BIG-TB: A benchmark for prediction and interpretability of sequence-based machine learning using *Mycobacterium tuberculosis* genomes

Mahbuba Tasmin[1,*] **Saishradha Mohanty**[1,*] **Sanjana Kulkarni**[2]
Maha R. Farhat[2,3] **Anna G. Green**[1]

[1]Manning College of Information and Computer Sciences, University of Massachusetts, Amherst MA 01002
[2]Department of Biomedical Informatics, Harvard Medical School, 25 Shattuck St, Boston MA, 02115
[3]Division of Pulmonary & Critical Care, Massachusetts General Hospital, 55 Fruit St, Boston, MA, 02114
[*]Equal contribution

mtasmin@umass.edu saishradhamo@umass.edu skulkarni@g.harvard.edu
maha_farhat@hms.harvard.edu annagreen@umass.edu

Abstract

Foundation models aim to learn useful representations of biological sequences. However, the applicability of these representations for a wide range of tasks, including phenotype prediction and variant discovery, is still in question, in large part due to the relatively small set of benchmark tasks. To this end, we present the Benchmarks for Interpretable prediction from Genomes of Tuberculosis, **BIG-TB**. We curate over 17,000 genomes with high-quality short read sequencing data and experimentally measured antibiotic resistance phenotypes, combined with a curated list of canonical resistance-conferring variants, and provide these data in an ML-ready format. BIG-TB defines two tasks for interrogating the utility of foundation models: (1) predictive performance of antibiotic resistance phenotypes, and (2) attribution of predictions to known resistance-conferring variants from an expert-curated dataset.

Using our benchmark, we show that DNA-based foundation models do not yet outperform simple machine learning baselines (mean test $AUC = 0.888$ vs. 0.846 for best CNN variant vs. best DNABERT variant across drugs). Our benchmark also supports protein-based models, where performance is worse than DNA-based models due to the loss of representation of non-coding variants, but where foundation models are more competitive with simple ML. We show that the models with highest predictive performance do not necessarily perform the best at canonical resistance variant discovery - indicating that in many cases the improved performance may be due to non-causal associations between variants and phenotype. Finally, we show that the choice of embedding representation has a major impact on foundation model performance, and that representations that average over sequence position perform poorly at both prediction and canonical resistance variant discovery. Overall, BIG-TB provides a new type of benchmark for foundation models of biological sequences, facilitating comparison of representations across multiple tasks.

Code available at
`https://github.com/SAGE-Lab-UMass/Big-TB-benchmark`.

1 Introduction

Foundation models for biological sequences promise broadly reusable representations of protein [1–3] and DNA [4–8] sequences. Increasing numbers of benchmarks have been introduced to evaluate the accuracy of these models across ranges of tasks and organisms [4, 6–10]. Recent evaluations demonstrate that pre-trained sequence encoders can improve predictive accuracy across diverse biological applications, including regulatory genomics, variant-effect prediction, and protein function modeling [5, 11, 12] though some studies find no or marginal utility of foundation models compared to traditional models [13]. But, predictive accuracy alone is not enough; interpretability, which aims to understand which signals the model uses to make predictions [14–16], is a central goal in biological sequence models [17]. Interpretability methods can be used to discover and prioritize sequence features underlying biological processes of interest, such as gene regulation [18–20], gene–gene and gene–environment interactions [21], and antibiotic resistance [22, 23]. However, it is difficult to evaluate interpretability methods because the ground-truth causal sequence features are often not known in advance [17].

We provide a benchmark dataset to simultaneously evaluate sequence models for their predictive capabilities and attribution to ground-truth causal features. Our phenotype of interest is antibiotic resistance, because resistance is an organism-level phenotype that is causally attributed to specific genetic variation. In the case of *Mycobacterium tuberculosis*, both genome-level phenotypes and curated variant-level labels are available due to extensive studies on the genetic bases of drug resistance [24–27], enabling joint assessment of predictive performance and interpretability on the same isolates. Tuberculosis, the disease caused by *M. tuberculosis*, remains a highly burdensome infectious disease [28]; rapid, interpretable predictions from sequence can inform treatment and surveillance [29, 30].

Here, we introduce the Benchmark for Interpretable prediction from Genomes of Tuberculosis, BIG-TB, a dataset and benchmark that unifies prediction and interpretability evaluation for antibiotic resistance in *M. tuberculosis*. Specifically, BIG-TB (1) provides curated genotype and phenotype labels for over 17,000 clinical isolates, (2) maps genotype data to a catalog of ground-truth resistance conferring variants, (3) defines a canonical variant discovery task for quantitative interpretability (first introduced in 31) and (4) standardizes both phenotype prediction and canonical variant discovery in this dataset. Using our benchmark, we evaluate three model families for both protein and DNA-level predictions: simple machine learning on reference-alternate encodings, CNN and transformer architectures on one-hot encoded sequences, and CNNs on learned embeddings from foundation models of sequences (ESM and DNA-BERT, respectively). We show that choice of embedding representation has substantial effect on predictive performance, as standard mean-pooling obscures relevant variation. However, we show that foundation models do not yet consistently outperform simpler ML approaches on either phenotype prediction or causal variant discovery. This benchmark provides a unified, publicly reusable resource to guide the development of accurate and interpretable sequence-based models for biological prediction.

2 Methods

2.1 Datasets

Phenotype data We compile laboratory-determined phenotype data for *M. tuberculosis* clinical isolates from previously published studies [32–44, 27, 45–48]. These isolates were tested for resistance or susceptibility to eleven antibiotics, resulting in binary phenotype labels for the following drugs: rifampicin, isoniazid, ethambutol, pyrazinamide, streptomycin, capreomycin, amikacin, ethionamide, moxifloxacin, levofloxacin, and kanamycin.

High Quality Annotated VCFs For isolates with published phenotype data, we compile publicly available whole-genome sequences for *M. tuberculosis*. We downloaded the raw read data from the NCBI SRA and performed our own variant calling to ensure consistency across datasets generated by multiple groups. We perform variant calling against the H37Rv reference genome [49] and variants were called according to a pipeline introduced in Ezewudo et al. [50] and updated in Freschi et al. [51]. We trim and filter reads with FASTP [52], remove contaminated isolates using Kraken2, and align to the reference using BWA-MEM2 [53]. We remove duplicate reads using Picard

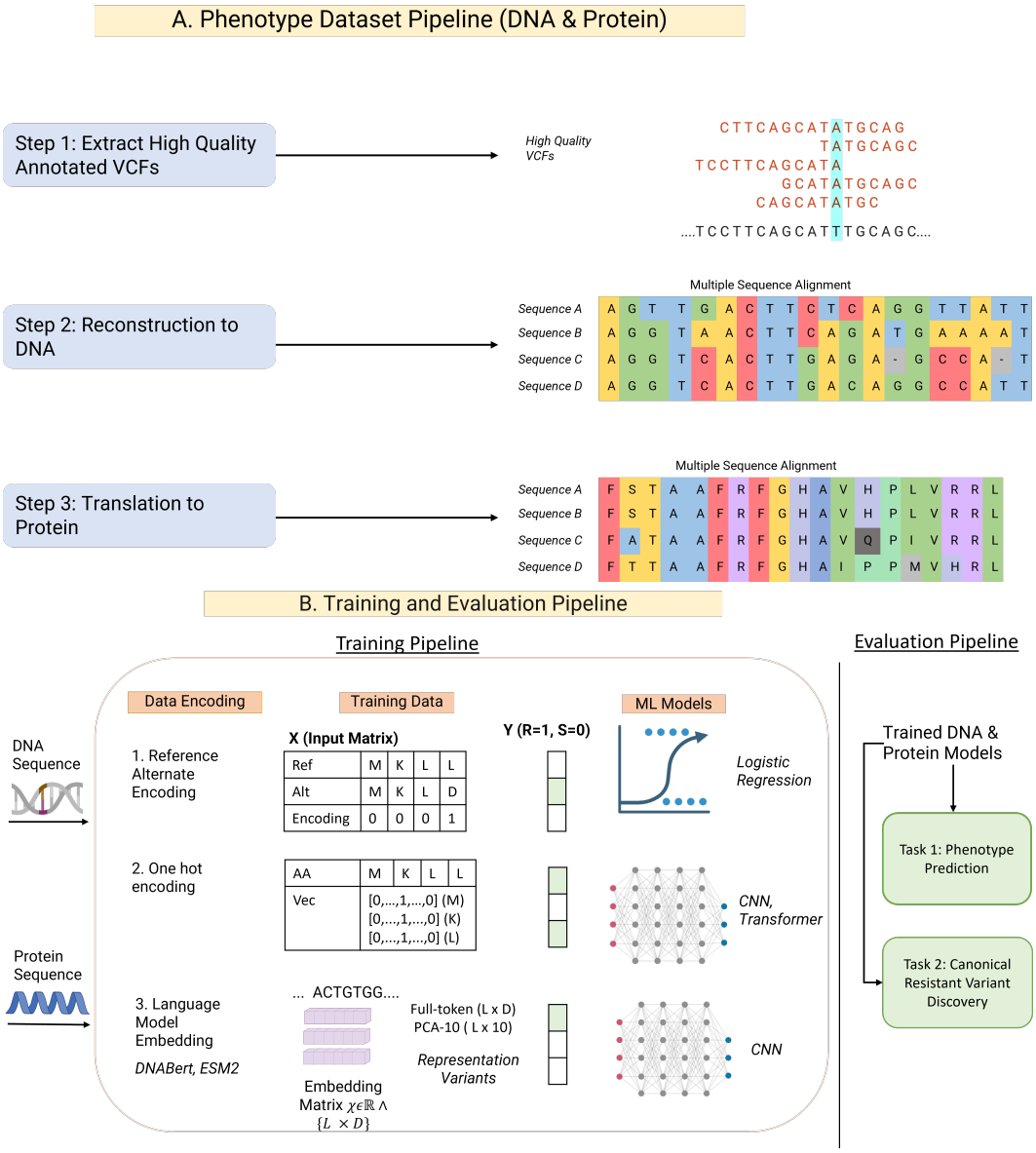


Figure 1: BIG-TB genotype and evaluation pipeline. (A) Genotype dataset pipeline: High-quality VCFs are reconstructed into gene-centric DNA sequences, aligned into multiple sequence alignments (MSAs), and translated into protein MSAs. (B) Training and evaluation pipeline: DNA and protein sequences are encoded using three representation strategies—(1) reference–alternate encoding for mutation indicators used in logistic regression, (2) one-hot encoding over 20 amino acids for CNN and Transformer models, and (3) contextual embeddings from pretrained language models (DNABERT for DNA, ESM2 for protein) processed via CNNs. All model families are trained to predict resistance phenotypes and discover canonical resistance variants

(http://broadinstitute.github.io/picard/), and we drop isolates with <95% coverage of the reference genome at 20× coverage. This results in one Variant Call Format (VCF) file per isolate that contains one row per mutation observed relative to the reference sequence.

Multiple Sequence Alignments Both foundation models and standard machine learning models for antibiotic resistance prediction use DNA or protein sequences as input. To support this, we provide a custom pipeline that outputs FASTA-formatted sequences from VCF data for each isolate as shown in Figure 1A (Appendix E). For models that take unaligned sequences as input, the sequences are generated without gaps. For downstream modeling in this study, we follow standard practice [54, 55] and select only genomic regions suspected to be involved in antibiotic resistance (Appendices D and E). While we have hand-selected regions for downstream modeling, we note that our code supports extraction of arbitrary genomic regions.

For models trained on protein data, DNA sequences for each gene region are translated into protein sequences using BioPython [56], as shown in Figure 1A (Appendix F). We generate a final CSV as shown in Table 1 containing per-isolate DNA sequences, protein sequences, phenotype labels, and a flag used to indicate sequences with a frameshift.

Table 1: Representative examples of *rpoB* mutations from isolates with distinct primary mutation sites. Each row lists the isolate ID, phenotype for a representative drug (RIF), the start of the full gene sequence, and the corresponding DNA and protein windows centered on the mutation site. Only the first detected mutation per isolate is reported for display.

Filename	Phenotype	DNA Sequence	Protein Sequence	Mutation Pos.
SAMN02585979	R	ACAAGC…	HKRRLSALGP…	Q429H
SAMEA2533629	S	ACAAGC…	HKRRLSALGL…	P454L
SAMEA3558815	R	CCCAGC…	PQRRLSALGP…	L443-
SAMN08436285	S	ACAAGC…	HKRRLSALGP…	E344Q
SAMN03648257	S	ACAAGC…	HKRRLSALGP…	D270E

Canonical Resistance Variant Discovery Data The World Health Organization (WHO) curates a comprehensive catalogue of observed mutations and their effect on antibiotic resistance in *M. tuberculosis* [26]. Mutations are graded into five evidence groups: (1) associated with resistance, (2) associated with resistance—interim, (3) uncertain significance, (4) not associated with resistance—interim, and (5) not associated with resistance (consistent with susceptibility). We implement a custom pipeline to merge this catalogue with our VCF files containing individual isolate genotypes, and shown in Table 3 (Appendix C).

2.2 Experimental Design & Evaluation

Following prior work on genotype-to-phenotype prediction in *M. tuberculosis* [57], we trained and evaluated all models to predict antibiotic resistance phenotypes from sequence data. We evaluated performance along two complementary objectives: (i) phenotype prediction—assessing predictive accuracy for resistance classification—and (ii) canonical resistance variant recovery—measuring whether learned representations recover known resistance-conferring mutations.

2.2.1 Task 1: Phenotype Prediction

Phenotype prediction was formulated either as a single-task binary classification problem, predicting resistance (`R`) or susceptibility (`S`) to a specific drug, or as a multi-task binary classification problem, predicting an eleven-drug resistance profile across all antibiotics. We refer to these as single-drug and multi-drug models, respectively.

Each isolate was represented by DNA- or protein-level features derived from one or more genes associated with resistance. For single-drug models, representations were drawn from genes relevant to that antibiotic; for multi-drug models, features from all resistance-associated genes were concatenated (Table 8). This framework was applied consistently across both input modalities (DNA and protein) and all architectures (logistic regression, CNN, Transformer, and ESM variants).

2.2.2 Task 2: Canonical Resistance Variant Discovery

Task 2 quantitatively evaluates each model's ability to recover known resistance-associated mutations curated in the WHO 2023 catalogue [26]. While past work had use the WHO catalogue to validate model predictions post-hoc[55, 58], we follow the framework introduced by Tasmin and Green [31], in which model explanations are benchmarked against human-annotated high confidence resistance variants.

We generate position-level importance scores from model attributions and rank these scores to identify the top-k most influential sites. We then measure how well these top-ranked positions overlap with WHO-listed resistance mutations using Precision@k, Recall@k, and F1@k, computed per drug across multiple k values (primary endpoint k=10).

Attribution methods depend on model class. Nonlinear sequence models (CNN, Transformer, ESM) employ DeepSHAP [15, 16] to compute token-level contributions on predicted probabilities, while linear reference–alternate baselines use the LinearExplainer variant of SHAP. Implementation details at H.4.

2.3 Computational Pipeline

Inputs and Encodings. For each isolate, we extract the candidate gene(s) associated with the drug and construct features using three encoding families: (i) reference–alternate (ref–alt) encodings, indicating nucleotide or amino-acid substitutions relative to the reference genome; (ii) one-hot encodings, where each base (DNA, 5 symbols including gaps) or residue (protein, 20 symbols) is represented by a binary vector; and (iii) foundation model embeddings, derived from DNABERT-2 (d=768 per token) for DNA and ESM-2 (esm2_t6_8M_UR50D) (d=320 per token) for protein [4, 59].

For both DNA (DNABERT-2) and protein (ESM) models, the starting point is a per-token embedding matrix $\mathbf{H} \in \mathbb{R}^{L\times d}$, where L denotes sequence length and d the embedding dimension ($d = 768$ for DNABERT-2, $d = 320$ for ESM-2 8M). While these full embeddings retain maximal information, they are memory-intensive to store for tens of thousands of isolates. Therefore, we evaluate three representation strategies with embeddings (Table): (i) PCA-k: project per-token embeddings onto the top principal components ($k = 10$ for DNA, $k \in \{10\}$ for protein); (ii) channel-mean: average across embedding channels per token, yielding a 1D descriptor per residue; and (iii) sequence-mean: average across the length dimension to obtain a single d-vector per sequence. The latter is widely used in NLP-based baselines and past ESM-1v benchmarks [12] but discards positional structure entirely [60, 61]. For DNA models, we tested all four variants in single-drug CNNs, but only sequence-mean in multi-drug models due to GPU memory limits. For protein models, we tested all four variants across drugs. We note that sequence-mean cannot be used for canonical variant discovery since it discards residue-level structure.

Table 2: Embedding Representation Strategies grouped by granularity. Per-site methods yield token/residue detail; per-sequence methods yield global summaries.

Per-site embeddings (retain token-level detail)			
Variant	utput Shape	Compression	Description
Full-d	$L \times d$	None	Raw per-token embeddings. ($d = 768$ for DNABERT, $d = 320$ for ESM).
PCA-k	$L \times k$	PCA	Incremental PCA on all tokens, project to top k components.
Channel-mean	$L \times 1$	Mean	Average across d channels per token.
Per-sequence embedding (global descriptor)			
Sequence-mean	$1 \times d$	Mean	Average across length dimension; discards positional detail.

Architectures. Logistic regression on reference–alternate encoded sequences to provide a simple and interpretable linear baseline. We train convolutional models on one-hot encoded sequences following Green et al. [55] for DNA-based models and implementing a similar architecture for protein-based models. For foundation models, we use shallow 1D CNNs with the same architecture,

but with input channels corresponding to the embedding dimension (d=320 for ESM-2 or d=768 for DNABERT-2). For complete layer dimensions and parameter counts, see Table 11.

Input loci. For single-drug models, input loci are fused by concatenation along the sequence axis. For the multi-drug DNA one-hot CNN model, input loci are stacked along the channel dimension. For multi-drug DNABERT-2 models, seq-mean embeddings from each gene are concatenated into a $(G \times 768)$ feature matrix and channel-mean embeddings from each gene are concatenated into a $(G \times L)$ feature matrix where L denotes sequence length.

Cross-validation experimental design We employed a 5-fold stratified cross-validation (CV) protocol to estimate generalization performance. For each fold, models were trained on the training split and evaluated on the held-out validation split, with a separate test set reserved for final assessment. The number of folds (5) and random seed (42) were fixed across all experiments to ensure comparability. We compute the ROC-AUC for each fold. We summarize the ROC-AUC as mean $\pm$ standard deviation across folds to quantify variability in performance.

Statistical testing across models To assess whether differences between models were statistically significant, we applied the Wilcoxon signed-rank test (two-sided) in a paired fashion across folds.

For each drug and model, we report the mean and standard deviation of AUCs, the 95% CI ($\mu \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$), and the p-value from the Wilcoxon test against the baseline. Significance was set at $p < 0.05$. Specificity and sensitivity were also computed for selected baselines.

This cross-validation and testing framework was applied consistently across all model types (LogReg, CNN, Transformer, DNA-BERT, ESM-PCA10, and ESM-Full320) in both DNA and protein settings.

3 Results

3.1 Dataset Overview

Table 3: Summary of phenotype and mutation variant statistics per drug. The *Phenotyped Isolates* columns report isolate-level counts for resistant (`R`) and susceptible (`S`) isolates based on laboratory drug susceptibility testing; isolates without phenotype labels are excluded. The `Resistance Rate` denotes the proportion of resistant isolates among all phenotyped isolates for each drug. The *WHO Catalogue Variants* columns summarize mutation annotations from the WHO resistance catalogue, reporting the total number of catalogued variant positions per drug, the subset classified as resistance-associated (confidence levels 1–2), and variants annotated as uncertain or not associated with resistance (confidence levels 3–5). R Isolates w/ WHO Variant denotes the fraction of phenotypically resistant isolates that harbor at least one WHO resistance-associated variant. WHO R Variants Observed (Ratio) denotes the fraction of WHO-listed resistance-associated variant positions for each drug that are observed at least once in the isolate-derived FASTA sequences.

Drug	Phenotyped Isolates (#)			WHO Catalogue Variants			R Isolates w. WHO Variant (Ratio)	WHO R Variants Observed (Ratio)
	Resistant	Susceptible	Resistance Rate	Total	Assoc. w. Resistance	Not/Uncertain		
Rifampicin	4869	12712	0.277	7266	136	7130	0.947	0.58
Isoniazid	5825	11610	0.334	7269	142	7127	0.906	0.316
Ethambutol	2945	12165	0.195	7109	13	7096	0.792	1.00
Pyrazinamide	2131	10772	0.165	2562	341	2221	0.803	0.75
Streptomycin	2416	5266	0.315	3052	159	2893	0.732	0.635
Capreomycin	675	3179	0.175	2648	70	2578	0.692	0.214
Amikacin	707	2998	0.191	2183	4	2179	0.696	1.00
Ethionamide	690	1463	0.320	2739	286	2453	0.787	0.49
Moxifloxacin	388	2480	0.135	2717	18	2699	0.675	1.00
Levofloxacin	76	193	0.282	3040	18	3022	0.908	1.00
Kanamycin	703	2873	0.197	2233	8	2225	0.821	1.00

We compiled resistance phenotype labels and corresponding genomic variant data for 17,942 clinical isolates of *M. tuberculosis* (Table 3). Each isolate is classified as resistant (`R`) or susceptible (`S`) to one or more antibiotics based on laboratory drug susceptibility testing. At the isolate level, we report the number of resistant and susceptible isolates for each drug and the corresponding resistance rate. For example, 27.7% of isolates are phenotypically resistant to rifampicin. At the variant level, we quantify the prevalence of canonical resistance variants across isolates. Because the presence of canonical resistance variants is evaluated independently of phenotypic resistance labels, the fraction of isolates harboring such variants may exceed the observed phenotypic resistance rate. For rifampicin, 28.51%

of isolates carry at least one resistance variant, including 94.7% of phenotypically resistant isolates . This confirms that presence of genotypic R variant agrees with the phenotypic resistant labels. For other drugs, a lower fraction of resistant isolates harbor canonical resistance variants, ranging from 67.5% for moxifloxacin to to 90.8% for levofloxacin. Finally, we report the fraction of WHO-listed resistance-associated variants that are observed in our isolate-derived FASTA sequences.

3.2 Benchmarking foundation models

We seek to ask whether foundation models can perform accurate antibiotic resistance phenotype prediction and can recover known canonical resistance variants. To this end, we evaluate single–drug and multi-drug models on two standardized task: Task 1 (Phenotype Prediction), and Task 2 (canonical resistance variant discovery).

3.2.1 Task 1: Phenotype Prediction

We compare all single-drug DNA-based models against the Logistic Regression baseline [55] using paired Wilcoxon signed-rank tests across identical cross-validation folds (Figure 2a). The single-drug SD-CNN on one-hot encoded sequences performed comparably to the baseline SD-LogReg (Mean AUC $= 0.886$ vs 0.874, $p \geq 0.05$), consistent with prior findings. We then benchmark SD-DNABERT-CNN models trained on token level embeddings. All corresponding DNABERT-CNN variants (CNN-768 and PCA10) perform comparably or worse than SD-LogReg overall across all drugs with a mean AUC of 0.852 ($p \geq 0.05$) for CNN-768 and 0.850 for PCA10 ($p = 0.04$). Both variants also perform significantly worse than the baseline ML model, SD-CNN, across all drugs (Table 14). Overall, among the single drug models, SD-CNN and SD-LogReg are the most performant, with foundation models significantly underperforming simple baselines.

For the multi-drug case, there is no Logistic Regression Baseline, hence, we compare all the multi-drug models against the existing, state-of-the-art, Multi Drug CNN (MD-CNN) model [55] trained on one-hot encoded sequences. Across most drugs, MD-DNABERT-CNN ($AUC = 0.824$) yields significantly lower predictive performance compared to MD-CNN ($AUC = 0.924$, $p = 0.0009$). We confirm that ML is a necessary approach to resistance prediction from genomes, by evaluating the WHO rules-based heuristic for sensitivity and specificity at the DNA level. Results show lower specificity (0.93) and sensitivity (0.85), leaving many resistant isolates incorrectly called as susceptible (Table 10). The MD-CNN model has the best sensitivity and specificity as 0.94 and 0.89, respectively.

For protein-based models, we compared all architectures against the Logistic Regression baseline using paired Wilcoxon signed-rank tests across identical cross-validation folds (Figure 2(b)). Full per-drug AUCs, standard deviations, confidence intervals, and paired Wilcoxon p-values are reported in Supplementary Table 15. Across drugs, no architecture achieved statistically significant gains over the baseline ($p > 0.05$). For drugs with a large number of labeled isolates (isoniazid, rifampicin, ethambutol; Table 3), CNN and embedding-based models reached AUCs ≥ 0.9, but did not show significant gains over Logistic Regression. One-hot CNNs generally matched baseline performance within statistical variation, while one-hot Transformers underperformed across all drugs—for example, pyrazinamide (0.66 vs. 0.84) and streptomycin (0.76 vs. 0.86)—although these differences did not reach statistical significance (p = 0.06 in both cases). Among the ten antibiotic resistance phenotypes predicted with protein data, amikacin and capreomycin yielded near-random AUCs (≈ 0.5) because their main resistance determinants reside in ribosomal RNA, hence the input proteins contain insufficient signal.

Test set accuracy We then measure the predictive accuracy of our DNA and protein models on the held-out test set (Table 4). Consistently, MD-CNN achieves the highest AUC values for the majority of drugs, with $\mathrm{AUC} > 0.9$ for isoniazid, ethambutol, and streptomycin. This indicates that concatenated, multi-gene representations are particularly effective when modeling resistance mechanisms jointly across drugs. In contrast, single-drug DNABERT embeddings both the full 768-dimensional variant and its PCA-compressed form tend to underperform classical one-hot CNNs.

The protein sequence–based results show a more heterogeneous performance landscape across antibiotics. CNNs trained directly on one-hot encoded sequences (CNN-OHE) and models based on ESM embeddings (ESM-320 and ESM-PCA10) achieved the highest AUCs for most drugs, with CNN-OHE reaching near-perfect separation for levofloxacin (AUC = 0.997) and maintaining strong

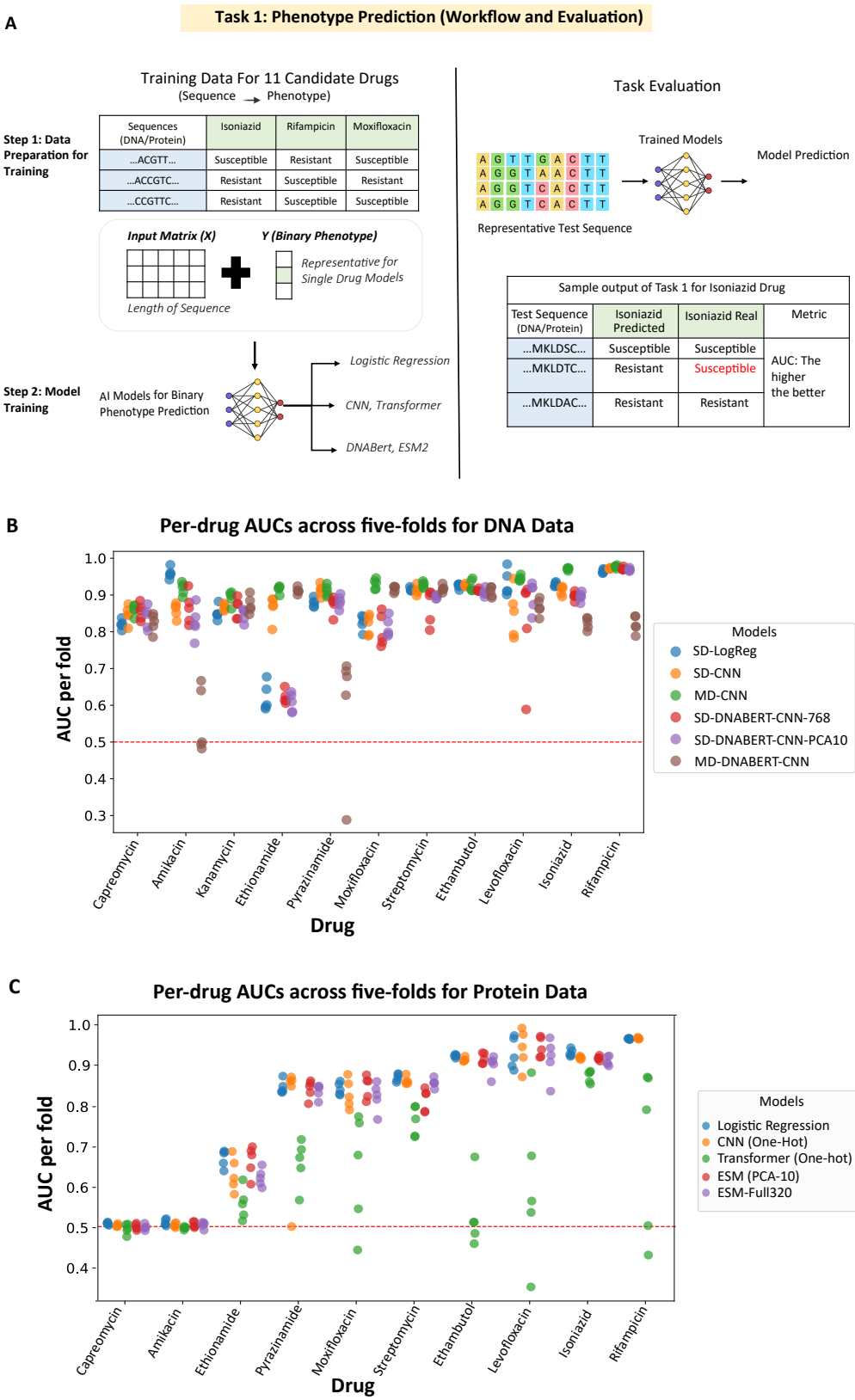


Figure 2: Overview of Task 1: phenotype prediction from DNA and protein sequence features. (A) Workflow showing model training and evaluation setup. (B) DNA-based model performance across 11 drugs; each point represents ROC–AUC on one held-out fold per model. (C) Protein-based model performance across 10 drugs. Red dashed line indicates random-chance performance (AUC = 0.5).

performance for ethambutol and rifampicin. Logistic Regression using Ref–Alt encodings remained competitive, particularly for isoniazid (0.932) and rifampicin (0.961). Transformers trained on one-hot inputs consistently underperformed CNNs and ESM-based models. ESM-based embeddings offered modest gains for specific drugs such as pyrazinamide and ethionamide.

Table 4: Test AUC across first- and second-line antibiotics using nucleotide (DNA) and protein sequence inputs. DNA results (top) are based on single- or multi-drug modeling; protein results (bottom) use per-drug inputs of single genes or concatenations per known mechanisms. Protein models which excludde the major resistance-determining regions of the small and large ribosomal RNA are marked with *. Protein results for Kanamycin were not available since its locus is not protein-coding (marked with "—"). Columns are aligned such that analogous model families across data modalities (e.g., CNN(OHE), DNABERT vs. ESM for DNA and protein models respectively) occupy the same positions.

DNA					
Drug	SD-LogReg (Ref–Alt)	SD-CNN (OHE)	MD-CNN (OHE)	SD-DNABERT-CNN-768	SD-DNABERT-CNN-PCA10
Amikacin	0.807	0.845	**0.899**	0.820	0.812
Capreomycin	0.846	**0.873**	0.829	0.819	0.812
Ethambutol	0.915	0.931	**0.936**	0.916	0.908
Ethionamide	0.708	0.670	**0.709**	0.581	0.585
Isoniazid	0.920	0.917	**0.951**	0.901	0.899
Levofloxacin	0.835	0.839	0.850	**0.853**	0.830
Moxifloxacin	0.864	0.886	**0.887**	0.800	0.792
Pyrazinamide	0.896	**0.930**	0.908	0.865	0.875
Rifampicin	0.975	0.977	0.981	**0.985**	0.969
Streptomycin	0.924	0.911	**0.938**	0.899	0.892
Kanamycin	0.834	0.849	**0.875**	0.867	0.857

Protein					
Drug	LogReg (Ref–Alt)	CNN (OHE)	Transformer (OHE)	ESM-320	ESM-PCA10
Amikacin*	**0.510**	0.501	0.500	0.503	0.505
Capreomycin*	**0.503**	0.501	0.489	0.502	0.499
Ethambutol	0.885	**0.923**	0.640	0.750	0.911
Ethionamide	0.534	0.620	0.405	0.638	**0.663**
Isoniazid	**0.920**	0.919	0.887	0.912	0.907
Levofloxacin	0.854	**0.997**	0.856	0.864	0.921
Moxifloxacin	0.796	0.807	0.613	0.804	**0.820**
Pyrazinamide	0.626	0.820	0.728	**0.851**	0.844
Rifampicin	0.966	0.961	0.533	**0.969**	0.969
Streptomycin*	0.784	0.840	0.728	**0.858**	0.832
Kanamycin*			—		

Comparison between DNA- and protein-based models. DNA-based models consistently achieved higher overall AUCs than protein-based models for the same drugs, likely because they incorporate both coding and non-coding regions, including regulatory and rRNA sequences that are excluded from protein-based analyses. Protein models performed well only for drugs whose known resistance-conferring mutations occur within protein-coding genes (*rpoB*, *katG*, *embB*), achieving AUCs above 0.9, but degraded to near-random performance for rRNA-mediated drugs such as amikacin and capreomycin.

3.2.2 Task 2: Canonical Resistance Variant Discovery

We test whether model-derived residue importance recovers WHO-annotated resistance loci (2023 catalogue). Table 5 summarizes the fraction of WHO resistance-conferring residues recovered at $k = 10$ across DNA- and protein-based models.

Among DNA models, SD-CNN perform best overall, with the exception of kanamycin where all the models perform equivalently. Across most first line drugs, the OHE based SD-CNN consistently identifies more true positive variants and achieves higher precision, particularly for ethambutol (0.70) and rifampicin (0.50) where several well-known canonical resistance variants (e.g., *embB* p.Met306Val), *rpoB* p.Ser450Leu, are recovered. However, recall is low because the DNA dataset contains many unique canonical resistance variants relative those recovered at k=10, as shown in table 16. Amikacin achieves the highest recall (0.75) after recovering 3/4 canonical resistance variants with one of them being a non-coding rrs-associated resistance site (*rrs* n.1401A>G). Interestingly, all the single-drug models recover significant number of canonical resistance variants for pyrazinamide, with SD-LogReg achieving the highest precision (0.80) by correctly retrieving five variants. Recovery rates of DNABERT variants are poor compared to regression and CNN based DNA models with low recovery for most second line drugs.

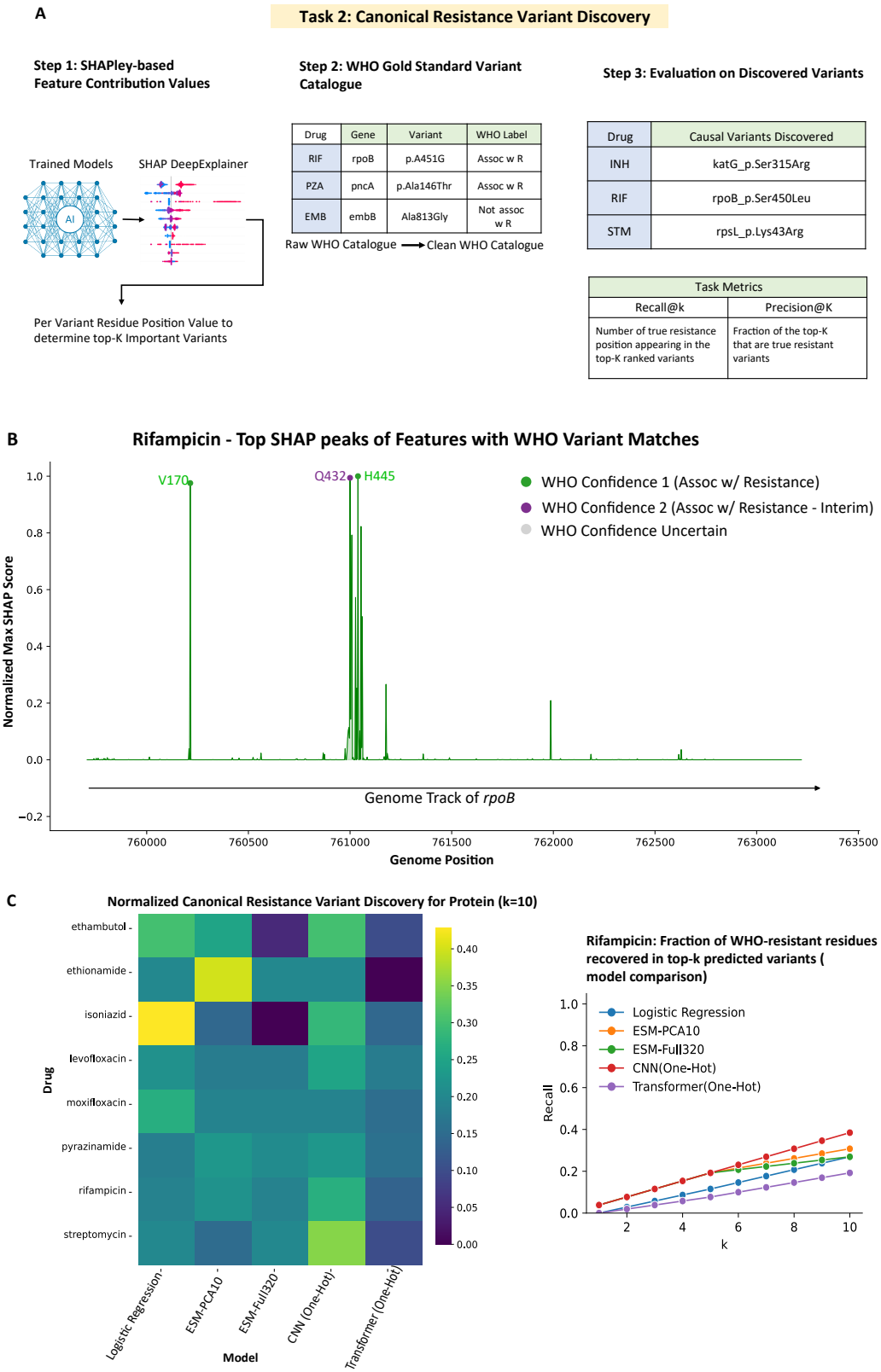


Figure 3: Canonical Resistance Variant Discovery. (a) Workflow for identifying causal variants using SHAP-based model interpretation, the WHO 2023 resistance catalogue, and top-k evaluation. (b) Genome tracks for rifampicin showing SHAP peaks at the *rpoB* locus annotated with WHO resistance variants. (c) Model comparison using recall@k curves for rifampicin and row-normalized heatmap summarizing recall@k=10 across all drugs and models.

Table 5: Fraction of WHO resistance-conferring residues discovered at $k = 10$ (Recall@k) and corresponding precision values (Prec@10) for each drug–model pair, comparing DNA- and protein-based approaches. Bold values indicate the best performance per drug within each block.

DNA											
Drug	Res. Pos.	SD-LogReg		SD-CNN (OHE)		MD-CNN (OHE)		SD-DNABERT-CNN-768		SD-DNABERT-CNN-PCA10	
		Prec@10	**Recall@10**	**Prec@10**	**Recall@10**	**Prec@10**	**Recall@10**	**Prec@10**	**Recall@10**	**Prec@10**	**Recall@10**
Amikacin	4	0.20	0.50	**0.30**	**0.75**	0.10	0.25	0.20	0.25	0.10	0.25
Kanamycin	8	0.20	0.125	0.10	0.125	0.10	0.125	0.20	0.125	0.10	0.125
Capreomycin	16	**0.20**	**0.063**	**0.20**	**0.063**	0.10	0.063	0.00	0.00	0.00	0.00
Ethambutol	9	0.50	0.076	**0.70**	**0.444**	0.10	0.153	0.40	0.076	0.20	0.076
Ethionamide	150	**0.50**	**0.007**	0.30	0.013	0.10	0.007	0.00	0.00	0.00	0.00
Isoniazid	42	**0.30**	**0.119**	**0.30**	**0.119**	0.20	0.044	0.10	0.022	0.00	0.00
Levofloxacin	14	0.40	0.071	**0.50**	**0.214**	0.10	0.071	0.00	0.00	0.00	0.00
Moxifloxacin	14	0.40	0.071	**0.50**	**0.214**	0.10	0.071	0.20	0.00	0.20	0.00
Pyrazinamide	194	**0.80**	**0.00**	0.50	0.018	0.00	0.00	0.50	0.007	0.50	0.004
Rifampicin	47	0.30	0.066	0.50	0.149	0.30	0.066	0.60	0.044	**0.70**	**0.044**
Streptomycin	101	**0.50**	**0.019**	0.40	0.039	0.30	0.029	0.30	0.025	0.20	0.025

Protein											
Drug	Res. Pos.	CNN (OHE)		ESM-320 (Full)		LogReg (Ref–Alt)		ESM-PCA10		Transformer (OHE)	
		Prec@10	**Recall@10**	**Prec@10**	**Recall@10**	**Prec@10**	**Recall@10**	**Prec@10**	**Recall@10**	**Prec@10**	**Recall@10**
Amikacin	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kanamycin	–	–	–	–	–	–	–	–	–	–	–
Capreomycin	2	**0.10**	**0.50**	0.00	0.00	0.00	0.00	**0.10**	**0.50**	0.00	0.00
Ethambutol	6	**0.60**	**1.00**	0.10	0.17	**0.60**	**1.00**	0.50	0.83	0.20	0.33
Ethionamide	10	0.10	0.10	0.10	0.10	**0.10**	**0.10**	0.20	0.20	0.00	0.00
Isoniazid	3	0.20	0.67	0.00	0.00	**0.30**	**1.00**	0.10	0.33	0.10	0.33
Levofloxacin	10	**0.70**	**0.70**	0.50	0.50	0.60	0.60	0.50	0.50	0.50	0.50
Moxifloxacin	10	0.50	0.50	0.50	0.50	**0.70**	**0.70**	0.50	0.50	0.40	0.40
Pyrazinamide	95	**1.00**	**0.11**	0.90	0.09	0.80	0.08	**1.00**	**0.11**	0.70	0.07
Rifampicin	26	**1.00**	**0.38**	0.70	0.27	0.70	0.27	0.80	0.31	0.50	0.19
Streptomycin	10	**0.70**	**0.70**	0.40	0.40	0.40	0.40	0.30	0.30	0.20	0.20

Protein models exhibit greater variability across models and drugs, with CNNs and ESM-PCA10 generally performing best. For fluoroquinolones, recovery is particularly strong: both levofloxacin and moxifloxacin achieve recall values up to 0.70–1.00, demonstrating the capacity of protein-based models to prioritize well-known *gyrA* and *gyrB* hotspots. Similarly, isoniazid achieves complete recovery under Ref–Alt logistic regression, reflecting the strong signal from canonical *katG* and *inhA* mutations found in protein-coding genes. See Supplementary Table 16 for full per-drug variant recovery details).

Linking predictive accuracy and canonical variant discovery performance. In the DNA modality, models that excel at phenotype prediction do not necessarily identify the most canonical resistance variants. The multi-drug CNN (MD–CNN) achieves the highest predictive accuracy overall, yet it recovers fewer canonical resistance mutations than the single-drug CNN (SD–CNN), which yields stronger correspondence with WHO-annotated loci. This divergence between predictive and causal performance is analyzed in detail in Section 3.2.3.

A similar pattern is observed in the protein modality: the architectures with the best predictive accuracy are not the ones that recover the most biologically validated resistance sites. In Task 1, CNN–OHE achieves the highest (AUC $\geq$ 0.9) for well-characterized targets such as *rpoB*, *katG*, and *embB*. However, in Task 2, CNN–OHE is not uniformly superior: Ref–Alt Logistic Regression achieves complete recovery for isoniazid, and ESM–PCA10 matches or exceeds CNN–OHE for fluoroquinolones.

3.2.3 Effect of increasing input loci in DNA models

We investigate whether expanding the set of resistance-associated loci incorporated into the models enhances their predictive capacity and enables them to better discover resistance variants. Extending the observations by [55], we evaluate a multi-drug DNA model (MD-CNN) and single-drug DNA models (SD-CNN), and multi-drug and single-drug foundation model variants.

In the one-hot encoded (OHE) CNN models, SD-CNN and MD-CNN architectures perform comparably across first-line (mean $\Delta\text{AUC} = +0.011$) and second line drugs (mean $\Delta\text{AUC} = +0.015$). Despite the superior predictive performance of MD-CNN, SD-CNN outperforms at the canonical resistance variant discovery task. For first-line drugs with well-characterized resistance mechanisms, such as ethambutol and rifampicin, SD-CNN recovers (7 vs 1) and (5 vs 3) canonical resistance variants, respectively. Notably, SD-CNN also recovers five canonical resistance variants for pyrazinamide, whereas MD-CNN fails to recover any. For the multi-drug models, many top-ranked features

corresponded to variants that cause resistance to other antibiotics, rather than direct mechanistic causes of the resistance in question.

We find that this trade-off between prediction and canonical resistance variant discovery also appears in the foundation model. In the DNABERT embedding based CNN models (seq-mean variant), MD-DNABERT-CNN slightly outperforms SD-DNABERT-CNN across first-line drugs (mean $\Delta\mathrm{AUC} = +0.05$). Among second-line drugs (excluding levofloxacin), the multidrug model demonstrates a clear performance advantage (mean $\Delta\mathrm{AUC} = +0.188$). The marked dip in performance for Levofloxacin in the DNA foundation model likely reflects overfitting in the multi-drug case, given that only approximately 8% of isolates carry known resistance mutations in *gyrA* or *gyrB*.

3.3 Ablation Study of Foundation Model Embedding Variants

Foundation models for biological sequences are intended to encode rich contextual information in their embeddings of biological sequences, but how to represent these embeddings for downstream tasks remains an open question. Using our newly introduced BIG–TB benchmark, we systematically evaluate alternative representation strategies for compressing and aggregating token-level embeddings from DNABERT–2 (DNA) and ESM–2 (protein). These representations differ in the degree to which they preserve positional detail versus global context (Table 2). Specifically, we compare: (*i*) full per-token embeddings that retain all token-level and contextual information; (*ii*) PCA-k projections that reduce dimensionality while maintaining token-level information; (*iii*) channel-mean embeddings that collapse across channels per position; and (*iv*) sequence-mean embeddings that pool across the entire sequence, producing a single vector descriptor. Sequence-mean embeddings are widely used in NLP for efficiency [61], but they discard positional structure and are therefore unsuited for residue-level interpretability. Here, we assess how these representation choices affect both predictive accuracy (Task 1) and canonical resistance variant discovery (Task 2) across DNA and protein modalities.

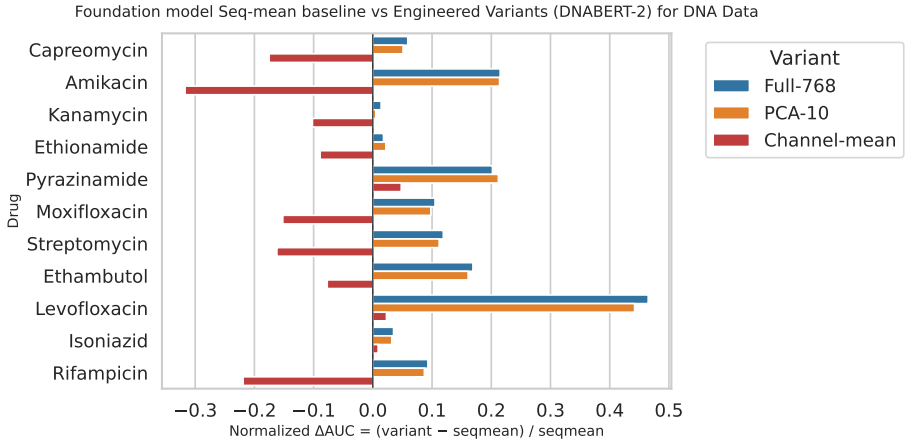
3.3.1 Representation quality across tasks (AUC and recall@k)

Predictive performance (AUC). For DNA models, preserving per-token structure through DNABERT-2 embeddings consistently improves accuracy over the seq–mean (global pooling) baseline. The SD-DNABERT CNN-768 representation improves upon the seq-mean baseline across every drug (mean $\Delta\mathrm{AUC} = +0.136$). The PCA-10 representation also yields consistent improvement across all eleven drugs (mean $\Delta\mathrm{AUC} = +0.131$), showing that projecting per-site embeddings to ten principal components preserves almost all predictive signal. Channel-mean aggregation is consistently detrimental for most drugs (-0.110) except isoniazid, pyrazinamide and levofloxacin indicating that retaining specific channel dimensions is necessary to effectively capture signals relevant to resistance prediction. Overall, within SD-DNABERT variants, full-768 achieves the highest AUC in 9/11 drugs, PCA-10 in 2/11, channel-mean and seq-mean never lead.

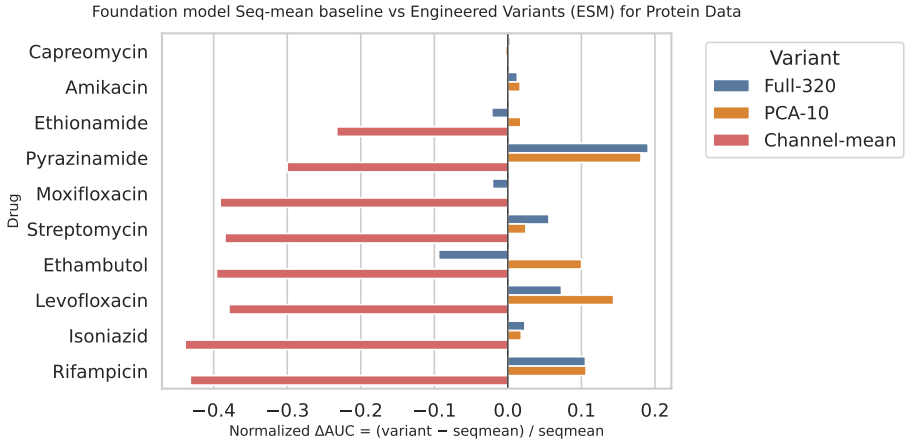
For protein models, preserving per-residue structure while compressing channels via PCA consistently improves accuracy over the seq-mean (global pooling) baseline (Figure 4b). Across the ten drugs, PCA-10 outperforms seq-mean in eight cases (mean $\Delta\mathrm{AUC} = +0.048$), while full-320 improves in seven ($+0.025$ on average). Channel-mean aggregation is consistently harmful (-0.24). Within-ESM comparisons show that PCA-10 achieves the highest AUC in 5/10 drugs, full-320 in 4/10; seq-mean and channel-mean never lead. Exact normalized ΔAUC values are provided in Supplementary Figure 6.This trend indicates that moderate compression (PCA-10) effectively preserves the discriminative content of full embeddings while improving stability across drugs.

Canonical resistance variant discovery. For DNA models, full-768 and PCA-10 representations yield comparable results at canonical variant discovery. Channel-mean fails to capture resistant variants, indicating that higher-dimensional per-site features are needed to detect resistance-associated sites. Representation strategies that pool across the sequence (e.g., sequence-mean) discard positional structure entirely and are excluded from interpretability analyses.

For protein models, PCA-10 and full-320 ESM embedding representations achieve competitive recovery of resistance-conferring residues (Table 5). Sequence-mean is again excluded from the evaluation due to lack of positional structure. Overall, these ablations reveal that retaining per-token positional structure—either through full or PCA-compressed embeddings—is critical for capturing canonical resistance variants: while global sequence pooling retains enough information for accurate classification, it fails to localize canonical resistance variants because spatial context is lost.



(a) DNABERT-2 embedding representation strategies (DNA). Full-768 embeddings outperform seq-mean in 9/11 drugs (mean $\Delta\text{AUC} = +0.136$), and PCA-10 improves in 2/11 ($+0.131$ on average), showing that low-rank projection preserves most predictive signal. Channel-mean pooling is consistently harmful (-0.110) except in isoniazid, pyrazinamide and levofloxacin.



(b) ESM-2 embedding representation strategies (Protein). PCA-10 achieves consistent gains across 8/10 drugs (mean $\Delta\text{AUC} = +0.048$), with notable improvements for pyrazinamide ($+0.181$), levofloxacin ($+0.144$), rifampicin ($+0.106$), and ethambutol ($+0.100$). Full-320 provides smaller but stable gains ($+0.025$ on average), while channel-mean aggregation is consistently detrimental (-0.24).

Figure 4: Ablation of foundation model embedding representation strategies for DNA and protein sequences. Normalized change in AUC (ΔAUC) is shown for each drug relative to the sequence–mean baseline. Positive values indicate improvement over seq-mean. Both DNABERT (a) and ESM (b) models show that preserving per-token structure (Full or PCA-10) yields clear accuracy gains, whereas channel-mean and global pooling substantially reduce predictive power. Exact normalized ΔAUC matrices are provided in Supplementary Figures 5,6

4 Discussion & Conclusion

We present BIG-TB, a foundation model benchmark for antibiotic resistance prediction and canonical resistance variant discovery in *M. tuberculosis*. Our benchmark enables DNA-level representations of any genomic region, and protein-level representations of coding regions, combined with phenotype data for 11 antibiotics. We uniformly reprocessed publicly available *M. tuberculosis* sequencing data using stringent quality thresholds, and then mapped each annotated mutation to the WHO catalogue of known resistance mutations [26]. Our dataset enables joint evaluation of predictive accuracy (Task 1) and interpretability through canonical resistance variant discovery (Task 2), enabling systematic comparison of classical machine-learning and foundation-model architectures. Our benchmark links genotype, phenotype, and curated resistance annotations for interpretable machine learning.

Using our new benchmark, we evaluated DNA- and protein-based foundation models versus simpler machine learning methods for predicting antibiotic resistance. For DNA-based models, simpler approaches outperformed DNABERT-derived embeddings for most drugs. For protein-based models, compact foundation model variants yielded the highest AUCs across drugs, though these improvements were marginal over simpler models (Figure 4a,4b). DNA-based modeling produced higher predictive accuracy than protein-based modeling. We attribute this to the fact that antibiotic resistance is caused by both coding and non-coding variation, hence the best performance can be obtained with models that consider both types of input variation.

Our benchmark includes a new task, canonical resistance variant discovery, which assesses whether models have learned the true biological causes of a phenotype. This task is performed after model training, by evaluating model attribution using SHAP-based methods, and comparing the highest-scoring features to known, biologically causal features of the phenotype. While SHAP is the current standard practice for model interpretation, our dataset could be used to evaluate future advancements in this space, including those from causal modeling and mechanistic interpretability. Ongoing work to incorporate additional biological signal, such as protein structure, sequence annotation, and known interaction partners, may improve the ability to foundation models to reason about the effects of mutations.

Expanding the DNA modeling framework to multi-drug, multi-gene prediction improved predictive accuracy at the expense of canonical variant recovery. Our analyses revealed that in multi-drug models, many top-ranked features corresponded to variants correlated with resistance to other drugs rather than direct mechanistic causes, due to correlated resistance phenotypes in MDR/XDR isolates [55, 62]. These results caution against naïve multitask integration, and instead encourage explicit modeling of shared or correlated resistance pathways [62]. They also urge caution when incorporating additional loci into predictive models – as any locus with correlated, non-causal variation may result in improved predictive accuracy at the expense of biological fidelity.

Ablation experiments across foundation model embedding variants showed that compact position-preserving representations nearly match or outperform both high-dimensional and globally pooled embeddings. For DNA, projecting per-token DNABERT embeddings to 10 principal components (PCA-10) retains almost all predictive signal while improving efficiency. For proteins, ESM-PCA10 achieves consistent $\Delta\mathrm{AUC}$ gains (+0.048 on average) relative to sequence-mean baselines and often matches per-token performance. Channel-mean aggregation-averaging over embedding dimension is consistently harmful across both modalities. These findings suggest that most biologically relevant variation is encoded in a small, low-rank subspace of foundation model embeddings, and that preserving residue positions is essential for both predictive accuracy and biological interpretability.

Our findings establish a new benchmark for sequence-based antibiotic resistance modeling. In both DNA and protein modalities, preserving positional structure proves indispensable for identifying canonical resistance variants and achieving biologically meaningful predictions. These results emphasize that foundation models, while powerful, require careful adaptation to the statistical and mechanistic constraints of genomic resistance data. By integrating predictive accuracy and biological interpretability within a unified evaluation framework, this study provides a blueprint for developing faithful, transparent models in pathogen genomics.

Acknowledgments and Disclosure of Funding

The authors thank members of the UMass SAGE lab and the Farhat lab at HMS DBI for valuable input. This work utilized resources from Unity, a collaborative, multi-institutional high-performance computing cluster managed by UMass Amherst Research Computing and Data. SK was supported by the National Science Foundation Graduate Research Fellowship DGE2140743.

References

[1] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL `https://www.science.org/doi/abs/10.1126/science.ads0018`.

[2] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

[3] Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978, November 2023. doi: https://doi.org/10.1016/j.cels.2023.10.002. URL `https://www.cell.com/cell-systems/fulltext/S2405-4712(23)00272-7`.

[4] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome, 2023. URL `@inproceedings{zhou2024dnabert,title={{DNABERT}-2: EfficientFoundationModelandBenchmarkForMulti-SpeciesGenomes},author={ZhihanZhouandYanrongJiandWeijianLiandPratikDuttaandRamanaVDavuluriandHanLiu}, booktitle={TheTwelfthInternationalConferenceonLearningRepresentations}, year={2024},url={https://openreview.net/forum?id=oMLQB4EZE1}}`.

[5] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL `https://doi.org/10.1093/bioinformatics/btab083`.

[6] Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336. URL `https://www.science.org/doi/abs/10.1126/science.ado9336`.

[7] Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, and others. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, pages 2025–02, 2025.

[8] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, and others. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

[9] Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, 2023.

[10] Jacob West-Roberts, Joshua Kravitz, Nishant Jha, Andre Cornman, and Yunha Hwang. Diverse Genomic Embedding Benchmark for functional evaluation across the tree of life. *Cold Spring Harbor Laboratory*, July 2024. doi: 10.1101/2024.07.10.602933. URL `http://dx.doi.org/10.1101/2024.07.10.602933`.

[11] Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. *bioRxiv*, 2021. doi: 10.1101/2021.02.12.430858. URL `https://www.biorxiv.org/content/10.1101/2021.02.12.430858v1`.

[12] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf`.

[13] Ziqi Tang, Nirali Somia, Yiyang Yu, and Peter K Koo. Evaluating the representational power of pre-trained DNA language models for regulatory genomics. *Genome Biology*, 26(1), July 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03674-8. URL `https://doi.org/10.1186/s13059-025-03674-8`.

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

[15] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[16] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2:56–67, 2020. doi: 10.1038/s42256-019-0138-9.

[17] Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023.

[18] Jareth C. Wolfe, Liudmila A. Mikheeva, Hani Hagras, and Nicolae Radu Zabet. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila. *Genome Biology*, 22(1):308, 2021. doi: 10.1186/s13059-021-02508-1.

[19] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

[20] Shushan Toneyan and Peter K Koo. Interpreting cis-regulatory interactions from large-scale deep neural networks. *Nature genetics*, 56(11):2517–2527, 2024.

[21] Pål V. Johnsen, Signe Riemer-Sørensen, Andrew T. DeWan, Megan E. Cahill, and Mette Langaas. A new method for exploring gene–gene and gene–environment interactions in GWAS with tree ensemble methods and SHAP values. *BMC Bioinformatics*, 22(1):230, 2021. doi: 10.1186/s12859-021-04041-7.

[22] Anna G. Green, Chang Ho Yoon, Michael L. Chen, Yasha Ektefaie, Mack Fina, Luca Freschi, Matthias I. Gröschel, Isaac Kohane, Andrew Beam, and Maha Farhat. A convolutional neural network highlights mutations relevant to antimicrobial resistance in mycobacterium tuberculosis. *Nature Communications*, 13(1):3817, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-31236-0. URL `https://doi.org/10.1038/s41467-022-31236-0`.

[23] Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Anne-Christin Hauschild, Oliver Schwengers, and Dominik Heider. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, 38(2):325–334, 2022.

[24] Maha R Farhat, Reem Sultana, Oksana Iartchouk, S Bozeman, James Galagan, Patricia Sisk, Caroline Stolte, Hila Nebenzahl-Guimaraes, Kristina R Jacobson, Anna Sloutsky, and others. Genetic determinants of drug resistance in Mycobacterium tuberculosis and their diagnostic value. *American Journal of Respiratory and Critical Care Medicine*, 194(5):621–630, 2016. doi: 10.1164/rccm.201510-2091OC.

[25] Paolo Miotto, Belay Tessema, Elisa Tagliani, Leonid Chindelevitch, Angela M. Starks, Claudia Emerson, Debra Hanna, Peter S. Kim, Richard Liwski, Matteo Zignol, Christopher Gilpin, Stefan Niemann, Claudia M. Denkinger, Joy Fleming, Robin M. Warren, Derrick Crook, James Posey, Sebastien Gagneux, Sven Hoffner, Camilla Rodrigues, Iñaki Comas, David M. Engelthaler, Megan Murray, David Alland, Leen Rigouts, Christoph Lange, Keertan Dheda, Rumina Hasan, Uma Devi K. Ranganathan, Ruth McNerney, Matthew Ezewudo, Daniela M. Cirillo, Marco Schito, Claudio U. Köser, and Timothy C. Rodwell. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. *European Respiratory Journal*, 50(6), 2017. doi: 10.1183/13993003.01354-2017. URL `http://erj.ersjournals.com/content/erj/50/6/1701354`.

[26] World Health Organization. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance, second edition. 2023. URL `https://www.who.int/publications/i/item/9789240082410`.

[27] T M Walker, T A Kohl, S V Omar, J Hedge, C Del Ojo Elias, P Bradley, Z Iqbal, S Feuerriegel, K E Niehaus, D J Wilson, D A Clifton, G Kapatai, C L C Ip, R Bowden, F A Drobniewski, C Allix-Beguec, C Gaudin, J Parkhill, R Diel, P Supply, D W Crook, E G Smith, A S Walker, N Ismail, S Niemann, and T E A Peto. Whole-genome sequencing for prediction of mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. 15(10): 1193–1202, .

[28] World Health Organization. Global Tuberculosis Report 2024. 2024. URL `https://www.who.int/publications/i/item/9789240101531`.

[29] Christoph Lange, Dumitru Chesov, Jan Heyckendorf, Chi C. Leung, Zarir Udwadia, and Keertan Dheda. Drug-resistant tuberculosis: An update on disease burden, diagnosis and treatment. *Respirology*, 23(7):656–673, 2018. ISSN 1440-1843. doi: 10.1111/resp.13304. URL `https://pubmed.ncbi.nlm.nih.gov/29641838/`.

[30] Maha Farhat, Helen Cox, Marwan Ghanem, Claudia M Denkinger, Camilla Rodrigues, Mirna S Abd El Aziz, Handaa Enkh-Amgalan, Debrah Vambe, Cesar Ugarte-Gil, Jennifer Furin, and others. Drug-resistant tuberculosis: a persistent global health concern. *Nature Reviews Microbiology*, 22(10):617–635, 2024.

[31] Mahbuba Tasmin and Anna G. Green. Beyond Sequence-Only Models: Leveraging Protein Structure for Antibiotic Resistance Prediction in Sparse Genomic Datasets. In *Proceedings of the ICLR 2025 Machine Learning for Genomics Explorations Workshop*, 2025.

[32] Yann Blouin, Yolande Hauck, Charles Soler, Michel Fabre, Rithy Vong, Céline Dehan, Géraldine Cazajous, Pierre-Laurent Massoure, Philippe Kraemer, Akinbowale Jenkins, Eric Garnotel, Christine Pourcel, and Gilles Vergnaud. Significance of the identification in the horn of africa of an exceptionally deep branching mycobacterium tuberculosis clade. 7(12):e52841. Publisher: Public Library of Science.

[33] J M Bryant, A C Schurch, H van Deutekom, S R Harris, J L de Beer, V de Jager, K Kremer, S A van Hijum, R J Siezen, M Borgdorff, S D Bentley, J Parkhill, and D van Soolingen. Inferring patient to patient transmission of mycobacterium tuberculosis from whole genome sequencing data. 13:110.

[34] Anirvan Chatterjee, Kayzad Nilgiriwala, Dhananjaya Saranath, Camilla Rodrigues, and Nerges Mistry. Whole genome sequencing of clinical strains of *Mycobacterium tuberculosis* from mumbai, india: A potential tool for determining drug-resistance and strain lineage. 107:63–72. ISSN 1472-9792. doi: 10.1016/j.tube.2017.08.002. URL `https://www.sciencedirect.com/science/article/pii/S1472979217301786`.

[35] Taane G Clark, Kim Mallard, Francesc Coll, Mark Preston, Samuel Assefa, David Harris, Sam Ogwang, Francis Mumbowa, Bruce Kirenga, Denise M O'Sullivan, Alphonse Okwera, Kathleen D Eisenach, Moses Joloba, Stephen D Bentley, Jerrold J Ellner, Julian Parkhill, Edward C Jones-López, and Ruth McNerney. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. 8(12):e83012. Publisher: Public Library of Science.

[36] K A Cohen, T Abeel, A Manson McGuire, C A Desjardins, V Munsamy, T P Shea, B J Walker, N Bantubani, D V Almeida, L Alvarado, S B Chapman, N R Mvelase, E Y Duffy, M G Fitzgerald, P Govender, S Gujja, S Hamilton, C Howarth, J D Larimer, K Maharaj, M D Pearson, M E Priest, Q Zeng, N Padayatchi, J Grosset, S K Young, J Wortman, K P Mlisana, M R O'Donnell, B W Birren, W R Bishai, A S Pym, and A M Earl. Evolution of extensively drug-resistant tuberculosis over four decades: Whole genome sequencing and dating analysis of mycobacterium tuberculosis isolates from KwaZulu-natal. 12(9):e1001880.

[37] Francesc Coll, Jody Phelan, Grant A Hill-Cawthorne, Mridul B Nair, Kim Mallard, Shahjahan Ali, Abdallah M Abdallah, Saad Alghamdi, Mona Alsomali, Abdallah O Ahmed, Stephanie Portelli, Yaa Oppong, Adriana Alves, Theolis Barbosa Bessa, Susana Campino, Maxine Caws, Anirvan Chatterjee, Amelia C Crampin, Keertan Dheda, Nicholas Furnham, Judith R Glynn, Louis Grandjean, Dang Minh Ha, Rumina Hasan, Zahra Hasan, Martin L Hibberd, Moses Joloba, Edward C Jones-López, Tomoshige Matsumoto, Anabela Miranda, David J Moore, Nora Mocillo, Stefan Panaiotov, Julian Parkhill, Carlos Penha, João Perdigão, Isabel Portugal, Zineb Rchiad, Jaime Robledo, Patricia Sheen, Nashwa Talaat Shesha, Frik A Sirgel, Christophe Sola, Erivelton Oliveira Sousa, Elizabeth M Streicher, Paul Van Helden, Miguel Viveiros, Robert M Warren, Ruth McNerney, Arnab Pain, and Taane G Clark. Genome-wide analysis of multi- and extensively drug-resistant mycobacterium tuberculosis. 50(2):307–316.

[38] James J Davis, Alice R Wattam, Ramy K Aziz, Thomas Brettin, Ralph Butler, Rory M Butler, Philippe Chlenski, Neal Conrad, Allan Dickerman, Emily M Dietrich, Joseph L Gabbard, Svetlana Gerdes, Andrew Guard, Ronald W Kenyon, Dustin Machi, Chunhong Mao, Dan Murphy-Olson, Marcus Nguyen, Eric K Nordberg, Gary J Olsen, Robert D Olson, Jamie C Overbeek, Ross Overbeek, Bruce Parrello, Gordon D Pusch, Maulik Shukla, Chris Thomas, Margo VanOeffelen, Veronika Vonstein, Andrew S Warren, Fangfang Xia, Dawen Xie, Hyunseung Yoo, and Rick Stevens. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. 48:D606–D612. ISSN 0305-1048. doi: 10.1093/nar/gkz943. URL https://doi.org/10.1093/nar/gkz943.

[39] Keertan Dheda, Jason D. Limberis, Elize Pietersen, Jody Phelan, Aliasgar Esmail, Maia Lesosky, Kevin P. Fennelly, Julian te Riele, Barbara Mastrapa, Elizabeth M. Streicher, Tania Dolby, Abdallah M. Abdallah, Fathia Ben-Rached, John Simpson, Liezel Smith, Tawanda Gumbo, Paul van Helden, Frederick A. Sirgel, Ruth McNerney, Grant Theron, Arnab Pain, Taane G. Clark, and Robin M. Warren. Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. 5(4):269–281. ISSN 2213-2600, 2213-2619. doi: 10.1016/S2213-2600(16)30433-7. URL https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(16)30433-7/fulltext. Publisher: Elsevier.

[40] Jennifer L Gardy, James C Johnston, Shannan J Ho Sui, Victoria J Cook, Lena Shah, Elizabeth Brodkin, Shirley Rempel, Richard Moore, Yongjun Zhao, Robert Holt, Richard Varhol, Inanc Birol, Marcus Lem, Meenu K Sharma, Kevin Elwood, Steven J M Jones, Fiona S L Brinkman, Robert C Brunham, and Patrick Tang. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. 364(8):730–739.

[41] Nathan D. Hicks, Jian Yang, Xiaobing Zhang, Bing Zhao, Yonatan H. Grad, Liguo Liu, Xichao Ou, Zhili Chang, Hui Xia, Yang Zhou, Shengfen Wang, Jie Dong, Lilian Sun, Yafang Zhu, Yanlin Zhao, Qi Jin, and Sarah M. Fortune. Clinically prevalent mutations in mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. 3(9):1032–1042. ISSN 2058-5276. doi: 10.1038/s41564-018-0218-3. URL https://www.nature.com/articles/s41564-018-0218-3. Publisher: Nature Publishing Group.

[42] M Merker, C Blin, S Mona, N Duforet-Frebourg, S Lecher, E Willery, M G Blum, S Rusch-Gerdes, I Mokrousov, E Aleksic, C Allix-Beguec, A Antierens, E Augustynowicz-Kopec,

M Ballif, F Barletta, H P Beck, C E Barry, 3rd, M Bonnet, E Borroni, I Campos-Herrero, D Cirillo, H Cox, S Crowe, V Crudu, R Diel, F Drobniewski, M Fauville-Dufaux, S Gagneux, S Ghebremichael, M Hanekom, S Hoffner, W W Jiao, S Kalon, T A Kohl, I Kontsevaya, T Lillebaek, S Maeda, V Nikolayevskyy, M Rasmussen, N Rastogi, S Samper, E Sanchez-Padilla, B Savic, I C Shamputa, A Shen, L H Sng, P Stakenas, K Toit, F Varaine, D Vukovic, C Wahl, R Warren, P Supply, S Niemann, and T Wirth. Evolutionary history and global spread of the mycobacterium tuberculosis beijing lineage. 47(3):242–249.

[43] Jody E. Phelan, Dodge R. Lim, Satoshi Mitarai, Paola Florez de Sessions, Ma Angelica A. Tujan, Lorenzo T. Reyes, Inez Andrea P. Medado, Alma G. Palparan, Ahmad Nazri Mohamed Naim, Song Jie, Edelwisa Segubre-Mercado, Beatriz Simoes, Susana Campino, Julius C. Hafalla, Yoshiro Murase, Yuta Morishige, Martin L. Hibberd, Seiya Kato, Ma Cecilia G. Ama, and Taane G. Clark. Mycobacterium tuberculosis whole genome sequencing provides insights into the manila strain and drug-resistance mutations in the philippines. 9(1):9305. ISSN 2045-2322. doi: 10.1038/s41598-019-45566-5. URL https://www.nature.com/articles/s41598-019-45566-5. Publisher: Nature Publishing Group.

[44] Timothy M. Walker, Camilla LC Ip, Ruth H. Harrell, Jason T. Evans, Georgia Kapatai, Martin J. Dedicoat, David W. Eyre, Daniel J. Wilson, Peter M. Hawkey, Derrick W. Crook, Julian Parkhill, David Harris, A. Sarah Walker, Rory Bowden, Philip Monk, E. Grace Smith, and Tim EA Peto. Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study. 13(2):137–146, . ISSN 1473-3099, 1474-4457. doi: 10.1016/S1473-3099(12)70277-3. URL https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(12)70277-3/fulltext. Publisher: Elsevier.

[45] Kurt R. Wollenberg, Christopher A. Desjardins, Aksana Zalutskaya, Vervara Slodovnikova, Andrew J. Oler, Mariam Quiñones, Thomas Abeel, Sinead B. Chapman, Michael Tartakovsky, Andrei Gabrielian, Sven Hoffner, Aliaksandr Skrahin, Bruce W. Birren, Alexander Rosenthal, Alena Skrahina, and Ashlee M. Earl. Whole-genome sequencing of mycobacterium tuberculosis provides insight into the evolution and genetic composition of drug-resistant tuberculosis in belarus. 55(2):457–469. doi: 10.1128/jcm.02116-16. URL https://journals.asm.org/doi/full/10.1128/jcm.02116-16. Publisher: American Society for Microbiology.

[46] Hongtai Zhang, Dongfang Li, Lili Zhao, Joy Fleming, Nan Lin, Ting Wang, Zhangyi Liu, Chuanyou Li, Nicholas Galwey, Jiaoyu Deng, Ying Zhou, Yuanfang Zhu, Yunrong Gao, Tong Wang, Shihua Wang, Yufen Huang, Ming Wang, Qiu Zhong, Lin Zhou, Tao Chen, Jie Zhou, Ruifu Yang, Guofeng Zhu, Haiying Hang, Jia Zhang, Fabin Li, Kanglin Wan, Jun Wang, Xian-En Zhang, and Lijun Bi. Genome sequencing of 161 mycobacterium tuberculosis isolates from china identifies genes and intergenic regions associated with drug resistance. 45(10):1255–1260.

[47] Matteo Zignol, Andrea Maurizio Cabibbe, Anna S. Dean, Philippe Glaziou, Natavan Alikhanova, Cecilia Ama, Sönke Andres, Anna Barbova, Angeli Borbe-Reyes, Daniel P. Chin, Daniela Maria Cirillo, Charlotte Colvin, Andrei Dadu, Andries Dreyer, Michèle Driesen, Christopher Gilpin, Rumina Hasan, Zahra Hasan, Sven Hoffner, Alamdar Hussain, Nazir Ismail, S. M. Mostofa Kamal, Faisal Masood Khanzada, Michael Kimerling, Thomas Andreas Kohl, Mikael Mansjö, Paolo Miotto, Ya Diul Mukadi, Lindiwe Mvusi, Stefan Niemann, Shaheed V. Omar, Leen Rigouts, Marco Schito, Ivita Sela, Mehriban Seyfaddinova, Girts Skenders, Alena Skrahina, Sabira Tahseen, William A. Wells, Alexander Zhurilo, Karin Weyer, Katherine Floyd, and Mario C. Raviglione. Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. 18(6):675–683. ISSN 1473-3099, 1474-4457. doi: 10.1016/S1473-3099(18)30073-2. URL https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(18)30073-2/fulltext. Publisher: Elsevier.

[48] CRyPTIC Consortium. A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics. *PLoS biology*, 20 (8):e3001721, 2022.

[49] Stewart T Cole, Roland Brosch, Julian Parkhill, Thierry Garnier, Carol Churcher, Duncan Harris, Stephen V Gordon, Karin Eiglmeier, Sacha Gas, Clifton E Barry III, and others. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393 (6685):537–544, 1998. doi: 10.1038/31159.

[50] Matthew Ezewudo, Amanda Borens, Álvaro Chiner-Oms, Paolo Miotto, Leonid Chindelevitch, Angela M Starks, Debra Hanna, Richard Liwski, Matteo Zignol, Christopher Gilpin, and others. Integrating standardized whole genome sequence analysis with a global Mycobacterium tuberculosis antibiotic resistance knowledgebase. *Scientific reports*, 8(1):15382, 2018.

[51] Luca Freschi, Roger Vargas Jr, Ashaque Husain, SM Mostofa Kamal, Alena Skrahina, Sabira Tahseen, Nazir Ismail, Anna Barbova, Stefan Niemann, Daniela Maria Cirillo, and others. Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. *Nature communications*, 12(1):6099, 2021.

[52] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, September 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty560. URL https://doi.org/10.1093/bioinformatics/bty560.

[53] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, 2013. URL https://arxiv.org/abs/1303.3997.

[54] Yu Wang, Zhonghua Jiang, Pengkuan Liang, Zhuochong Liu, Haoyang Cai, and Qun Sun. TB-DROP: deep learning-based drug resistance prediction of Mycobacterium tuberculosis utilizing whole genome mutations. *BMC Genomics*, 25(1):167, 2024. ISSN 1471-2164. doi: 10.1186/s12864-024-10066-y. URL https://pubmed.ncbi.nlm.nih.gov/38347478/.

[55] Anna G. Green, Chang Ho Yoon, Michael L. Chen, Yasha Ektefaie, Mack Fina, Luca Freschi, Matthias I. Gröschel, Isaac Kohane, Andrew Beam, and Maha Farhat. A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis. *Nature Communications*, 13(1):3817, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-31236-0. URL https://doi.org/10.1038/s41467-022-31236-0.

[56] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cyril J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and others. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. doi: 10.1093/bioinformatics/btp163. URL https://doi.org/10.1093/bioinformatics/btp163.

[57] Matthias I Gröschel, Martin Owens, Luca Freschi, Roger Vargas, Maximilian G Marin, Jody Phelan, Zamin Iqbal, Avika Dixit, and Maha R Farhat. GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome medicine*, 13:1–14, 2021.

[58] Sanjana G. Kulkarni, Anna G. Green, Brendon C. Mann, Samantha Malatesta, Suchitra Kulkarni Goodwin, Nina Cesare, Shandukani Mulaudzi, Noorjahn Rawoot, MIC-ML Consortium, Robin Warren, Karen R. Jacobson, and Maha R. Farhat. Convolutional neural networks quantify antibiotic resistance in Mycobacterium tuberculosis with diagnostic grade accuracy and predict treatment response. *medRxiv*, page 2025.08.05.25333066, January 2025. doi: 10.1101/2025.08.05.25333066. URL http://medrxiv.org/content/early/2025/08/07/2025.08.05.25333066.abstract.

[59] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

[60] Elana Simon and James Zou. InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders. 2025. doi: 10.1038/s41592-025-02836-7. URL https://www.nature.com/articles/s41592-025-02836-7.

[61] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2019. URL https://arxiv.org/abs/1908.10084.

[62] Julian G Saliba, Wenshu Zheng, Qingbo Shu, Liqiang Li, Chi Wu, Yi Xie, Christopher J Lyon, Jiuxin Qu, Hairong Huang, Binwu Ying, et al. Enhanced diagnosis of multi-drug-resistant microbes using group association modeling and machine learning. *Nature Communications*, 16 (1):2933, 2025.

[63] J.T. Den Dunnen, R. Dalgleish, D.R. Maglott, and others. HGVS recommendations for the description of sequence variants: 2016 update. *Human Mutation*, 37(6):564–569, 2016. doi: 10.1002/humu.22981.

[64] Pablo Cingolani, Alannah Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiaoqi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Taylor \& Francis*, 6(2):80–92, 2012. doi: 10.4161/fly.19695. URL `https://doi.org/10.4161/fly.19695`.

[65] Adamandia Kapopoulou, Jocelyne M. Lew, and Stewart T. Cole. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinburgh, Scotland)*, 91(1):8–13, 2011. doi: 10.1016/j.tube.2010.09.006. URL `https://doi.org/10.1016/j.tube.2010.09.006`.

[66] Roger Jr Vargas, Michael J Luna, Luca Freschi, Maximillian Marin, Ruby Froom, Kenan C Murphy, Elizabeth A Campbell, Thomas R Ioerger, Christopher M Sassetti, and Maha R Farhat. Phase variation as a major mechanism of adaptation in Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences*, 120(28):e2301394120, 2023. doi: 10.1073/pnas.2301394120.

[67] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, pages 265–283, 2016. URL `https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf`.

[68] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[69] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A Supplemental

B Catalogue Overview

We downloaded the 2nd edition (2023) of the WHO Mutation Catalogue for *M. tuberculosis* complex and its association with drug resistance in Excel format (version *WHO-UCN-TB-2023.6-eng.xlsx*) [26]. The catalogue includes over 30,000 observed variants from 52,000 isolates worldwide, detailing the frequency of each mutation and its resistance or susceptibility associations, for 15 anti-TB drugs. We created a cleaned, user-friendly CSV version of this catalogue. Each row in the cleaned dataset represents a mutation and its corresponding resistance associations, with one row per mutation per drug. Our cleaned WHO catalogue has the following columns:

1. Variant: Describes the mutation, including the affected gene, the type of mutation (e.g., nucleotide change, amino acid change, or loss of function (`LoF`)), and its position, including any upstream or downstream regions. The notation follows the Human Genome Variation Society (HGVS) [63] guidelines. Examples of the variant column representations are shown below:
   - `bacA_c.1527C>T`: A nucleotide substitution in the bacA gene of the H37Rv reference genome, where cytosine (C) is replaced by thymine (T) at position 1527 in the gene's coding region.
   - `bacA_c.-21G>A`: A nucleotide substitution in the bacA gene of the H37Rv reference genome, where guanine (G) is replaced by adenine (A) 21 bases upstream of the start codon (ATG) in the untranslated region of the gene.
   - `bacA_p.Ala197Val`: An amino acid substitution in the bacA gene of the H37Rv reference genome, where alanine (Ala) is replaced by valine (Val) at position 197 in the protein sequence.
   - `bacA_LoF`: A Loss of Function (LoF) mutation in the bacA gene.
   - `rrs_n.1000G>A`: A nucleotide substitution in the rrs gene of the H37Rv reference genome, where guanine (G) is replaced by adenine (A) at position 1000 in the gene's non-coding RNA region of the sequence.

   The positions in this notation represent the relative positions with respect to the coding regions or start positions of the gene, unlike the genome index column, which represents an absolute position with respect to the H37Rv reference genome.
2. Drug: The antibiotic being considered. The catalogue includes 15 first-line drugs: isoniazid, rifampicin, pyrazinamide, and ethambutol, as well as second-line drugs: streptomycin, ciprofloxacin, delamanid, moxifloxacin, clofazimine, capreomycin, amikacin, kanamycin, ethionamide, bedaquiline, and linezolid.
3. Confidence: The confidence level assigned to the mutation in causing resistance to the drug under consideration. The confidence categories are:
   - 1) Assoc w R
   - 2) Assoc w R - Interim
   - 3) Uncertain significance
   - 4) Not assoc w R - Interim
   - 5) Not assoc w R

   Categories 1 and 2 are associated with resistance, while categories 4 and 5 are not associated and therefore consistent with susceptibility when observed. Category 3 indicates uncertainty regarding the association with resistance or susceptibility.
4. Genome_index: The starting index of the nucleotide subsequence on the forward strand of the H37Rv reference genome containing the mutation. For example, if `ATG` changes to `TGC` at position 32,467, this denotes the absolute position of `A` in the reference genome, marking the start of the mutated subsequence.
5. Gene: The gene in which the mutation has occurred.
6. Ref_nt: Refers to the original nucleotide subsequence in the H37Rv genome, before any mutation at the given genomic index.

7. Alt_nt: Refers to the alternate nucleotide subsequence observed as a mutation from the H37Rv genome, starting at the given genomic index.

The `ref_nt` and `alt_nt` represent nucleotide subsequences even for amino acid changes, as these correspond to mutations at the nucleotide level, regardless of their effect on the reference genome sequence.

The original WHO mutation catalogue consists of two sheets. The *Catalogue_master_file* contains columns for `Gene`, `Variant`, `Drug`, `Confidence`, and additional information. The *Genomic_coordinates* sheet includes columns for `Variant`, `Genome_index`, `Ref_nt`, and `Alt_nt` for all observed *M. tuberculosis* variants. First, we removed rows with missing values in the variant and position columns from the *Catalogue_master_file*. Entries with "combo" in the confidence column were excluded to ensure only clearly defined mutations remained. Then, we merged the sheets into a single dataframe using the variant column as the key. The dataframe includes all the aforementioned columns and was exported as *WHO_resistance_variant_all_2023.csv*.

Table 6: Distribution of resistance-associated variants in the WHO catalogue across all the confidence categories for 15 anti-TB drugs.

Drug	Cat1	Cat2	Cat3	Cat4	Cat5
Amikacin	2	5	8294	620	396
Capreomycin	1208	338	10398	434	388
Ethambutol	52	0	16926	4195	273
Ethionamide	2045	1136	6486	950	10
Isoniazid	3141	378	17089	3400	191
Kanamycin	13	2	8671	652	53
Levofloxacin	43	19	7172	2860	146
Linezolid	7	7	3912	111	13
Moxifloxacin	37	25	6291	2646	129
Pyrazinamide	1477	864	5035	1611	77
Rifampicin	95	358	15836	7933	337
Streptomycin	795	586	8530	963	74
Bedaquiline	638	2044	7754	789	2
Clofazimine	606	1986	19148	1140	40
Delamanid	0	11143	2029	748	0

C Mutation–Phenotype Interpretability Dataset Construction

Codon level mutation grouping in the VCF file A VCF file records each nucleotide change individually, while the WHO mutation catalogue includes both nucleotide and amino acid level mutations. To correctly merge the VCF data with the WHO catalogue, nucleotide mutations affecting the same amino acid residue must be grouped together. This ensures nucleotide-level mutations in the VCF can be mapped to amino acid-level mutations in the WHO catalogue. The steps for combining codons in the VCF files are outlined below:

1. F Valid SNPs are selected for codon-level combination by removing mutations with the `IMPRECISE` flag, excluding insertions (`IC`) and deletions (`DC`), and retaining only those with equal-length `REF` and `ALT` alleles (`len(ALT) = len(REF)`). Indels are retained for separate downstream processing.
2. Each mutation is mapped to its codon by first identifying its gene using a genome position-to-gene dictionary. If mutation is in a coding region, the gene's coordinates are used to locate the codon based on the reading frame. For overlapping genes, codons from all relevant genes are retrieved.
3. SNPs within the same codon are combined into a single entry by merging the `REF` and `ALT` nucleotides into one sequence. For non-consecutive mutations, missing positions are filled with the reference nucleotide, and the position is set to the first nucleotide of the codon. The resulting entries are exported to an updated VCF file named `{filename}_combined_codons.vcf`.

Table 7: Snapshot of mutations for one isolate in CSV format, showing annotations and isoniazid `INH` specific phenotype and confidence. These fields are drug-specific and are populated only when the mutation is listed for the respective drug in the WHO catalogue. The complete file for an isolate includes phenotypes for all 11 drugs.

Mutation	Gene	Position	Mutation Type	INH	INH_Confidence
inhA_c.-154G>A	inhA	1674048	intergenic_variant	R	1) Assoc w R
Rv2752c_p.Glu142*	Rv2752c	3065768	stop_gained	N/A	3) Uncertain significance
dnaA_c.1131C>A	dnaA	4139246	synonymous_variant	S	4) Not assoc w R - Interim
glpK_p.Val460Ala	glpK	4138377	missense_variant	S	5) Not assoc w R
rpsL_c.-165T>C	rpsL	781395	intergenic_region		

VCF annotation. The updated VCF files are annotated with additional mutation details for seamless integration with the WHO catalogue. We use the `snpEff` tool [64] to annotate each mutation with the gene in which it occurs, the resulting amino acid change due to nucleotide substitution(s), and the predicted mutation effect (e.g., missense, nonsense, frameshift). The annotated VCFs are then exported to a new file named `{filename}_combined_codons.eff.vcf`. Mutations affecting multiple genes are annotated with each gene's information and stored as separate records.

Variant standardization and mapping. Each annotated mutation in the VCF is assigned a standardized variant name based on the HGVS guidelines to match the WHO naming. This is achieved by combining the `gene` and `HGVS.c` or `HGVS.p` columns in the annotated VCF as `{gene}_{HGVS.c or p}`, depending on whether the mutation is at the protein level.

The annotated VCF and WHO catalogue files are mapped using the variant name, mutation position, reference allele, and altered allele, with separate mappings for missense and other variants. The final mapped file includes the following columns:

- Variant: The variant name in the HGVS format (e.g., `rpoB_p.Ser450Leu`).
- Gene: The gene where the mutation occurs.
- Mutation_position: The genomic index of the starting position of mutation in the H37Rv reference genome.
- Mutation_type: The effect of the mutation (e.g., `missense_variant`, `synonymous_variant`).
- Drug-specific confidence: Columns for each of the 11 anti-TB drugs, indicating the confidence category for the mutation against the drug.
- Drug-specific phenotype: Columns for each of the 11 anti-TB drugs, indicating whether the mutation confers resistance to the drug:
  - `R` for mutations with confidence `1) Assoc with R` or `2) Assoc with R - Interim`.
  - `S` for mutations with confidence `4) Not assoc with R - Interim` or `5) Not assoc with R`.
  - `N/A` for mutations with confidence `3) Uncertain significance`.

We also provide an API to apply stringent quality thresholds to each mutation in the annotated VCF file before mapping, to filter out low-confidence mutations. The threshold values are detailed in appendix section E.

D Resistance-Conferring Gene Extraction

To manage the resource-intensive nature of our ML models, we select a subset of genes from the H37Rv genome. Following standard practice, we choose genes with putative resistance-conferring mutations. For each drug in the WHO Catalogue [26], we select genes with at least one mutation classified under confidence category `1) Assoc w R`. Operon coordinates were obtained from MycoBrowser [65] and extended $\pm$100 bp to include potential regulatory regions. For genes that fall contiguously on the DNA strand, padding is applied only upstream or downstream to not repeat any regions from the subsequent gene. Example: As *gyrB* and *gyrA* are placed contiguously on the DNA strand, 100bp are added to only upstream of the *gyrB* strand and 100bp are added to only the

downstream of the *gyrA* strand. Table 8 lists the genomic coordinates for each gene, the strand from which the original nucleotide sequence was extracted and direction in which the operon padding is applied.

Table 8: Padded start and end coordinates for each gene or operon region. Strand column depicts the strand from which the original gene specific nucleotide sequence was taken for further processing.

Gene or Operon	Padded Start (genome)	Padded End (genome)	Strand	Padding Applied
rrs	1471746	1473382	pos	+100 upstream
rrl	1473658	1476895	pos	+100 downstream
tlyA	1917840	1918846	pos	+100 upstream and downstream
embC	4239763	4243147	pos	+100 upstream
embA	4243233	4246517	pos	none
embB	4246514	4249910	pos	+100 downstream
ethA	4325904	4327473	neg	+100 upstream
ethR	4327549	4328299	pos	+100 downstream
fabG1	1673340	1674183	pos	+100 upstream
inhA	1674202	1675111	pos	+100 downstream
katG	2153789	2156211	neg	+100 upstream and downstream
eis	2714024	2715432	neg	+100 upstream and downstream
gyrB	5140	7267	pos	+100 upstream
gyrA	7302	9918	pos	+100 downstream
pncA	2288581	2289341	neg	+100 upstream and downstream
rpoB	759707	763325	pos	+100 upstream
rpoC	763370	767420	pos	+100 downstream
gid	4407428	4408302	neg	+100 upstream and downstream
rpsL	781460	782034	pos	+100 upstream and downstream

E Generating Multiple Sequence Alignments from VCF Files

The following sections describe the construction of high-quality multiple sequence alignments for each gene region using VCF files. Two alignment formats are generated: *aligned*, which includes gap characters to ensure all sequences for a gene region have the same length, and *unaligned*, which does not preserve the original sequence lengths and may vary across isolates for the same gene region.

Reference Genome Region Extraction We extract relevant genomic regions from the H37Rv reference genome (GenBank accession: `NC_000962.3` [49]) using the start and end positions of the selected gene regions. The positions within the gene regions are 0-indexed.

Mutation Categorization and Quality Control Filtering Each VCF file is sequentially processed to extract mutations, which are categorized into three types: `ref` for positions with no mutation (sequence matches the reference), `alt` for mutations where the reference nucleotide is replaced by an alternate, and `missing` for low-quality or ambiguous calls (e.g., marked as "N").
To retain only high-confidence mutations, we apply stringent quality filters based on the recommendations in [66]. Mutations are reclassified as `missing` if they fail any of the following thresholds: `QUAL` $>$ 10 for variant call quality, `DP` $>$ 5 to ensure sufficient coverage, `MQ` $>$ 30 to avoid misaligned reads, and `BQ` $>$ 20 for SNPs and MNPs to ensure reliable base support. Additionally, Mutations flagged as `Amb` in the `FILTER` field are classified based on allele frequency (`AF`): `AF` $>$ 0.75 as `alt`, `AF` $<$ 0.25 as `ref`, and `0.25 < AF < 0.75` as `missing`. Mutations with the `IMPRECISE` flag in the `INFO` field are also classified as `missing`.

Mutation Introduction into the Reference Sequence To generate an isolate genome, mutations from the VCF file are systematically applied to the corresponding H37Rv reference gene region. Mutation types, including SNPs and indels, are carefully introduced to ensure accurate alignment. The process involves the following steps:

1. Substitution-Based Mutation Handling: For Sigle Nucleotide Polymorphisms (SNPs), the reference (`REF`) and alternative (`ALT`) alleles are extracted from the VCF. If the mutation is categorized as `alt`, the reference base is replaced with the alternate base; if categorized as `missing`, it is replaced with "`N`". The substitution is applied in place, so the sequence

length remains unchanged. For multi-nucleotide polymorphisms (MNPs), where multiple adjacent bases change, the entire affected region is replaced with the alternate sequence, while preserving sequence length.

2. Deletion Handling: Deletions remove nucleotides from the sequence, shortening it (`ALT` < `REF`). The affected reference nucleotides are identified and replaced with `"-"` to preserve alignment only in the *aligned* sequences. Large deletions extending beyond the defined genetic region are replaced entirely with gaps. If a deletion removes an entire gene or region, the affected sites are logged for reference.

3. Insertion Handling: Insertions introduce nucleotides not present in the reference, resulting in a longer sequence (`ALT` > `REF`). The inserted bases from `ALT` are added immediately after the reference position. If an insertion exceeds `100 bp`, it is replaced with `"N"` to avoid alignment issues. To preserve alignment, corresponding gap characters (`"-"`) are inserted into the reference sequence only in the *aligned* sequences. A dictionary is maintained to record the positions within the gene where insertions occur and the number of nucleotides added. Before generating the final aligned sequences, these insertion sites are revisited to ensure all sequences remain the same length by padding shorter sequences with gap characters.

FASTA Export For each locus, the aligned sequences are exported to a FASTA file. The reference sequence is saved under the label `>MT_H37Rv`, and each isolate's mutated sequence is saved under its unique identifier. Additionally, a CSV file is generated for each gene region to record insertion site positions for easy tracking.

F Protein Translation

The protein sequence generation pipeline starts by selecting alignment slices from the isolate FASTA files using the reference H37Rv coordinates. Then, nucleotide segments are extracted with non-gapped indices and translated into protein sequences using a gap- and ambiguity-tolerant function. We implement the following rules: codons with ambiguous bases (e.g., 'N') were translated to 'X', gaps ('-') leading to in-frame mutations were included, and sequences with frameshifts were flagged and handled appropriately for the downstream model.

We introduce an additional quality filter to ensure no unexpected behavior in protein translation: we require that the translated sequence length matches the expected reference protein length within a 5% margin of error, and that no internal stop codons (except at the terminal position) are present. Sequences failing these criteria are flagged as frameshifted and handled separately. For example, we found that for the *rpoB* dataset, `100%` of translations matched expected lengths.

G Curated List of Protein-Substitution Variants from WHO Catalogue

We filter the WHO Catalogue Master File [26] to capture only protein-level substitutions, characterized by the format gene_p.Aaa123Bbb, where a reference amino acid is substituted at a specific position. For each variant, we parse the 3-letter amino acid codes, convert them to the corresponding 1-letter representation, and extract the mutation position to generate a the mutation label (e.g., p.A109C). This processing yields a clean dataset containing both the original variant notation and its 1-letter mutation equivalent, alongside associated drug, phenotype, and confidence information. The final catalogue is exported as a machine-readable CSV to facilitate downstream protein-level analysis of resistance-conferring mutations in *M. tuberculosis*.

H Computational Pipeline and Models

H.1 Heuristic Model

Rules-based heuristic approaches are in regular use to predict resistance of *M. tuberculosis* from genome sequences [43**?**]. The heuristic model generates resistance predictions for each isolate using mutation-to-phenotype mappings (Table 7). We label an isolate as resistant if it carries at least one

Table 9: Representative snapshot of WHO-listed resistance-associated protein variants across key M. *tuberculosis* genes.

Gene	Drug	Variant	Effect	Confidence	Mutation
eis	Amikacin	eis_p.Trp36Arg	missense_variant	3) Uncertain significance	p.W36R
embB	Ethambutol	embB_p.Met306Val	missense_variant	1) Assoc w R	p.M306V
ethA	Ethionamide	ethA_p.Arg207Gly	missense_variant	1) Assoc w R	p.R207G
gid	Streptomycin	gid_p.Ala134Glu	missense_variant	1) Assoc w R	p.A134E
gyrA	Levofloxacin	gyrA_p.Asp94Gly	missense_variant	1) Assoc w R	p.D94G
inhA	Isoniazid	inhA_p.Ser94Ala	missense_variant	2) Assoc w R - Interim	p.S94A
katG	Isoniazid	katG_p.Ser315Thr	missense_variant	1) Assoc w R	p.S315T
pncA	Pyrazinamide	pncA_p.His57Asp	missense_variant	1) Assoc w R	p.H57D
rpoB	Rifampicin	rpoB_p.Ser450Leu	missense_variant	1) Assoc w R	p.S450L

Table 10: Heuristic model performance per drug. The columns represent total number of isolates (N), true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), sensitivity, and specificity.

Drug	N	TP	FP	TN	FN	Sensitivity	Specificity
Isoniazid	17435	5275	222	6419	300	0.946	0.967
Rifampicin	17581	4612	281	11372	225	0.953	0.976
Ethambutol	15110	2333	802	11233	610	0.793	0.933
Pyrazinamide	12903	1711	339	7584	339	0.835	0.957
Streptomycin	7682	1768	405	4829	648	0.732	0.923
Amikacin	3705	492	56	2908	214	0.697	0.981
Capreomycin	3854	467	119	3060	207	0.693	0.963
Kanamycin	3576	577	60	163	10	0.983	0.731
Moxifloxacin	2868	262	97	255	16	0.942	0.724
Levofloxacin	269	69	18	175	7	0.908	0.907
Ethionamide	2153	543	421	505	82	0.869	0.545

category 1 or 2 resistance mutation found in the WHO catalogue for the chosen drug; otherwise, it is labeled sensitive.

H.2 Neural Network Architectures

Table 11: Neural network architectures used across encoding types.

Model	Layer / Operation	Dimensions or Parameters
CNN (one-hot DNA / protein)	Input tensor	$(C \times L)$, where C=5 (DNA) or 20 (protein)
	Stem conv	$\text{Conv1d}(C, 64, 1)$
	Conv block 1	$\text{Conv1d}(64, 64, 12)$, ReLU, $\text{MaxPool1d}(3)$
	Conv block 2	$\text{Conv1d}(64, 32, 3)$, ReLU, $\text{Conv1d}(32, 32, 3)$, ReLU, $\text{MaxPool1d}(3)$
	Fully connected	$\text{Linear}(\text{flat}, 256)$, ReLU, $\text{Linear}(256, 256)$, ReLU, $\text{Linear}(256, 1)$
CNN (foundation embeddings)	Input tensor	$(d \times L)$, d=320 (ESM-2) or 768 (DNABERT-2)
	Stem conv	$\text{Conv1d}(d, 64, 1)$
	Conv block 1	$\text{Conv1d}(64, 64, 12)$, ReLU, $\text{MaxPool1d}(3)$
	Conv block 2	$\text{Conv1d}(64, 32, 3)$, ReLU, $\text{Conv1d}(32, 32, 3)$, ReLU, $\text{MaxPool1d}(3)$
	Fully connected	$\text{Linear}(\text{flat}, 256)$, ReLU, $\text{Linear}(256, 256)$, ReLU, $\text{Linear}(256, 1)$
Mean-MLP (sequence-mean)	Input vector	(d), d=320 (ESM) or 768 (DNABERT)
	Layers	$\text{Linear}(d, 128)$, ReLU, $\text{Dropout}(0.2)$, $\text{Linear}(128, 1)$

H.3 Model training

For all single-run experiments outside cross-validation, data were split 80/20 into training and test sets. Model hyperparameters are summarized in Section 2.3.

All protein neural models—including CNNs, Transformers, DNABERT, and ESM-based classifiers—were trained using the Adam optimizer with a learning rate of 5×10^{-4}, batch size of 32, and a maximum of 20 epochs. For DNA-based models implemented in TensorFlow, CNNs were trained for up to 175 epochs with learning rate 1.23×10^{-4} and batch size 128, while DNABERT models implemented in PyTorch were trained for 30 epochs with learning rate 5×10^{-5}. SD-DNABERT and MD-DNABERT models used batch sizes of 12 and 10, respectively. A binary cross-entropy loss with class-imbalance weighting (ratio of susceptible to resistant isolates) was used for all neural models. Output-layer biases were initialized to the log-odds of the positive class to stabilize early training under skewed label distributions. Logistic regression baselines were optimized using up to 1000 (DNA) or 5000 (protein) iterations with class weighting set to `balanced`. All runs converged without warnings.

All models were implemented in Python. The SD-LogReg, SD-CNN, and MD-CNN models were implemented using TensorFlow v2.14 [67], whereas all protein models were implemented in PyTorch [68] and scikit-learn [69]. Training and inference were executed on the UMass Unity Supercomputing Cluster. Each job ran on NVIDIA A100 GPUs with 2 GPUs, 8 CPU cores, and 8 GB of system memory for single-drug models, whereas multi-drug DNA models were allocated 100 GB of system memory. Foundation model embedding extraction (ESM and SD-DNABERT) was performed on high-memory GPUs (up to 40 GB), while MD-DNABERT required at least 80 GB due to token-level tensor representations.

Logistic regression models used `max_iter` (5000 for protein, 1000 for DNA) as the stopping criterion, and all runs terminated well before this limit without emitting convergence warnings. DNA-based CNN and DNABERT models employed early stopping based on validation loss, and models which did not stop early were manually inspected for convergence. Early stopping was triggered for all protein CNN and Transformer models. On average, CNNs reached their best validation epoch after 72% of the total training schedule, whereas Transformer models converged even faster, typically by 50% of their allotted epochs. Across all architectures, final validation losses plateaued and ROC–AUC values remained stable in the last epochs, demonstrating that optimization had reached equilibrium without overfitting.

H.4 Explaining sequence models with SHAP

Motivation and design principles. SHAP's DeepExplainer requires a background distribution; too small yields unstable estimates, too large becomes computationally prohibitive. To ensure coverage without duplication, we deduplicate all isolates and partition into disjoint sets: a background (10%, capped at 160) and an explained set (90%). This guarantees that each unique sequence is explained exactly once and never used as its own reference. Splits are stratified by resistance label (at least one R and one S if both exist), seeded, and logged for reproducibility. Our attribution design balances (i) faithfulness to the trained model, (ii) biological fidelity by covering all unique isolates including rare variants, and (iii) computational practicality given high-dimensional embeddings.

Deduplication and disjoint split. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the deduplicated pool with $y_i \in \{0, 1\}$. We sample without replacement to form background indices B (size m) and set the explained indices $E = \{1, \ldots, N\} \setminus B$. We use stratification to ensure both R and S must appear in B when available. Default parameters are $\alpha = 0.10$, $B_{\max} = 160$, and $m_E^{\min} = 1$. An implementation sketch is provided below.

```
def disjoint_bg_explain_indices(N, bg_frac=0.10, max_bg=150, seed=42):
    rng    = np.random.default_rng(seed)
    perm   = rng.permutation(N)
    bgn    = int(round(bg_frac * N))
    bgn    = min(max(bgn, 1), N-1, max_bg)
    bg     = perm[:bgn]
    ex     = perm[bgn:]
    return bg, ex
```

Neural models (CNNs over embeddings). We apply SHAP `DeepExplainer` to trained CNNs over ESM embeddings. Models are wrapped to output $(B, 1)$ tensors, and SHAP values are computed on the explained set. Attributions are aggregated across classes and sliced into gene segments using known per-gene padded lengths.

Linear models. For logistic, ridge, and lasso classifiers, we use SHAP's `LinearExplainer` with two maskers: *Independent* (interventional, "true to the model") and *Impute* (correlation-aware, "true to the data"). This allows comparison with CNN-based models under the same discovery metric. Scores are obtained for each residue in each gene by the maximum absolute attribution across explained isolates.

Post-processing and evaluation. For each gene, residues are ranked by their maximum absolute SHAP value across explained isolates. Recovery is benchmarked against the WHO catalogue, restricted to entries labeled "Assoc w R" or "Assoc w R – Interim" with `intersectional=True`. WHO positions are 1-indexed; we shift by +1 to align with our 0-indexed residues. We report Precision@k, Recall@k, and F1@k per drug.

```
# Build deduplicated pool of train U test
bg_idx, ex_idx = disjoint_bg_explain_indices(N, bg_frac=0.10, max_bg=150)
background = X[bg_idx]; explained = X[ex_idx]

explainer = shap.DeepExplainer(model, [background])
phi = explainer.shap_values([explained], check_additivity=False)[0]
```

```
# Aggregate per-residue and slice by gene
for gene, (s,e) in gene_slices.items():
    scores = np.abs(phi[:, s:e]).max(axis=0)
    evaluate_topk(scores, WHO_sites[gene])
```

Defaults and sensitivity. We fix the background fraction at 10%, cap at 150 sequences, and use a fixed seed. Increasing background size beyond 150 yields marginal differences in rankings but substantially increases runtime. All splits and metadata (#background, #explained, seed) are logged for reproducibility.

H.5 DNABERT Embedding Variants

A single DNA sequence of length $L = 3500$ requires approximately 6 MB of memory when stored in `float16` precision, corresponding to roughly 72 MB for a batch size of 12 sequences. To mitigate this memory bottleneck, embeddings are chunked and stored using memory-mapped files, enabling efficient access without loading the full dataset into RAM. All embedding configurations for the SD-DNABERT-CNN variants are supported through a unified training interface (`train_token`), with `mode` $\in \{$`full`, `pca`, `mean`$\}$.

For MD-DNABERT-CNN, stacking sequence embeddings across all 19 genes requires approximately 1 GB for a batch size of 9 (in `float16`). To manage this overhead, embeddings are similarly chunked and stored as `.npy` files for efficient loading during training.

H.6 ESM Embedding Variants

A single protein sequence of length $L = 800$ requires approximately 0.5 MB of memory when stored in `float16` precision, corresponding to roughly 64 MB for a batch size of 128 sequences. To mitigate this bottleneck, embeddings are chunked and stored using memory-mapped files with similar unified training interface as the DNA data.

I Results and Statistical Significance

I.1 Embedding variants

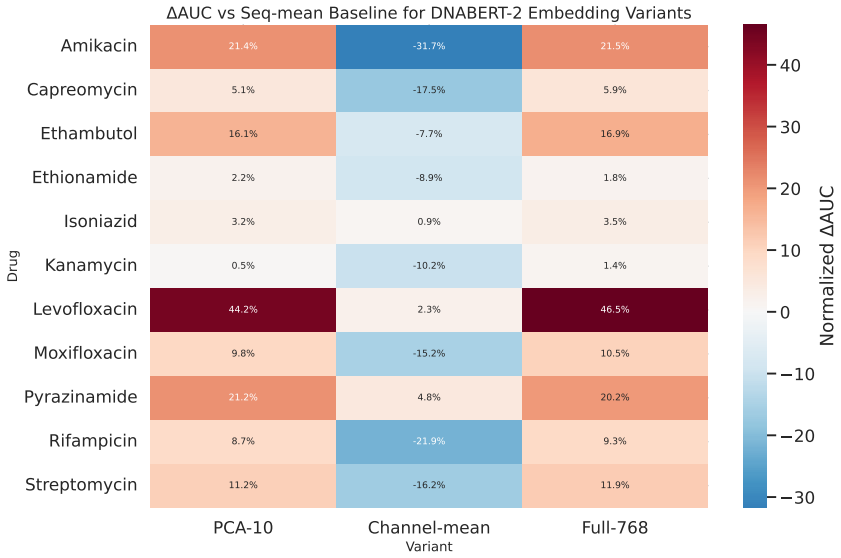


Figure 5: ΔAUC for DNABERT-2 embedding representations. Full-768 and PCA-10 embeddings show consistent gains.

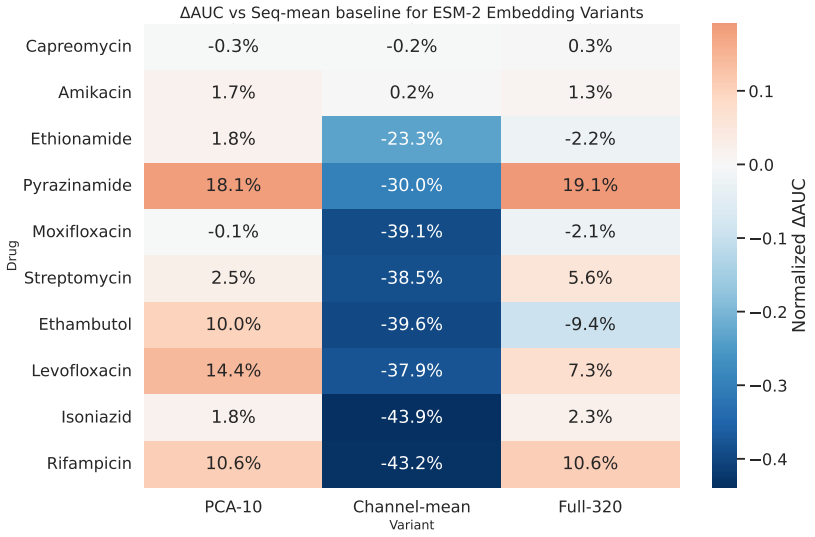


Figure 6: ΔAUC for ESM embedding representations. PCA-10 yields the largest gains for key drugs.

I.2 Predictive Performance: Statistical Significance

Table 12: DNA per-drug mean AUC (± SD) and paired Wilcoxon two-sided p-values versus Logistic Regression baseline. Each value represents the mean performance across five cross-validation folds. The most performant model per drug is shown along with its 95 % confidence interval and corresponding p-value. No multiple-testing correction was applied since comparisons were limited to one comparison per drug.

Drug	Best Model	Mean AUC	95% CI	p-value vs LogReg
Amikacin	SD-LogReg	0.9590 ± 0.0148	[0.9461, 0.9720]	–
Capreomycin	MD-CNN	0.8599 ± 0.0147	[0.8470, 0.8727]	0.125
Ethambutol	SD-CNN	0.9258 ± 0.0032	[0.9230, 0.9286]	0.813
Ethionamide	MD-CNN	0.9161 ± 0.0101	[0.9072, 0.9249]	0.063
Isoniazid	MD-CNN	0.9708 ± 0.0034	[0.9678, 0.9737]	0.063
Levofloxacin	MD-CNN	0.9450 ± 0.0072	[0.9387, 0.9513]	0.625
Moxifloxacin	MD-CNN	0.9298 ± 0.0135	[0.9180, 0.9416]	0.063
Pyrazinamide	SD-CNN	0.9129 ± 0.0154	[0.8994, 0.9264]	0.063
Rifampicin	MD-CNN	0.9769 ± 0.0039	[0.9734, 0.9803]	0.063
Streptomycin	MD-CNN	0.9276 ± 0.0079	[0.9206, 0.9346]	0.125
Kanamycin	MD-CNN	0.8925 ± 0.0183	[0.8764, 0.9086]	0.125

Table 13: Protein per-drug mean AUC (± SD) and paired Wilcoxon p-values versus Logistic Regression baseline. Each value represents the mean performance across five cross-validation folds. The most performant model per drug is shown along with its 95 % confidence interval and corresponding p-value. No multiple-testing correction was applied since comparisons were limited to per-drug model contrasts.

Drug	Best Model	Mean AUC	95% CI	p-value vs LogReg
Amikacin	CNN-PCA10	0.505 ± 0.007	[0.498, 0.511]	0.063
Capreomycin	CNN	0.503 ± 0.003	[0.500, 0.505]	0.125
Ethambutol	CNN-PCA10	0.917 ± 0.013	[0.905, 0.928]	0.438
Ethionamide	CNN-PCA10	0.663 ± 0.038	[0.630, 0.696]	1.000
Isoniazid	CNN	0.918 ± 0.003	[0.915, 0.921]	0.063
Levofloxacin	CNN-PCA10	0.944 ± 0.025	[0.922, 0.966]	0.438
Moxifloxacin	CNN-PCA10	0.846 ± 0.028	[0.821, 0.871]	0.813
Pyrazinamide	CNN-PCA10	0.840 ± 0.023	[0.820, 0.860]	0.813
Rifampicin	CNN	0.967 ± 0.002	[0.965, 0.968]	0.313
Streptomycin	CNN	0.861 ± 0.010	[0.853, 0.869]	0.313

Table 14: Per-drug dna model performance relative to the Logistic Regression baseline.Mean ROC–AUC (± SD) and 95% confidence intervals are reported across five cross-validation folds. p-values correspond to paired Wilcoxon signed-rank tests versus Logistic Regression on identical folds. No multiple-testing correction was applied.

Drug	Model	Mean AUC	SD	95% CI	p vs. LogReg
Amikacin	SD-LogReg	0.9590	0.0148	[0.9461, 0.9720]	–
	SD-CNN	0.8589	0.0195	[0.8419, 0.8760]	0.063
	MD-CNN	0.9176	0.0171	[0.9026, 0.9326]	0.313
	SD-DNABERT-CNN-768	0.8634	0.0425	[0.8261, 0.9006]	0.063
	SD-DNABERT-CNN-PCA10	0.8258	0.0427	[0.7884, 0.8632]	0.063
	MD-DNABERT-CNN	0.5565	0.0891	[0.4784, 0.6346]	0.063
Capreomycin	SD-LogReg	0.8200	0.0125	[0.8090, 0.8309]	–
	SD-CNN	0.8467	0.0242	[0.8255, 0.8679]	0.125
	MD-CNN	0.8599	0.0147	[0.8470, 0.8727]	0.125
	SD-DNABERT-CNN-768	0.8548	0.0226	[0.8350, 0.8745]	0.063
	SD-DNABERT-CNN-PCA10	0.8383	0.0305	[0.8116, 0.8650]	0.438
	MD-DNABERT-CNN	0.8239	0.0251	[0.8020, 0.8459]	0.813
Ethambutol	SD-LogReg	0.9243	0.0067	[0.9184, 0.9302]	–
	SD-CNN	0.9258	0.0032	[0.9230, 0.9286]	0.813
	MD-CNN	0.9253	0.0158	[0.9114, 0.9392]	1.000
	SD-DNABERT-CNN-768	0.9130	0.0039	[0.9092, 0.9169]	0.063
	SD-DNABERT-CNN-PCA10	0.9066	0.0102	[0.8976, 0.9155]	0.063
	MD-DNABERT-CNN	0.9087	0.0137	[0.8967, 0.9207]	0.063
Ethionamide	SD-LogReg	0.6216	0.0377	[0.5886, 0.6546]	–
	SD-CNN	0.8648	0.0340	[0.8349, 0.8946]	0.063
	MD-CNN	0.9161	0.0101	[0.9072, 0.9249]	0.063
	SD-DNABERT-CNN-768	0.6216	0.0178	[0.6060, 0.6372]	1.000
	SD-DNABERT-CNN-PCA10	0.6072	0.0255	[0.5848, 0.6295]	0.813
	MD-DNABERT-CNN	0.9116	0.0084	[0.9042, 0.9190]	0.063
Isoniazid	SD-LogReg	0.9288	0.0054	[0.9240, 0.9335]	–
	SD-CNN	0.9123	0.0119	[0.9019, 0.9227]	0.063
	MD-CNN	0.9708	0.0034	[0.9678, 0.9737]	0.063
	SD-DNABERT-CNN-768	0.8972	0.0109	[0.8877, 0.9068]	0.063
	SD-DNABERT-CNN-PCA10	0.8953	0.0120	[0.8847, 0.9058]	0.063
	MD-DNABERT-CNN	0.8228	0.0159	[0.8088, 0.8367]	0.063
Kanamycin	SD-LogReg	0.8510	0.0275	[0.8269, 0.8751]	–
	SD-CNN	0.8667	0.0294	[0.8409, 0.8924]	0.125
	MD-CNN	0.8925	0.0183	[0.8764, 0.9086]	0.125
	SD-DNABERT-CNN-768	0.8640	0.0357	[0.8327, 0.8953]	0.625
	SD-DNABERT-CNN-PCA10	0.8455	0.0278	[0.8211, 0.8699]	0.313
	MD-DNABERT-CNN	0.8810	0.0168	[0.8663, 0.8958]	0.125

Table 14: Per-drug dna model performance relative to the Logistic Regression baseline.Mean ROC–AUC (± SD) and 95% confidence intervals are reported across five cross-validation folds. p-values correspond to paired Wilcoxon signed-rank tests versus Logistic Regression on identical folds. No multiple-testing correction was applied.

Drug	Model	Mean AUC	SD	95% CI	p vs. LogReg
Levofloxacin	SD-LogReg	0.9316	0.0351	[0.9008, 0.9624]	–
	SD-CNN	0.8504	0.0661	[0.7925, 0.9083]	0.125
	MD-CNN	0.9450	0.0072	[0.9387, 0.9513]	0.625
	SD-DNABERT-CNN-768	0.8250	0.1391	[0.7031, 0.9470]	0.188
	SD-DNABERT-CNN-PCA10	0.8966	0.0399	[0.8616, 0.9316]	0.313
	MD-DNABERT-CNN	0.8672	0.0219	[0.8480, 0.8864]	0.063
Moxifloxacin	SD-LogReg	0.8253	0.0206	[0.8072, 0.8433]	–
	SD-CNN	0.8190	0.0272	[0.7951, 0.8428]	0.438
	MD-CNN	0.9298	0.0135	[0.9180, 0.9416]	0.063
	SD-DNABERT-CNN-768	0.8038	0.0448	[0.7645, 0.8431]	0.438
	SD-DNABERT-CNN-PCA10	0.8116	0.0262	[0.7886, 0.8346]	0.625
	MD-DNABERT-CNN	0.9164	0.0085	[0.9090, 0.9238]	0.063
Pyrazinamide	SD-LogReg	0.8789	0.0114	[0.8689, 0.8888]	–
	SD-CNN	0.9129	0.0154	[0.8994, 0.9264]	0.063
	MD-CNN	0.9113	0.0146	[0.8986, 0.9241]	0.063
	SD-DNABERT-CNN-768	0.8734	0.0238	[0.8525, 0.8943]	1.000
	SD-DNABERT-CNN-PCA10	0.8816	0.0174	[0.8664, 0.8969]	1.000
	MD-DNABERT-CNN	0.5987	0.1761	[0.4443, 0.7530]	0.063
Rifampicin	SD-LogReg	0.9649	0.0051	[0.9605, 0.9694]	–
	SD-CNN	0.9724	0.0014	[0.9712, 0.9736]	0.063
	MD-CNN	0.9769	0.0039	[0.9734, 0.9803]	0.063
	SD-DNABERT-CNN-768	0.9714	0.0038	[0.9681, 0.9747]	0.063
	SD-DNABERT-CNN-PCA10	0.9687	0.0037	[0.9654, 0.9719]	0.313
	MD-DNABERT-CNN	0.8202	0.0228	[0.8002, 0.8402]	0.063
Streptomycin	SD-LogReg	0.9154	0.0046	[0.9114, 0.9194]	–
	SD-CNN	0.9127	0.0120	[0.9022, 0.9232]	0.438
	MD-CNN	0.9276	0.0079	[0.9206, 0.9346]	0.125
	SD-DNABERT-CNN-768	0.8698	0.0481	[0.8277, 0.9120]	0.063
	SD-DNABERT-CNN-PCA10	0.8954	0.0050	[0.8910, 0.8998]	0.063
	MD-DNABERT-CNN	0.9172	0.0092	[0.9091, 0.9253]	0.813

Table 15: Per-drug protein model performance relative to the Logistic Regression baseline.Mean ROC–AUC ($\pm$ SD) and 95% confidence intervals are reported across five cross-validation folds. p-values correspond to paired Wilcoxon signed-rank tests versus Logistic Regression on identical folds. No multiple-testing correction was applied.

Drug	Model	Mean AUC	SD	95% CI	p vs. LogReg
Amikacin	LogReg	0.510	0.008	[0.503, 0.517]	–
	CNN-PCA10	0.505	0.007	[0.498, 0.511]	0.063
	CNN-320	0.504	0.008	[0.497, 0.511]	0.313
	CNN	0.503	0.005	[0.499, 0.508]	0.063
	Transformer	0.497	0.003	[0.494, 0.500]	0.063
Capreomycin	LogReg	0.507	0.003	[0.504, 0.510]	–
	CNN-PCA10	0.499	0.007	[0.493, 0.505]	0.188
	CNN-320	0.498	0.007	[0.492, 0.504]	0.125
	CNN	0.503	0.003	[0.500, 0.505]	0.125
	Transformer	0.494	0.013	[0.483, 0.505]	0.063
Ethambutol	LogReg	0.923	0.004	[0.920, 0.926]	–
	CNN-PCA10	0.917	0.013	[0.905, 0.928]	0.438
	CNN-320	0.901	0.024	[0.879, 0.922]	0.125
	CNN	0.914	0.005	[0.909, 0.918]	0.125
	Transformer	0.527	0.084	[0.453, 0.601]	0.063
Ethionamide	LogReg	0.670	0.022	[0.651, 0.689]	–
	CNN-PCA10	0.663	0.038	[0.630, 0.696]	1.000
	CNN-320	0.621	0.022	[0.602, 0.640]	0.125
	CNN	0.630	0.042	[0.593, 0.667]	0.313
	Transformer	0.556	0.039	[0.522, 0.591]	0.063
Isoniazid	LogReg	0.930	0.009	[0.923, 0.938]	–
	CNN-PCA10	0.917	0.005	[0.913, 0.922]	0.063
	CNN-320	0.910	0.011	[0.900, 0.920]	0.063
	CNN	0.918	0.003	[0.915, 0.921]	0.063
	Transformer	0.872	0.015	[0.859, 0.885]	0.063

Table 15: Per-drug protein model performance relative to the Logistic Regression baseline.Mean ROC–AUC ($\pm$ SD) and 95% confidence intervals are reported across five cross-validation folds. p-values correspond to paired Wilcoxon signed-rank tests versus Logistic Regression on identical folds. No multiple-testing correction was applied.

Drug	Model	Mean AUC	SD	95% CI	p vs. LogReg
Levofloxacin	LogReg	0.929	0.039	[0.894, 0.963]	–
	CNN-PCA10	0.944	0.025	[0.922, 0.966]	0.438
	CNN-320	0.916	0.050	[0.872, 0.960]	0.625
	CNN	0.941	0.048	[0.899, 0.983]	0.625
	Transformer	0.601	0.195	[0.430, 0.772]	0.063
Moxifloxacin	LogReg	0.841	0.015	[0.828, 0.855]	–
	CNN-PCA10	0.846	0.028	[0.821, 0.871]	0.813
	CNN-320	0.822	0.035	[0.790, 0.853]	0.438
	CNN	0.829	0.036	[0.798, 0.861]	0.625
	Transformer	0.638	0.142	[0.514, 0.763]	0.063
Pyrazinamide	LogReg	0.845	0.017	[0.830, 0.859]	–
	CNN-PCA10	0.840	0.023	[0.820, 0.860]	0.813
	CNN-320	0.836	0.017	[0.821, 0.850]	0.813
	CNN	0.788	0.161	[0.647, 0.930]	0.625
	Transformer	0.658	0.058	[0.607, 0.708]	0.063
Rifampicin	LogReg	0.965	0.002	[0.964, 0.967]	–
	CNN-PCA10	0.965	0.004	[0.961, 0.969]	0.625
	CNN-320	0.964	0.003	[0.962, 0.967]	0.625
	CNN	0.967	0.002	[0.965, 0.968]	0.313
	Transformer	0.692	0.211	[0.507, 0.876]	0.063
Streptomycin	LogReg	0.870	0.009	[0.861, 0.878]	–
	CNN-PCA10	0.815	0.028	[0.790, 0.839]	0.063
	CNN-320	0.856	0.011	[0.846, 0.866]	0.063
	CNN	0.861	0.010	[0.853, 0.869]	0.313
	Transformer	0.762	0.037	[0.729, 0.794]	0.063

I.3 Canonical Variant Discovery

Table 16: WHO variants recovered by the best CNN–OHE models at top-k=10. For each drug, the table lists the number of true positives (TP@10), the harmonic mean of precision and recall (F1@10), and representative recovered variants. DNA-based CNN results are shown for single-drug (SD) models, and protein-based CNN results are shown for per-drug models.

Drug	TP@10	F1@10	Representative Identified Variants
DNA SD-CNN–OHE (Task 2: Canonical Variant Discovery)			
Amikacin	3	0.43	*rrs_n.1401A>G*
Kanamycin	1	0.11	*rrs_n.1401A>G*
Capreomycin	2	0.09	*rrs_n.1401A>G*
Ethambutol	7	0.54	*embB_p.Met306Val*, *embB_p.Asp354Ala*
Ethionamide	3	0.03	*inhA_c.-77C>T*
Isoniazid	3	0.17	*katG_p.Ser315Thr*, *inhA_c.-777C>T*
Levofloxacin	5	0.30	*gyrA_p.Asp94Ala*
Moxifloxacin	5	0.30	*gyrA_p.Asp94Ala*
Pyrazinamide	5	0.03	*pncA_p.Asp49Gly*
Rifampicin	5	0.23	*rpoB_p.Ser450Leu*, *rpoB_p.His445Asp*
Streptomycin	4	0.07	*rpsL_p.Lys43Arg*, *gid_p.Ala183Val*
Protein CNN–OHE (Task 2: Canonical Variant Discovery)			
Amikacin	0	0.00	–
Kanamycin	–	–	–
Capreomycin	1	0.17	*tlyA_p.Gly232Asp*
Ethambutol	6	0.75	*embB_p.Asp328Tyr*, *embB_p.Asp354Ala*
Ethionamide	1	0.10	*inhA_p.Ser94Ala*
Isoniazid	2	0.31	*katG_p.Ser315Arg*, *inhA_p.Ser94Ala*
Levofloxacin	7	0.70	*gyrB_p.Asn499Asp*, *gyrB_p.Ser447Phe*
Moxifloxacin	5	0.50	*gyrA_p.Ala90Val*, *gyrA_p.Asp89Asn*
Pyrazinamide	10	0.19	*pncA_p.Asp8Ala*, *pncA_p.Cys138Arg*
Rifampicin	10	0.56	*rpoB_p.Arg448Gln*, *rpoB_p.Asp435Ala*
Streptomycin	7	0.70	*rpsL_p.Lys43Arg*, *gid_p.Lys88Arg*