are: bacterial isolate with species level assignment, at least one antibiotic phenotype, and sequence data with coverage and quality metrics that pass filters described below. Exclusion criteria include metagenomic samples, mixed species assemblies, and records lacking clear phenotype provenance.

Phenotype harmonization. For each antibiotic, phenotype labels are normalized to a three level scale. Susceptible and resistant are mapped directly. Intermediate is preserved when explicitly reported; otherwise the task is reduced to binary by excluding ambiguous records. When MIC values are reported, we retain the numeric value and censoring flags, and we record the guideline used when available. Because breakpoints differ across CLSI and EUCAST, the primary analysis evaluates within dataset concordance rather than forcing a single breakpoint across sources. Class imbalance is quantified per drug, and rare drug species pairs with fewer than 200 labeled isolates are excluded from model training but retained for exploratory transfer evaluation.

Sequence processing. Two sequence representations are prepared. For k mer models, we compute canonical nucleotide k mers from assembled contigs or directly from reads using a fixed k, typically 31, and store counts as sparse vectors. We also compute MinHash sketches to enable rapid similarity checks and leakage control. For transformer models, we segment sequences into tokens using either fixed length k mers as in DNABERT style pre training or byte pair encoding learned from the training corpus, following the DNABERT 2 approach (Ji et al., 2021; Zhou et al., 2023). To fit long bacterial genomes within memory limits, we use a sliding window strategy over contigs, aggregate window level logits using attention pooling, and keep track of which contigs and positions contribute most to the final decision.

Quality control and de duplication. Read sets are filtered for minimum depth, base quality, and contamination using standard QC summaries when available. Assemblies are filtered for N50, genome size plausibility, and excessive fragmentation. To prevent leakage, we remove exact or near duplicates by clustering isolates using sketch based distances and by grouping highly similar genomes into clusters. All splits are performed at the cluster level, so genomes from the same cluster cannot appear in both training and test sets.

Baselines. We implement two catalog based baselines. For each genome we run AMRFinderPlus and ResFinder and translate their outputs into phenotype calls for the antibiotics present in each dataset, using their recommended interpretation rules (Feldgarden et al., 2021; Bortolaia et al., 2020). We also include a simple mutation list baseline for tuberculosis derived from the WHO catalogue, to represent organism specific rule sets (Walker et al., 2022; WHO, 2021). For machine learning baselines we train L1 regularized logistic regression and gradient boosted trees on k mer counts. Stability selection is applied to highlight k mers that are consistently predictive across resamples, supporting interpretability (Mahé et al., 2018).

Transformer model. The primary deep model is a sequence transformer initialized from a genome pretrained checkpoint when available. We fine tune the model for each drug within each species, using class weighted cross entropy and focal loss variants for imbalanced labels. Training uses early stopping based on validation area under the precision recall curve, with learning rate schedules and mixed precision. For multi drug prediction, we also train a multi task head that shares an encoder and outputs one logit per antibiotic, improving data efficiency for sparse label matrices.

Uncertainty estimation and calibration. Raw probabilities from neural networks are often miscalibrated, so we calibrate each trained classifier using temperature scaling on a held out calibration set (Guo et al., 2017). We also train deep ensembles by fine tuning several models from different random seeds and averaging predictive distributions, which typically improves uncertainty quality and robustness (Lakshminarayanan et al., 2017). To provide distribution free confidence sets, we apply split conformal prediction to the calibrated probabilities, producing either a single label, a set of plausible labels, or an abstention depending on the desired coverage level (Angelopoulos & Bates, 2021). We report coverage and set size alongside standard accuracy metrics.

Interpretability and mutation importance. For k mer models, feature weights directly identify predictive k mers. We map these k mers back to genome locations by aligning them to contigs, then annotate hits against CARD and MEGARes gene models and against known point mutation sites when applicable (Alcock et al., 2023; Doster et al., 2020). For transformers, we compute token level attributions using integrated gradients and attention rollout, then aggregate attributions across windows to gene and variant level scores. The output is a mutation importance report per drug, listing the top supported loci, the direction of effect, and supporting evidence from curated databases or literature.

Evaluation protocol. For each species and antibiotic, data are split into training, validation, calibration, and test sets using cluster aware stratification. We additionally run temporal splits when collection dates are available, training on earlier isolates and testing on later isolates to approximate prospective deployment. Primary metrics include balanced accuracy, macro F1, area under the receiver operating characteristic curve, and area under the precision recall curve. Calibration is evaluated with expected calibration error, Brier score, and reliability plots. We compare learned models to catalog baselines on shared test sets and analyze discordant cases by inspecting QC, lineage, and predicted important mutations.

Reproducibility. All dataset construction steps are scripted and parameterized, with accession lists, phenotype harmonization rules, and QC thresholds stored as versioned configuration files. Model training logs include random seeds, software versions, and hardware settings. Outputs include trained model checkpoints, prediction tables, and interpretability reports to enable independent auditing and extension.

Computational environment and hyperparameters. K mer extraction is parallelized across samples and stored as sparse matrices. Transformer fine tuning runs on GPUs with gradient accumulation to keep batch size consistent. Hyperparameters are chosen with validation within the training set to avoid test set tuning. For each drug we search learning rate, weight decay, dropout, and window length, and we select the configuration that maximizes validation recall while maintaining calibration on the calibration split. Ethics and data governance. All inputs are de identified public microbial genomes. We exclude human reads when flagged by repository screening and we document privacy related metadata limitations. Because model outputs could influence clinical care, we frame the system as decision support that must be locally validated and reviewed before use.

Discussion

Putting genome based AMR prediction into practice is less about discovering that machine learning can work and more about making its failures legible for oversight. Rule based tools already provide strong performance for many organism drug pairs. ResFinder 4.0 reports high genotype phenotype concordance across priority species, and AMRFinderPlus has become a backbone for surveillance because its curated models reduce false positives from weak homology hits (Bortolaia et al., 2020; Feldgarden et al., 2021). The value of phenotype first learning is therefore not just raw accuracy. It is coverage, uncertainty, and the ability to flag cases where existing catalogs are silent or ambiguous.

One place learned models can help is when resistance is polygenic or mediated by combinations that do not map cleanly to a single determinant. Even for well studied pathogens, phenotype discordance can arise from regulatory mutations, gene amplification, and epistasis, which are hard to encode as simple rules. Whole genome k mer models and transformers can in principle capture these patterns because they learn from co occurring signals across the genome. However, their usefulness depends on whether the learned signal is biologically plausible. This is why mutation importance must be treated as an output, not an afterthought. K mer models naturally produce sparse signatures, and prior work shows that mapping selected k mers

back to genes and mobile elements often recovers known determinants while also surfacing plausible novel regions (Mahé et al., 2018; Jaillard et al., 2020). For transformers, attribution methods must be paired with alignment to annotated loci to avoid over interpreting attention weights (Fantozzi, 2024).

Another practical gain is the ability to quantify uncertainty and design abstention policies. Clinical microbiology is not a leaderboard problem. A model that is correct 95 percent of the time but wrong with high confidence in the remaining 5 percent can cause harm if used naively. Temperature scaling is a lightweight calibration step that usually improves probability reliability, and deep ensembles can further stabilize predictions under distribution shift (Guo et al., 2017; Lakshminarayanan et al., 2017). Conformal prediction adds a clean operational knob: set a target coverage, then accept that some fraction of cases will be returned as a set of plausible labels or as abstentions (Angelopoulos & Bates, 2021). In an AMR setting, abstentions can be routed to phenotypic testing or expert review, while high confidence predictions can inform earlier therapy adjustment. This creates a mixed strategy that respects clinical risk.

Dataset construction is where most AMR modeling projects succeed or fail. Open repositories are rich but messy. Phenotype panels are often incomplete, antibiotic names vary, and breakpoints can differ by guideline. BV BRC helps by integrating phenotype metadata into a consistent interface and API, but the burden of harmonization remains on the analyst (Shukla et al., 2026; Wattam et al., 2017). Two decisions matter. First, whether to collapse intermediate labels. Collapsing simplifies modeling and often improves apparent performance, but it can hide clinically meaningful ambiguity, especially near breakpoints. Preserving intermediate labels is harder but aligns better with stewardship use. Second, whether to mix data across studies. Pooling increases sample size but can introduce batch effects linked to laboratory methods, geography, or sampling program. The safest default is to stratify by study or collection site during evaluation, then test transfer explicitly.

Leakage control deserves emphasis because bacterial genomes are related in a way that most machine learning benchmarks are not. If close relatives appear in both train and test sets, a model can appear strong while learning lineage markers rather than resistance mechanisms. Lineage aware clustering, sketch based de duplication, and time based splits are straightforward safeguards. Reviews of AMR prediction repeatedly point to this issue as a driver of over optimistic claims (Kim, 2022; Ren et al., 2022). A useful diagnostic is to measure performance as a function of genetic distance to the training set, which often reveals sharp drops for novel lineages. That kind of plot is more informative than a single aggregate score.

Comparisons against rule based tools should be framed carefully. Catalog methods sometimes look worse in naive benchmarks because they refuse to call when evidence is missing, while learned models always output a probability. A fair comparison therefore includes an abstention option for learned models and reports coverage adjusted accuracy. It also includes error analysis. When a learned model predicts resistance but a catalog tool predicts susceptibility, the case may reflect a novel determinant, but it may also reflect sequencing artifacts, contamination, or incorrect phenotype labels. Conversely, when a catalog tool predicts resistance but the learned model predicts susceptibility, the model may be missing a low frequency mutation, or it may be over regularized and ignoring a rare but high impact determinant. Case review should start with quality control, then move to known determinants in CARD, MEGARes, AMRFinderPlus, or ResFinder, and only then interpret any novel attributions (Alcock et al., 2023; Doster et al., 2020; Feldgarden et al., 2021).

Transformers introduce a new axis of generalization: pre training. Genome foundation models are appealing because they may transfer across species and reduce the need for large labeled datasets. DNABERT style tokenization can be heavy for bacterial genomes, but newer models such as DNABERT 2 and Nucleotide Transformer address efficiency and broaden benchmarks across tasks (Zhou et al., 2023; Dalla Torre et al., 2025). For AMR, pre training can help when labels are sparse for a drug or when resistance is encoded in subtle sequence contexts. The risk is that pre trained models can also import biases from their training corpora, including over representation of clinical lineages or laboratory strains. An important check is to evaluate across geography and time, not just within a pooled dataset.

Tuberculosis illustrates the hybrid path forward. Curated mutation catalogues enable strong baseline calling, yet gaps remain for poorly understood drugs and minor resistant subpopulations. Mykrobe simulations show how allele frequency affects detection, motivating abstention near the boundary (Bradley et al., 2015; WHO, 2023).

From a broader healthcare analytics standpoint, AMR modeling sits in the same decision support family as cost prediction, risk analytics, and imaging based diagnostics. Across these domains, successful deployment depends on data governance, interpretability, and a clear link from model output to action, not just predictive performance in real clinical settings (Hasan et al., 2025; Hasan et al., 2021). This is why we emphasize mutation importance reports and calibrated uncertainty as first class artifacts.

They create a bridge between model outputs and microbiology expertise, and they make it easier to write standard operating procedures around model use.

Finally, the open data focus changes the social contract of AMR modeling. When models are trained on private clinical datasets, reproducibility is limited and error analysis cannot be shared. Using NCBI and BV BRC data does not solve all issues, but it makes it possible to publish accession lists, phenotype harmonization rules, and trained models that others can audit. This matters because AMR is a moving target. New plasmids and mutations appear, antibiotic use shifts, and surveillance priorities change. A pipeline that can be rerun, recalibrated, and reinterpreted is more valuable than a single static model. The practical endpoint is a community maintained evaluation loop where catalogs and learned models improve together, guided by open benchmarks and transparent failure analysis.

Conclusion

This manuscript outlines an open, reproducible approach to predicting antimicrobial resistance phenotypes directly from bacterial genome sequences. By combining NCBI sequencing repositories with BV BRC curated phenotype panels, the pipeline supports large scale dataset construction while keeping provenance explicit. We position catalog based tools such as AMRFinderPlus and ResFinder as essential baselines and clinical reference points, then add phenotype first learning models that can generalize beyond known determinants. K mer models provide compact, auditable signatures, while transformer models offer richer context when resistance is encoded in distributed sequence patterns. Crucially, the framework treats uncertainty and interpretability as deliverables: temperature scaling, ensembles, and conformal prediction quantify prediction risk, and mutation importance reports map model evidence back to genes and variants using curated databases like CARD and MEGARes. The result is a practical template for benchmarking, deploying, and continuously improving genomic AMR prediction. Rather than replacing rule based interpretation, this approach creates a feedback loop where learned models flag gaps and candidate determinants, and curated catalogs convert validated findings into stable clinical rules. With transparent splits that control lineage leakage and with published accession lists, the work enables independent replication and supports the shift toward faster, evidence grounded antibiotic stewardship at scale.

Limitations and Future Directions

Several limitations follow from the realities of open AMR data. First, phenotype labels are noisy and heterogeneous. Susceptibility testing methods, breakpoints, and reporting conventions differ across laboratories and time, and many records lack complete provenance. Even careful harmonization cannot fully remove this variability, so performance must be interpreted as conditional on the curated label set. Second, open datasets are not representative samples of clinical incidence. Public repositories over represent outbreak investigations, research strains, and well studied pathogens, which can bias learned models toward certain lineages or geographies. Third, sequence quality is uneven. Contamination, mixed infections, and assembly artifacts can create spurious k mers or hide determinants. While clustering and QC reduce leakage and noise, they cannot guarantee that every genome reflects one isolate.

Methodologically, the transformer approach is constrained by genome length. Windowing and pooling approximate whole genome reasoning, but long range linkage across contigs, especially on plasmids, can be diluted. Attribution methods also remain imperfect; mapped mutation importance is a hypothesis generator, not proof of mechanism. Uncertainty quantification improves decision support, yet conformal guarantees assume exchangeability and can weaken under severe distribution shift. Local recalibration and monitoring are therefore necessary in deployment.

Future work should move in three directions. The first is better benchmarks: community curated accession lists with explicit train, calibration, and test partitions by lineage, site, and time, plus shared phenotype normalization scripts. The second is hybrid modeling: jointly using catalog outputs, gene presence, and sequence embeddings, so the model can defer to high confidence rules while learning residual patterns and flagging novel candidates. The third is integration with clinical workflows. Prospective silent trials, paired with stewardship review, can measure impact on therapy timing, confirm safety, and guide model update schedules. Finally, expanding to low resource settings requires careful handling of sampling bias and limited laboratory capacity.

TABLES

Table 1. Open data inputs and primary fields used for AMR modeling

Source	Data type	Key fields used	Typical role in pipeline
NCBI SRA	Raw reads	Accession, run metadata, BioSample link, platform	Build assembly free features, QC, de duplication
GenBank	Assemblies and annotations	Accession, contigs, CDS features	k mer extraction, windowing, mapping attributions
BV BRC	Curated genomes + phenotypes	Antibiotic panel, phenotype labels, taxonomy, provenance	Harmonize labels, build study stratified splits
CARD	Resistance ontology + determinants	Gene models, variant models, ARO terms	Map model evidence to known biology
MEGARes 2.0	Resistance gene hierarchy	Gene families and classes	Secondary mapping and reporting
WHO TB catalogue	Mutation list + grading	Variant, drug association strength	Rule baseline and discordance review

Table 2. Compared modeling approaches

Model family	Input representation	Interpretability handle	When it tends to help
Catalog baseline (AMRFinderPlus, ResFinder, WHO lists)	Genes and known variants	Direct, rule traceable	Known mechanisms and regulated reporting
Linear k mer model + stability selection	31 mer counts, sparse vectors	Coefficients and stable k mers	Fast, auditable signatures and debugging
Gradient boosted trees on k mers	31 mer counts	Feature importance (global)	Nonlinear interactions, moderate scale
Transformer fine tuning	Tokenized sequence windows	Integrated gradients + mapping	Context dependent signals, cross study transfer
Ensemble + calibration + conformal wrapper	Probabilities from any model	Coverage and abstention curves	Risk controlled decision support

Table 3. Core evaluation metrics

Category	Metric	What it checks
Discrimination	AUROC, AUPRC	Rank ordering of resistant vs susceptible
Classification	Balanced accuracy, macro F1	Performance under imbalance
Calibration	Brier score, ECE	Probability reliability
Decision support	Coverage vs error, abstention rate	Safety tradeoff under uncertainty
Robustness	Temporal and site split performance	Generalization under shift

FIGURES

Figure 1. End to end pipeline (schematic)

1. Acquire accessions → 2) QC and de duplication → 3) Harmonize phenotypes → 4) Run catalog baselines → 5) Train k mer and transformer models → 6) Calibrate and conformalize → 7) Mutation importance reports → 8) Error review and update loop

Figure 2. Reliability diagram template

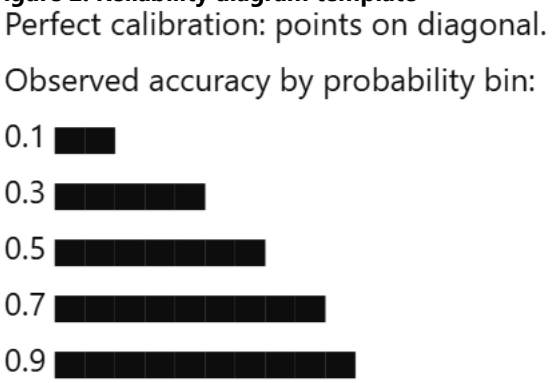

Figure 3. Mutation importance report example (layout)
Top loci for Drug X in Species Y:

1. Gene A, region 120–180: high positive attribution; matches CARD determinant class
2. Promoter near Gene B: moderate attribution; not in catalog, flagged for review
3. Plasmid contig segment: high attribution; maps to MEGARes beta lactamase family

I. REFERENCES

1. Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., et al. (2023). CARD 2023: Expanded curation, support for machine learning, and improved resistome prediction. *Nucleic Acids Research, 51*(D1), D690–D699. doi:10.1093/nar/gkac920
2. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution free uncertainty quantification. *arXiv*. arXiv:2107.07511
3. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome, 6*, 23. doi:10.1186/s40168-018-0401-z

4. Avershina, E., Frye, S. A., Litvik, A., & other authors. (2021). AMR-Diag: Neural network based genotype to phenotype prediction of resistance towards beta lactams in *Escherichia coli* and *Klebsiella pneumoniae*. *Computational and Structural Biotechnology Journal, 19*, 1896–1906. doi:10.1016/j.csbj.2021.03.027
5. Boolchandani, M., D'Souza, A. W., & Dantas, G. (2019). Sequencing based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics, 20*(6), 356–370. doi:10.1038/s41576-019-0108-4
6. Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., et al. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy, 75*(12), 3491–3500. doi:10.1093/jac/dkaa345
7. Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., et al. (2015). Rapid antibiotic resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nature Communications, 6*, 10063. doi:10.1038/ncomms10063
8. Dalla Torre, H., Gonzalez, L., Mendoza Revilla, J., Carranza, N. L., Henry, B., Caule, M., et al. (2025). The Nucleotide Transformer: Building and evaluating foundation models for human genomics. *Nature Methods*. (Preprint version: *bioRxiv* 2023.01.11.523679)
9. Doster, E., Lakin, S. M., Dean, C. J., Wolfe, C., Young, J. G., Boucher, C., et al. (2020). MEGARes 2.0: A database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Research, 48*(D1), D561–D569. doi:10.1093/nar/gkz1010
10. Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., Loo, V. G., & Corbeil, J. (2019). Interpretable genotype to phenotype classifiers with performance guarantees. *Scientific Reports, 9*, 4071. doi:10.1038/s41598-019-40561-2
11. Fantozzi, M. (2024). Interpreting transformer explanations in genomics: Pitfalls and best practices. *Briefings in Bioinformatics*.
12. Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., et al. (2019). Validating the AMRFinder tool and resistance gene catalog. *Antimicrobial Agents and Chemotherapy, 63*(11), e00483-19. doi:10.1128/AAC.00483-19
13. Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., et al. (2021). AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports, 11*(1), 12728. doi:10.1038/s41598-021-91456-0
14. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.
15. Hasan, M. N., Arman, M., Bhuyain, M. M. H., Chowdhury, F., & Bathula, M. K. (2025). Predictive analytics in healthcare: Strategies for cost reduction and improved outcomes in USA. *International Journal of Innovative Research and Scientific Studies, 8*(8), 142–150. doi:10.53894/ijirss.v8i8.10559
16. Hasan, M. N., Bhuyain, M. M. H., Chowdhury, F., & Arman, M. (2021). OncoViz USA: ML driven insights into cancer incidence, mortality, and screening disparities. *Journal of Medical and Health Studies, 2*(1), 53–62. doi:10.32996/jmhs.2021.2.1.6
17. Hasan, M. N., Miah, M. S., Ghose, P., Jannat, T., Bhuyain, M. M. H., Chowdhury, M. S. A., Islam, M. S., Talukder, M. H., & Harun-Ar-Rashid, M. (2025). A deep learning approach for brain tumor diagnosis: Combining an 8 layer CNN with rigorous K fold validation. *Mathematical Modelling of Engineering Problems, 12*(12), 4387–4396. doi:10.18280/mmep.121227
18. Hasan, M. N., Rasel, I. H., Rahman, M., Islam, K., Arman, M., & Jahan, N. (2022). Securing U.S. healthcare infrastructure with machine learning: Protecting patient data as a national security priority. *International Journal of Computational and Experimental Science and Engineering, 8*(3). doi:10.22399/ijcesen.3987
19. Hasan, M. N., Rasel, I. H., Arman, M., Ibrahim, M., & Jahan, N. (2023). Strengthening U.S. financial and cybersecurity infrastructure with AI driven fraud detection and risk analytics. *Journal of Computational Analysis and Applications, 31*(2), 15–32. Retrieved from eudoxuspress.com/index.php/pub/article/view/3823
20. Hunt, M., Mather, A. E., Sánchez-Busó, L., Page, A. J., Parkhill, J., Keane, J. A., & Harris, S. R. (2017). ARIBA: Rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial Genomics, 3*(10), e000131. doi:10.1099/mgen.0.000131
21. Hunt, M., Bradley, P., Lapierre, S. G., Heys, S., Thomsit, M., Hall, M. B., et al. (2019). Antibiotic resistance prediction for *Mycobacterium tuberculosis* from sequencing data. *Genome Medicine*.
22. Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: Pre trained bidirectional encoder representations from transformers model for DNA language in genome. *Bioinformatics, 37*(15), 2112–2120. doi:10.1093/bioinformatics/btab083
23. Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V., & Jacob, L. (2020). A machine learning approach for bacterial resistance prediction with interpretability. *PLoS Computational Biology*.
24. Khan, S. A., Shah, A., & Arman, M. (2024). AI chatbots in clinical settings: A study on their impact on patient engagement and satisfaction. *Journal of Management World, 2024*(3), 207–213. doi:10.53935/jomw.v2024i4.1201
25. Kim, J. (2022). Machine learning for predicting antimicrobial resistance from whole genome sequencing: Evaluation pitfalls and best practices. *Clinical Microbiology Reviews*.

26. Kodama, Y., Shumway, M., & Leinonen, R. (2012). The Sequence Read Archive: Explosive growth of sequencing data. *Nucleic Acids Research, 40*(D1), D54–D56. doi:10.1093/nar/gkr854
27. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*.
28. Mahé, P., Tournoud, M., & colleagues. (2018). Genome wide k mer based prediction of antimicrobial resistance with stability selection. *Bioinformatics*.
29. Papp, M., et al. (2022). Comparative assessment of antimicrobial resistance gene databases and their impact on prediction pipelines. *Briefings in Bioinformatics*.
30. Ren, Y., et al. (2022). Predicting antimicrobial resistance from bacterial genomes using machine learning: A review of methods, data, and evaluation. *Frontiers in Microbiology*.
31. Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2024). GenBank. *Nucleic Acids Research*.
32. Shukla, M., Wattam, A. R., Brettin, T., Davis, J. J., & colleagues. (2026). BV BRC: A unified bacterial and viral bioinformatics resource with expanded functionality and AI integration. *Nucleic Acids Research, 54*(D1), D715–D723. doi:10.1093/nar/gkaf1254
33. Walker, T. M., et al. (2022). The 2021 WHO catalogue of *Mycobacterium tuberculosis* complex mutations associated with drug resistance: A systematic review and dataset. *The Lancet Microbe*.
34. World Health Organization. (2021). *Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance*. World Health Organization.
35. World Health Organization. (2023). *Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance* (2nd ed.). World Health Organization. ISBN 9789240082410
36. Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., et al. (2017). Improvements to PATRIC, the all bacterial bioinformatics database and analysis resource center. *Nucleic Acids Research, 45*(D1), D535–D542. doi:10.1093/nar/gkw1017
37. Zhou, Z., Ji, Y., Li, X., & Davuluri, R. V. (2023). DNABERT 2: Efficient foundation model and tokenizer for DNA sequences. *arXiv*. arXiv:2306.15006