nature communications

Article

https://doi.org/10.1038/s41467-026-72225-x

# Convolutional neural networks quantify antibiotic resistance in *Mycobacterium tuberculosis* with diagnostic grade accuracy and predict treatment response

Received: 22 October 2025

Accepted: 10 April 2026

Published online: 24 April 2026

Check for updates

Sanjana G. Kulkarni[1], Anna G. Green[1,2], Brendon C. Mann[3], Samantha Malatesta[4], Suchitra Kulkarni-Goodwin[4], Nina Cesare[4], Shandukani Mulaudzi[1], Noorjahn Rawoot[3], MIC-ML Consortium*, Robin M. Warren[3], Karen R. Jacobson[4] & Maha R. Farhat[1,5]

There is considerable interest in training machine learning (ML) models on genomic data that achieve clinical grade diagnostic accuracy. Many successful ML models have been trained and validated on binary tasks because predicting biomedically relevant continuous variables is difficult to optimize. In this work, we present convolutional neural networks (CNNs) that predict minimum inhibitory concentrations (MICs) for eight antibiotics from *Mycobacterium tuberculosis* complex (MTBC) gene sequences. By including evolutionary information, protein biochemical properties, and data augmentation for rare variants, we build models that predict 89% of MICs within one drug concentration doubling. Although trained on ≤ 52% of the World Health Organization's (WHO) MTBC drug resistance mutation catalog data, the CNNs accurately predict the effects of 97% of the catalog's graded mutations. In a cohort of 373 patients with rifampicin-susceptible *M. tuberculosis* infections, higher CNN-predicted rifampicin MICs are associated with unfavorable treatment outcomes, providing additional evidence that subtle differences in MIC below the resistance threshold are clinically relevant. These results demonstrate the value of encoding multiple dimensions of biological data in machine learning of *M. tuberculosis* drug resistance phenotypes and that domain knowledge-inspired machine learning models can be both interpretable and reach clinical grade accuracy.

Pathogen genetic data promises to accelerate the diagnosis of antibiotic resistance in infectious diseases[1,2]. For the *Mycobacterium tuberculosis* complex (MTBC), predicting antibiotic resistance from sequence data is well motivated biologically because it is almost entirely genetically encoded[3,4]. Point-of-care MTBC genetic probe assays currently aid clinicians in identifying resistance to rifampicin within a few hours of sample collection[5,6]. However gaps currently remain in the prediction of

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, USA. [2]Manning College of Information & Computer Sciences, University of Massachusetts, Amherst, USA. [3]DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, SAMRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Depts of Biomedical Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa. [4]Section of Infectious Diseases, Boston Medical Center, Boston, USA. [5]Division of Pulmonary & Critical Care Medicine, Massachusetts General Hospital, Boston, USA.*A list of authors and their affiliations appears at the end of the paper. e-mail: maha_farhat@hms.harvard.edu

https://doi.org/10.1038/s41467-026-72225-x

resistance to other drugs, including recently introduced drugs like bedaquiline.

Drug resistance is typically diagnosed as present or absent at an expert-determined drug "critical concentration" (CC)[7,8]. Despite this convention of convenience, antibiotic resistance is a continuous phenotype, and binary tests can misclassify bacterial samples that lie close to the CC[9]. CCs are also periodically revised depending on the changing distribution of clinical resistance globally[10]. Current data suggests that low levels of resistance, even those below the CC, can increase the likelihood of complete drug resistance developing during treatment[11–13] and have been found to be associated with TB relapse[14]. However, low-level resistance is generally only detectable by measuring minimum inhibitory concentrations (MICs), which is not performed in clinical practice due to high labor and material cost, biohazardous testing conditions, and low reproducibility[15]. Rapid in silico prediction of MICs directly from the genetic sequence could enable tailored TB therapy, reduce treatment failure, and inform antibiotic stewardship globally.

Whole genome sequencing (WGS) generates very high-dimensional data, whereas clinically characterized resistant isolates are often limited in number. This requires developing approaches that can effectively integrate genomic information without relying on large numbers of labeled examples. Rule-based classifiers, which predict isolates to be resistant based on the presence of certain high-confidence mutations, are highly interpretable and easily applied, but cannot capture interactions between mutations or predict quantitative resistance. Synthesizing genetic variant data using statistical models provides greater flexibility for modeling quantitative resistance[7,16–19]. Specifically, machine learning (ML) models are useful because they can consider non-linear and higher order relationships[20]. These include tree-based models such as random forests and gradient boosting machines, which have been used to predict quantitative resistance from k-mers[21]. While k-mer-based approaches bypass variant calling and its associated biases, they treat sequence fragments as independent features and do not explicitly model positional structure. We propose that convolutional neural networks (CNNs), a class of deep learning models, are well suited to quantitative resistance prediction from genetic sequence inputs because they preserve the spatial ordering of nucleotides and learn local, hierarchical patterns[22–24]. This representation may improve generalization to novel variants and contexts.

In this work, we introduce interpretable state-of-the-art CNN models that rapidly and accurately predict MICs from WGS data[7]. Adding amino-acid biophysical properties and lineage information to model inputs improves accuracy over nucleotide-only models. CNNs predict MICs with accuracy on par with in vitro growth based drug susceptibility tests, generate hypotheses for protein active sites, and provide methodological insights that are generalizable to other contexts of machine learning on sequence data.

## Results

### WGS data and MIC distribution

Data from 14,834 MTBC isolates with MICs (Supplementary Data 1) were obtained from the Comprehensive Resistance Prediction for Tuberculosis: An International Consortium (CRyPTIC) project[25], the National Center for Biotechnology Information database, PATRIC[26], the Harvard TB CETR[4], the 2nd edition of the World Health Organization Resistance Mutation Catalog[27], and unpublished sources curated by members of the MIC-ML consortium (Supplementary Data 2). We focused on MICs for the four first-line anti-tubercular drugs – rifampicin, isoniazid, ethambutol, and pyrazinamide–and four second-line drugs–levofloxacin, moxifloxacin, ethionamide, and bedaquiline. Per drug, sample size ranged from 1021 to 12,069, and the frequency of resistance (MIC > critical concentration) ranged from 9.6% to 55% (Table 1, Supplementary Fig. 1). MICs measured in different media were scaled to a single medium for each drug[4]. The lineage (L) distribution is

**Table 1 | Phenotypic summaries and genetic loci used to train models for each of 8 anti-TB drugs**

| Drug | Abbreviation | critical concentration in 7H10 Media (μg/mL) | Number of Isolates | Number Resistant (%) | Highest Tested Concentration (μg/mL) | Drug Mechanism of Action | Input Loci |
|---|---|---|---|---|---|---|---|
| Bedaquiline | BDQ | 0.25 (7H11) | 9127 | 128 (1.4%) | 2 | Inhibits the atpE subunit of the ATP synthase enzyme | *mmpL5-mmpS5, Rv0678, pepQ, atpE* |
| Bedaquiline (+ binary augmentation) | | | 9954 | 955 (9.6%) | | | |
| Ethambutol | EMB | 5 | 9984 | 2747 (28%) | 64 | Disrupts cell wall synthesis by disrupting arabinogalactan synthesis. | *aftA-embC-embA-embB, aftB-ubiA-Rv3807c-glfT2-glf* |
| Ethionamide | ETO | 5 | 10,120 | 1682 (17%) | 50 | Activated by ethA, disrupts mycolic acid synthesis | *fabG1-inhA, ethA, mshA, ethR* |
| Isoniazid | INH | 0.2 | 12,069 | 6613 (55%) | 128 | Activated by katG, disrupts mycolic acid synthesis | *fabG1-inhA, katG-furA, oxyR'-ahpC-ahpD, fabD-acpM-kasA-kasB-accD6* |
| Levofloxacin | LFX | 1 | 7800 | 1465 (19%) | 8 | Inhibit DNA gyrase | *gyrB-gyrA* |
| Moxifloxacin | MXF | 0.5 | 9423 | 1741 (19%) | 32 | | |
| Pyrazinamide | PZA | 100 (MGIT) | 1021 | 528 (52%) | 500 | Activated by pncA, interferes in metabolism | *Rv2042c-pncA, clpC1, panD-panC-Rv3603c, rpsA* |
| Pyrazinamide binary dataset | | | 14,107 | 2820 (20%) | NA | | |
| Rifampicin | RIF | 0.5 | 11,527 | 5717 (50%) | 100 | Inhibits RNA polymerase | *rpoB-rpoC* |

MIC ranges and averages are computed using the midpoints of the MIC ranges of the isolates. MICs for pyrazinamide were in MGIT medium because all isolates were tested MGIT, MICs for bedaquiline were normalized to 7H11 medium, and all other drug MICs were normalized to 7H10 medium. Due to right-censoring, the largest MICs are greater than the highest tested concentration, but are recorded as, for example, ">2 μg/mL". We included upstream promoter regions up to the nearest predicted transcriptional start site[83] for all loci. Contiguous genes in a single locus are delimited by hyphens, and different loci are delimited by commas. *ethA* and *ethR* were encoded separately due to differing strand senses, but the overlapping intergenic region was included in both loci in each pair. Coordinates and gene names were extracted from Mycobrowser (Supplementary Table 1)[84]. The full list of isolates and their MICs is in Supplementary Data 1.

https://doi.org/10.1038/s41467-026-72225-x

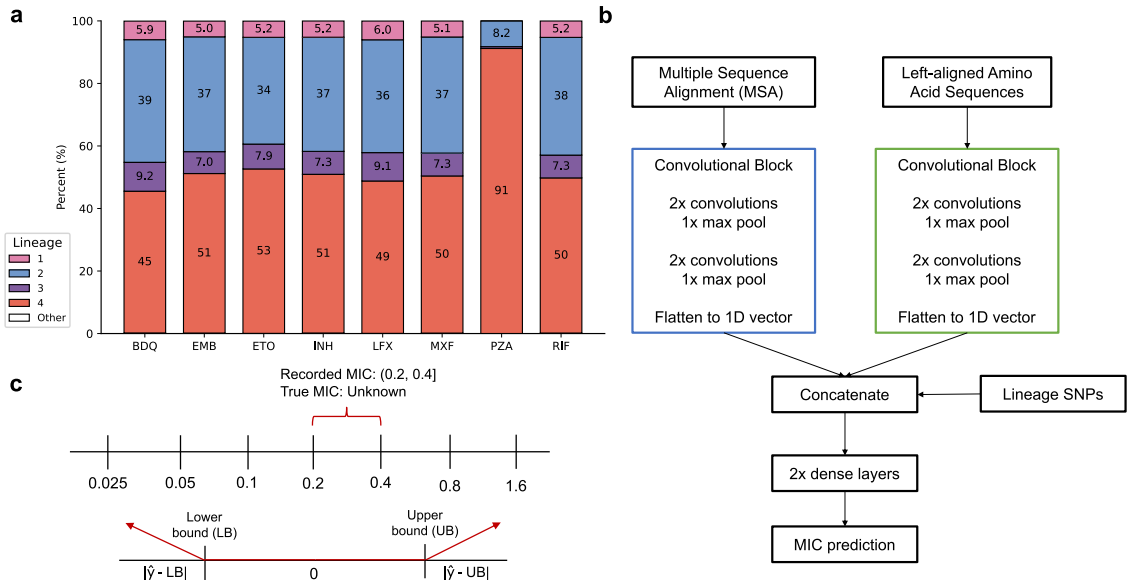



**Fig. 1 | Data and CNN overview. a:** Multi-modal CNN with two convolutional blocks and inclusion of lineage-defining SNPs. **b**: Custom loss function, which only penalizes predictions outside of the associated MIC range for each prediction, using mean absolute error. **c:** Lineage breakdown of the input datasets for the four most common lineages in the data. The "Other" category is composed of *M. bovis* and L6 isolates, which together are present in fewer than 1% of isolates in each dataset. Pyrazinamide MICs were not measured by the CRyPTIC study[25], so the distribution of lineage is highly skewed and not representative of the worldwide lineage distribution[29,85,86].

consistent with the global distribution of sequenced samples, with lineages 2 and 4 dominating (Fig. 1a). The top three sublineages in all the drug datasets are L2.2.1, L4.3.3, and L4.1.2.1. This is expected as these are the most prevalent lineages in the CRyPTIC dataset, which accounts for most of the data used here. For pyrazinamide, because of low MIC sample sizes especially in the susceptible range, we trained a separate CNN on binary phenotypes (Supplementary Fig. 3**, Methods**). Including this binary data, the sample size ranged from 7,800 to 14,107 across the 8 drugs (Table 1).

## CNN with multimodal input more accurately models genotype to phenotype than nucleotide-only input

We trained single-drug CNN models, and for optimal interpretability and computational tractability, relevant core genes were selected as input (Supplementary Table 1)[27]. We implemented a custom error function to accommodate the semi-continuous nature of MICs, including right and left-censoring at the highest and lowest tested drug concentrations(Fig. 1c, Methods). Initially we trained CNNs only on the nucleotide sequence, however this training resulted in models that did not sufficiently distinguish nonsense from missense mutation effects (e.g. for *pncA*, Supplementary Note 1). Inspired by EVEscape[28], we added molecular weight (g/mol), isoelectric point, and hydrophobicity (Eisenberg scale) for each amino acid in the proteins encoded by the relevant genes and passed them through independent convolutions (Fig. 1b). This multimodal input allowed the CNNs to learn from both the nucleotide sequence and the amino acid characteristics, and implicitly, the protein length. The nucleotide + amino acid pyrazinamide MIC model was able to distinguish between the effects of nonsense and missense mutations, unlike the corresponding nucleotide-only model (Supplementary Fig. 3). The former model also has significantly lower mean absolute error (MAE) than the latter model (ΔMAE = 0.33 log2-MIC units, $p = 0.03$, one-sided Welch's t-test).

## Incorporation of lineage boosts performance

The MTBC is a clonal group of species that is classified into 10 major genetically distinct populations (*i.e.*, lineages) and 9 animal-adapted lineages[29]. The genetic background of MTBC isolates, which is most easily described by lineage, is known to be associated with drug resistance phenotypes[30]. This may be due to differences in mutation rates, fitness, transmissibility, or other factors between lineages[29]. Due to linkage disequilibrium in bacteria, variants that affect drug resistance may be associated with lineage markers, making lineage a proxy variable that may improve resistance prediction[31]. We therefore tested if adding lineage information using a validated 62 silent mutation barcode[32] improves resistance prediction accuracy **(Methods)**.

Lineage information significantly improves model performance for bedaquiline, ethambutol, ethionamide, levofloxacin, and moxifloxacin (one-sided Welch's t-test $p < 0.04$, MAE decrease of 0.004-0.09 log2-MIC units in cross-validation), but significantly decreased pyrazinamide MIC performance ($p = 0.02$, MAE increase of 0.07 in cross-validation) due to the small pyrazinamide dataset size and strong lineage skew (Fig. 2a–h). For the binary pyrazinamide CNN, the lineage barcode did not significantly affect performance (two-sided Welch's t-test $p = 0.93$), possibly due to larger sample size and better lineage representation (Supplementary Fig. 3). The lineage barcode did not significantly alter performance for rifampicin and isoniazid ($p \geq 0.23$). Consequently, the final CNNs include lineage information for all drugs except pyrazinamide MIC. The learned relationships between lineages and MICs are generally supported by the average MICs by lineage (Supplementary Note 4). Lineage has the largest improvement on moxifloxacin MIC prediction, both in terms of MAE and the Spearman correlation between saliency scores and average observed MICs. For the two fluoroquinolones and pyrazinamide, the MAEs for lineage 2 isolates are significantly larger than the overall MAEs ($p < 0.03$, two-sided Mann-Whitney U test). For both rifampicin and ethambutol, MAEs for lineage 2 isolates are significantly lower than overall ($p < 0.003$), and MAEs for lineage 4 isolates are significantly larger ($p < 0.02$). Bedaquiline MAE for lineage 3 isolates is also significantly larger than overall ($p = 0.0003$). All other comparisons were not significant.

## CNNs outperform regression and predict MIC within 1 doubling for 89% of isolates

To evaluate the value of the CNN architecture, we benchmarked its error against a simpler linear regression approach (**Methods**). For all 8

https://doi.org/10.1038/s41467-026-72225-x

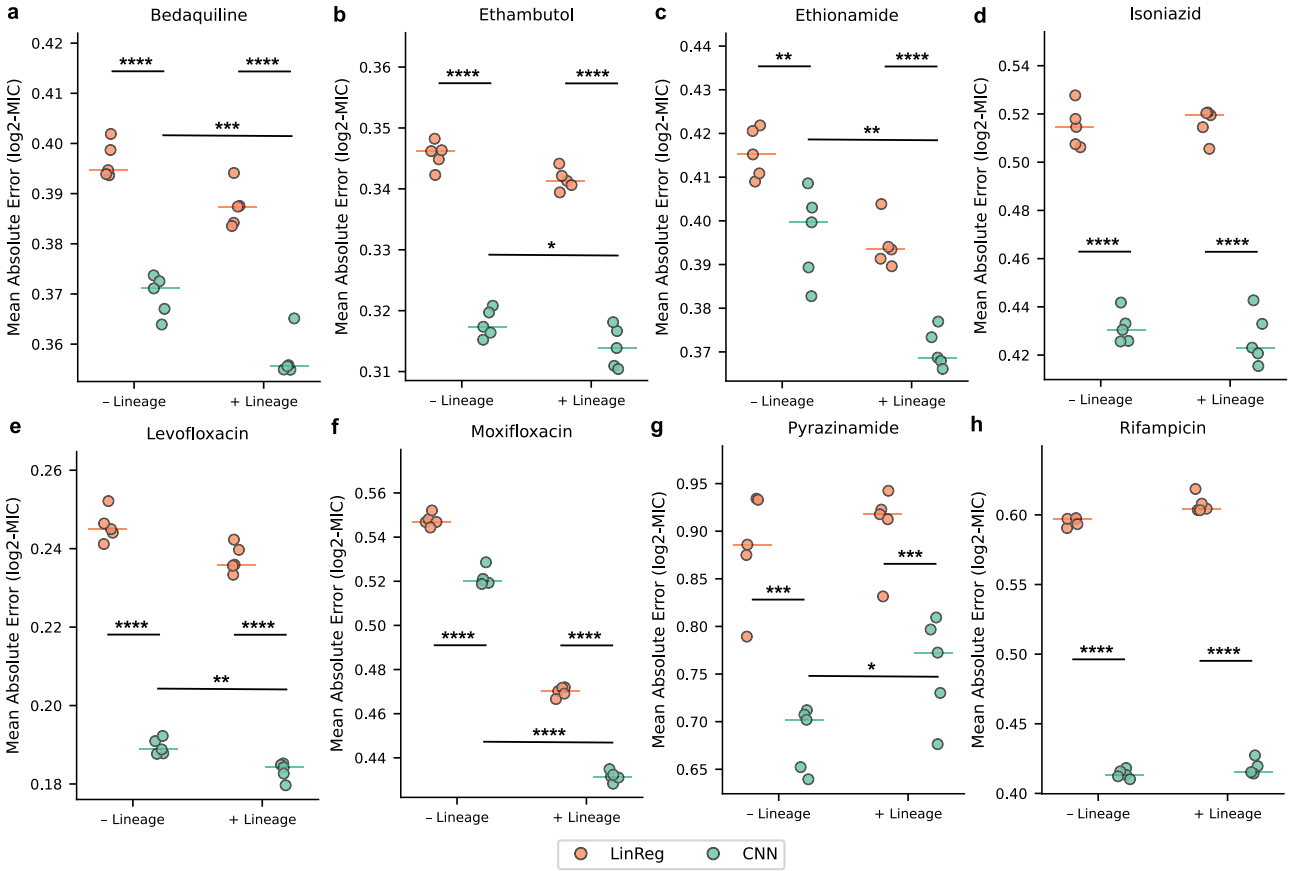



**Fig. 2 | Model metrics between CNN and linear regression on the test set. a–h** Mean absolute error of regression models and CNNs for all 8 drugs, with and without lineage SNPs. The units of mean absolute error are log2-MIC units. Replicates are 5 cross-validation splits. y-axes are different across drugs to better show the variability for single-drug models. *P* values are one-sided Welch's t-tests. For readability, p-value bars are only shown comparing performance between CNN models. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$. No annotation is $p \geq 0.05$.

drugs, the CNNs achieve significantly lower MAE than regression (one-sided Welch's t-test $p$ values $< 2 \times 10^{-4}$) (Fig. 2). CNNs' MAE on the hold-out test sets range from 0.18-0.67 $\log_2$(MIC) (*i.e.*, MIC doublings). Essential agreement (EA)–the proportion of predictions within 1 doubling dilution of the true MIC–is frequently used to evaluate antimicrobial susceptibility testing and its reproducibility as one doubling is the expected error margin for most drugs[33]. EA ranged from 79-93% (average 89%) on the test sets for single-drug CNNs, with the lowest being for pyrazinamide (Supplementary Data 3). The binary pyrazinamide CNN performed significantly better than logistic regression (Supplementary Fig. 3), with an average AUC difference of 2.5 percentage points (one-sided Welch's t-test $p = 2.4 \times 10^{-4}$).

## CNNs achieve state-of-the-art performance

We studied how the CNNs compare with the 2nd version of the WHO MTBC resistance mutation catalog[27], which was trained on larger datasets ranging from 14,000–48,000 isolates per drug. On the hold-out test sets, we predicted MICs, then dichotomized them at the critical concentrations for each drug to simulate binary resistance prediction. We then compared the sensitivities and specificities to the catalog classification method (Supplementary Table 2, Fig. 3a-b). For pyrazinamide, we performed the comparison to the binary CNN because the quantitative model had insufficient training data. Across the drugs, the CNNs have both higher sensitivity and specificity than the catalog (+2.2% and +1.6% on average for sensitivity and specificity, respectively). CNN sensitivities were higher for ethambutol, isoniazid and levofloxacin ($p = 2 \times 10^{-7}$, $1 \times 10^{-14}$ and 0.01 two proportion z-test), and CNN specificity higher for ethambutol ($p = 5.9 \times 10^{-8}$). The CNN sensitivity is lower for ethionamide ($p = 8.3 \times 10^{-4}$), with corresponding greater specificity ($p = 4.3 \times 10^{-14}$), and otherwise, including for the drug bedaquiline, CNN performances were on par with the catalog classification method ($p > 0.05$). The essential agreement rates of the CNNs are comparable to those reported by the CRyPTIC Consortium using XGBoost models trained on whole-genome k-mers, with the CNNs' EAs being on average 0.60 percentage points lower[21]. However, the XGBoost models used non-causal variants in their prediction and have substantially lower specificity for bedaquiline resistance prediction (Supplementary Note 3). For example, the top five k-mers for ethambutol are in *katG*, and none of the reported top k-mers for bedaquiline are in *Rv0678, mmpS5*, or *mmpL5*.

## CNNs attribute resistance to specific variants accurately and exhaustively

To investigate the most important variants identified by the CNNs, we created synthetic sequences with all substitutions and in-frame indels in the WHO MTBC drug resistance mutation catalog[27] and predicted MICs for them using the pre-trained models in **Results Section C**. The catalog consists of two categories of resistance-associated mutations and two categories of mutations that are not associated with resistance that differ by the strength of evidence (associated or associated-interim), as well as a fifth uncertain grading for mutations with insufficient evidence to be classified. The average predicted MIC (transformed to $\log_2$FC) for mutations in each catalog grading correlated with the strength of evidence and direction of association for that grade (Fig. 3c, Supplementary Data 4), including a higher average MIC for "associated with resistance" than for "associated with resistance

https://doi.org/10.1038/s41467-026-72225-x

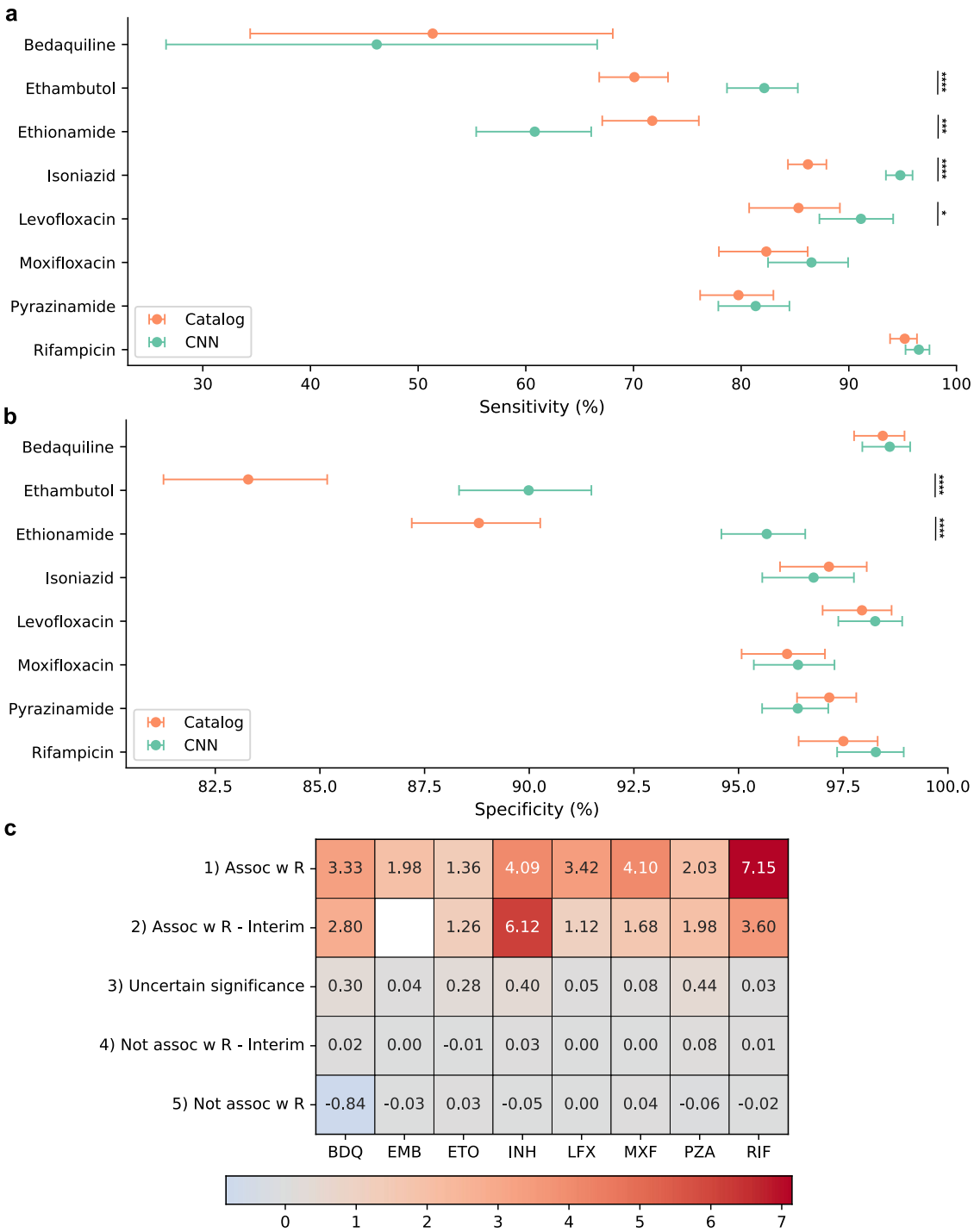



**Fig. 3 | Comparison of binary prediction metrics between the quantitative CNNs and catalog-based classification using the WHO catalog.** Sensitivity (**a**) and specificity (**b**) of binary resistance predictions of the WHO catalog classification method and the MIC predictions of the CNN, dichotomized at the critical concentrations. 95% confidence intervals (CIs) are exact binomial confidence intervals computed using the Clopper-Pearson method. The center of each CI is the exact value computed on the hold-out test set using each classification method. CI significance testing was performed using a one-sided two-proportion z-test for if the CNN value is greater than the catalog value, except for ethionamide in panel **a**, which is a one-sided test for if the CNN sensitivity is less than the catalog sensitivity. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$. No annotation is $p \geq 0.05$. Exact $p$ values in panel a in order down the list of drugs are 0.66 (bedaquiline), 2.3e-7, 8.3e-4, 1.0e-14, 1.5e-2, 6.3e-2, 0.25, and 5.6e-2 (rifampicin). Exact $p$ values in panel b in order down the list of drugs are 0.34 (bedaquiline), 5.9e-8, 4.3e-14, 0.69, 0.28, 0.35, 0.92, and 9.7e-2 (rifampicin). Source data are in Supplementary Data 6. c: Heatmap of the average $\log_2$(fold changes) of predicted MICs for mutations across the five catalog gradings.

interim." The exception was isoniazid, for which the category with the largest predicted average MIC was the 'interim' category. This was because this category contains many *katG* truncations that were graded interim due to rarity but which have higher predicted MICs than *katG* substitutions that predominate in the "associated with resistance" category. For the 28 mutations in the WHO catalog with literature evidence against drug resistance association[34,35], the average CNN-predicted $\log_2$FC was 0.02 (range = -0.73–0.89) (Supplementary

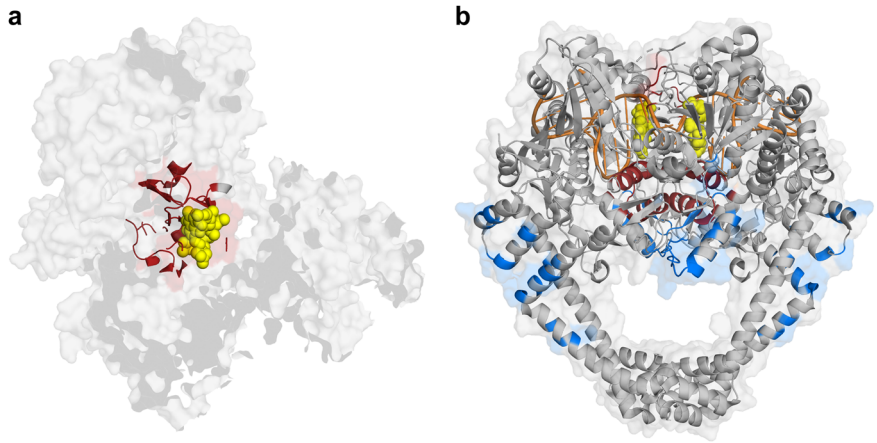



**Fig. 4 | In silico predictions for missense mutations identifies regions of essential genes that are highly associated with resistance.** For all protein structures, hotspot residues are colored red, and coldspot residues are colored blue. **a** Hotspots in rpoB. Yellow = rifampicin. There are no coldspots in rpoB. **b** Hot and coldspots in the gyrA-gyrB complex. Yellow = moxifloxacin, orange = DNA. In silico predictions and clustering results are in Supplementary Data 7.

Data 4). For the binary pyrazinamide CNN, we computed the proportion of variants in each category that were predicted to associate with resistance. This proportion is highest for Group 1 at 96% and decreases monotonically to Group 5 at 0% (Supplementary Data 4).

### CNNs identify mutations associated with subtle shifts in MIC

A key advantage of predicting resistance quantitatively is the identification of borderline or intermediate effects on antibiotic resistance. In the WHO catalog, 25 variants are associated with intermediate resistance – 13 moxifloxacin, 8 isoniazid, and 4 rifampicin (Supplementary Fig. 5). The MIC distribution of isolates with borderline resistance variants often straddle the critical concentration overlap with the susceptible and lower resistance range[36–38], and the CNNs reliably predict these intermediate effects for 7 of the 8 isoniazid variants (1 is absent from the training data and not predicted to elevate MIC) (Supplementary Fig. 5a)[39,40]. García-Marín *et al*[41] identified katG_p.Leu48Pro and katG_p.Gly273Arg as additional putative borderline isoniazid resistance variants, which are not in the WHO catalog. The CNN predicts MICs of 0.20 and 0.30 μg/mL for katG_p.Leu48Pro and katG_p.Gly273Arg, respectively, in agreement with the in vitro results. For moxifloxacin and rifampicin, all the aforementioned mutations are predicted to have intermediate resistance by the CNN, with predicted MICs of 0.10-1.9 μg/mL for moxifloxacin and 0.33-1.9 μg/mL for rifampicin (>1 μg/mL prediction is only for rpoB_p.Ile491Phe) (Supplementary Fig. 5b, c).

### CNNs learn drug binding pockets and the effects of antibiotic resistance variants in both essential and non-essential genes

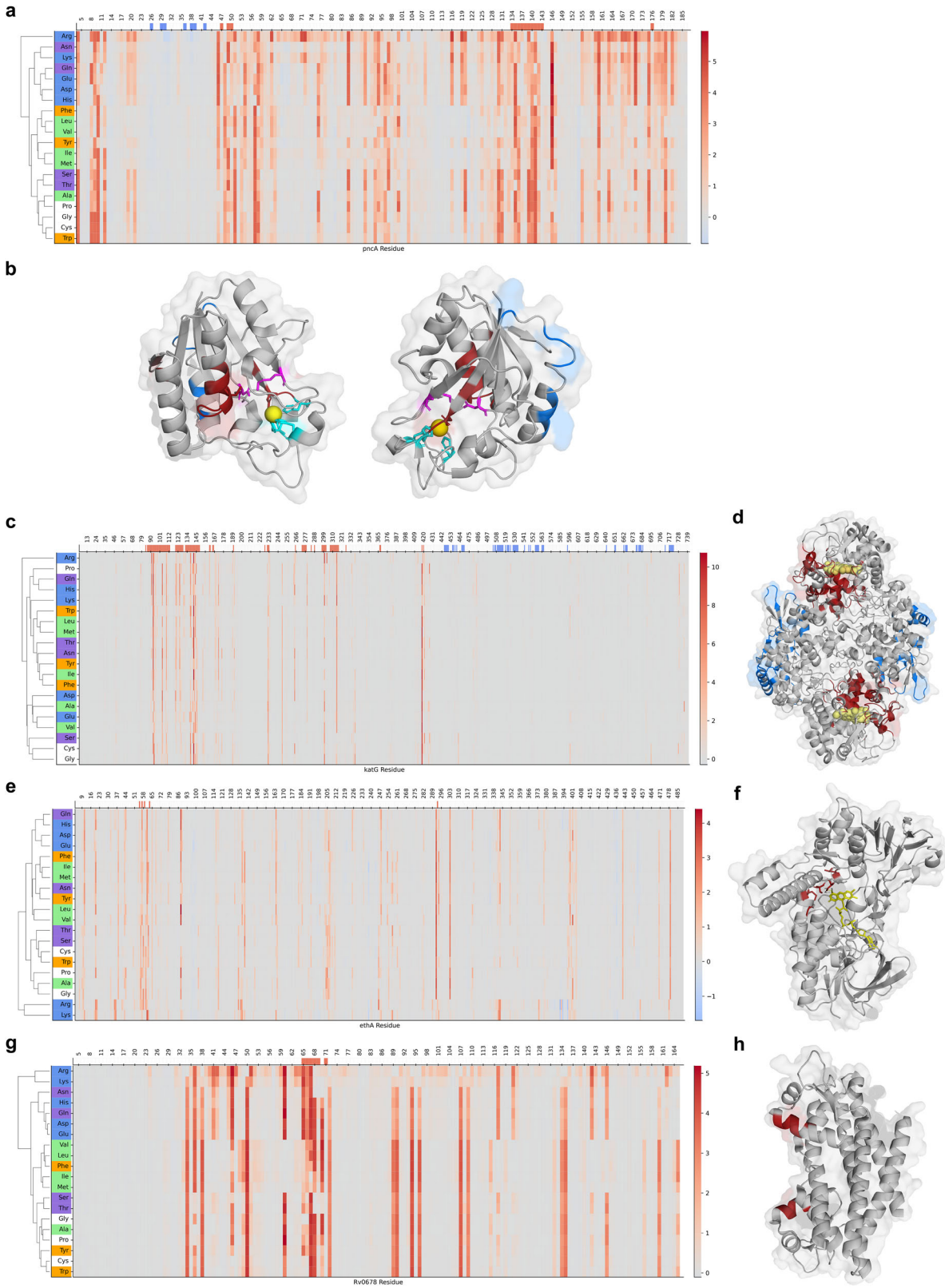
Because of the convolutional component, CNNs can potentially learn the protein region that impacts drug binding or effect. We confirm this and assess the interpretability of the model by studying the 3D distribution of mutations predicted to encode high MIC using the Getis-Ord ($G_i^*$) clustering statistic[42]. $G_i^*$ is computed for the $i^{th}$ residue using its predicted MIC (transformed to log2 of the fold change relative to the prediction for H37Rv) and its distance to every other residue in the protein. We tested for significant clustering of high or low levels of resistance around each residue using a permutation test (**Methods**). We considered the $i^{th}$ residue to be a resistance-associated "hotspot" if $G_i^* > 0$ and $\log_2 FC_i > 0$ and a susceptibility-associated "coldspot" if $G_i^* < 0$ and $\log_2 FC_i < 0$.

The CNN predictions demonstrate significant clustering around known drug binding pockets (also the active sites) in the essential drug targets RpoB (Fig. 4a) and GyrA and GyrB (Fig. 4b). All 27 residues of the rifampicin resistance determining region (RRDR) (coordinates 426-452) in *rpoB* are clustered (*a.k.a.* "hotspots"), and an additional 62 hotspots occur between residues 48 and 674 but are all located near the RRDR in 3D space (Fig. 4a). Similarly, we observe 55 hotspots in the gyrase subunits near the moxifloxacin- and DNA-binding sites, 35 of which are in the quinolone resistance determining regions of *gyrA* and *gyrB* (Fig. 4b)[43]. We do not observe hotspots outside of these active sites for essential proteins.

We observe similar structural clustering in non-essential proteins that are not drug targets. These include PncA and KatG, which activate the pro-drugs pyrazinamide and isoniazid, respectively, and Rv0678, a transcription factor that represses expression of the MmpL5/MmpS5 bedaquiline efflux pump[44]. In PncA, the clustering occurs in the pyrazinamide- and coordinating iron-binding sites (Fig. 5b). In KatG, the hotspot associated with elevated MIC occurs near the heme binding site that mediates drug activation in the N-terminus, and the coldspot associated with lower MIC occurs in the C-terminus (Fig. 5d). The C-terminus is likely important for protein stability and homodimer formation because C-terminus deletions also lead to INH resistance (Supplementary Fig. 3f)[37]. In each Rv0678 monomer, hotspots occur at residues 65-71, which are in the fourth alpha helix that is part of the DNA-binding domain (Fig. 5h)[45]. The full DNA-binding domain extends from residues 34-99, and many substitutions within it are predicted to elevate bedaquiline MIC (Fig. 5g). Mutations in the fourth alpha helix of Rv0678 may reduce DNA binding upstream of the bedaquiline efflux pumps, resulting in higher efflux pump transcription and elevated bedaquiline MIC.

We hypothesized that CNN prediction clustering can characterize new structure-function relationships for proteins known to mediate resistance without a well characterized mechanism. The non-essential gene *ethA*, encodes an FAD-dependent monooxygenase that activates the prodrug ethionamide. Disruptive mutations in *ethA* are the major source of resistance-causing mutations[37]. However, without a solved protein structure, the exact binding sites of EthA and its mechanism of drug activation are not known. Within the confident substructure of EthA predicted by AlphaFold (residues 2-483 with pLDDT > 70 of 489 total residues), we predicted the binding pocket of FAD using AlphaFill, which relies on homology to ligand- and cofactor-binding sites in protein complexes[46]. This prediction placed FAD within 2.1 Å of one of the EthA hotspots determined using Getis-Ord clustering of the predicted MICs (Fig. 5f). These results highlight the ability of machine learning models to identify structurally relevant regions of proteins without being explicitly trained on structural information.

https://doi.org/10.1038/s41467-026-72225-x




## CNNs identify novel resistance associations with ethionamide and other drugs

The WHO mutation catalog grades any loss of function (LoF) mutation in *ethA* as resistance associated by expert opinion that ethionamide activation from its prodrug to drug form is not possible if the main known activating enzyme EthA is not functional[27]. However, a substantial number of ethionamide-susceptible isolates with *ethA* LoF mutations have been previously reported[47,48]. We observe this in in our data (Supplementary Fig. 6a) and find a low specificity of the WHO catalog for ethionamide resistance due to the *ethA* LoF mutations (**Results Section C**). Additional putative activators, such as *Rv3083* (*mymA*)[49], *Rv0565c*[48], *mshA*[50–52], *ndh*[53] may directly or indirectly activate ethionamide in isolates with *ethA* LoF mutations, allowing for continued susceptibility to the drug. To study the role of

**Fig. 5 | In silico predictions for missense mutations identifies regions of non-essential genes that are highly associated with resistance. a**, **c**, **e**, **g** Heatmaps of MIC $log_2$FC for missense mutations, except residue 1, with hierarchical clustering on the amino acid axis for the non-essential proteins PncA (**a**), KatG (**c**), EthA (**e**), and Rv0678 (**g**). The top x-axis of each heatmap shows hot (red) and coldspots (blue) with significant clustering in 3D space. **b** Two views of PncA. Yellow = $Fe^{2+}$ coordinating ion. Cyan = Iron-coordinating residues (49, 51, 57, 71), one of which is a hotspot and is shown in red. Pink = catalytic triad, one of which is a hotspot and is shown in red. The side chains of the 4 iron-coordinating residues and the catalytic triad residues are shown. **d** KatG homodimer with heme substrates (yellow). **f** EthA AlphaFold predicted structure with the predicted FAD binding site by AlphaFill in yellow. The distance between residue 55 (a hotspot) and the FAD is shown. **h** Hotspots in the Rv0678 transcription factor, which is a homodimer. In panels **b**, **d**, **f**, and **h**, red and blue residues are hot- and coldspots, respectively, determined from Getis-Ord clustering. In silico predictions and clustering results are in Supplementary Data 7.

**Table 2 | 20 mutations in *Rv0565c, Rv3083, mshA*, and *ndh* from the 2023 WHO catalog with a predicted ethionamide MIC greater than 1 doubling relative to the wild-type prediction**

| Gene | Variant | Effect | Predicted MIC (μg/mL) |
|---|---|---|---|
| *Rv0565c* | Rv0565c_p.Tyr146dup | inframe_insertion | 5.8 |
| | Rv0565c_p.Trp395* | stop_gained | 3.0 |
| | Rv0565c_p.Gly189Arg | missense_variant | 2.9 |
| | Rv0565c_p.Arg462* | stop_gained | 2.8 |
| | Rv0565c_p.Gly83del | inframe_deletion | 2.7 |
| | Rv0565c_p.Lys408* | stop_gained | 2.6 |
| | Rv0565c_p.Gln351* | stop_gained | 2.5 |
| | Rv0565c_p.Trp266* | stop_gained | 2.5 |
| | Rv0565c_p.TrpLeu266* | stop_gained | 2.5 |
| *Rv3083* | Rv3083_p.Trp293* | stop_gained | 11.4 |
| | Rv3083_p.Trp261* | stop_gained | 2.3 |
| | Rv3083_p.Arg158_His165del | inframe_deletion | 2.3 |
| *mshA* | mshA_p.Gln128* | stop_gained | 7.7 |
| | mshA_p.Tyr123* | stop_gained | 7.1 |
| | mshA_p.Gly311* | stop_gained | 4.8 |
| | mshA_p.Gln205* | stop_gained | 4.4 |
| | mshA_p.Gln363* | stop_gained | 3.5 |
| | mshA_p.Glu220* | stop_gained | 2.9 |
| | mshA_p.Glu353* | stop_gained | 2.4 |
| *ndh* | ndh_p.Tyr342* | stop_gained | 3.7 |

All are graded as Group 3) Uncertain confidence in the catalog. Variants are ordered by gene, then by decreasing predicted MIC. The predicted MIC for H37Rv is 1.1 μg/mL, and the critical concentration for ethionamide is 5 μg/mL.

these genes, we built an expanded model with the WHO Tier 2[27] genes *Rv0565c, Rv3083*, and *ndh* added to the previous Tier 1 loci (*ethA, ethR, fabG1-inhA*, and *mshA*). The expanded model had higher, albeit not statistically significant, accuracy ($p = 0.11$, one sided Welch test), but allowed the identification of 20 novel ethionamide resistance associations in *mshA, Rv0565c, Rv3083*, and *ndh* (Table 2), all of which are graded uncertain in the catalog. Most (N = 16) are nonsense/LoF mutations (Table 2). Isolates with a LoF mutation in either *ethA, mshA, Rv3083, Rv0565c*, or *ndh* have significantly higher MIC than isolates with no LoF mutations in each gene (Supplementary Fig. 6f-j). The larger the proportion of activator enzymes that harbored an LoF, the higher the ethionamide MIC (3 of 5 with LoFs vs. 2 of 5 with LoFs: one-sided Welch's t-test $p = 3.6 \times 10^{-7}$; 2 of 5 vs. 1 of 5 LoFs: $p = 3.0 \times 10^{-3}$). LoFs were observed in up to 3 of the 5 ethionamide activator enzymes, but were not correlated between the activators (Spearman $\rho < 0.40$). These results support that multiple LoF mutations in ethionamide activation pathways are required to affect ethionamide resistance.

Expanding to all 8 drugs, the saliency analysis identifies 247 nucleotides (in 244 unique sites) with significant association with elevated predicted MIC, are observed at a frequency of at least 5 isolates, and are not currently associated in the WHO catalog (Supplementary Data 5). 74% of these are associated with ethionamide and isoniazid resistance, occurring in *ethR* and *katG* respectively, but we also find 5 in *mmpL5* and *Rv0678* related to bedaquiline. The models also identify 23 non-silent mutations from the WHO catalog as associated with hypersusceptibility (Table 3, Supplementary Note 6).

## CNNs can predict MICs below the critical concentrations with moderate accuracy

Predicting high-level resistance, *i.e.*, MICs above the critical concentration, has established value for clinical management of TB. We explore here the accuracy and value of predicting MICs below the critical concentrations in light of a recent study identifying a strong association between sub-critical concentration rifampicin and isoniazid MICs and TB relapse[14]. We leveraged data from a cohort study of 373 participants with rifampicin-susceptible TB confirmed by the commercial assay GeneXpert-TB/RIF or -Ultra and finely measured MICs **(Methods)**[54]. Of the bacterial isolates collected from 372 participants (one did not have MICs measured, and only 262 had measured pyrazinamide MICs), 25 (7%) had isoniazid resistance, 2 (0.5%) had ethambutol resistance, and 2 (0.8%) had pyrazinamide resistance (Fig. 5a). The MIC measurements span 4 doublings below the critical concentration for each drug (Supplementary Fig. 7d).

Essential agreement rates between measured and predicted MICs are 84%, 97%, 94%, and 96% for rifampicin, isoniazid, ethambutol, and pyrazinamide, respectively (Supplementary Fig. 7e–h). With the exception of isoniazid, there are few genetic differences in canonical resistance loci of these isolates (Supplementary Note 8, Supplementary Data 8).

## Predicted rifampicin MIC below the critical concentration is associated with composite TB treatment outcome

Using Cox proportional hazards regression, we studied both the time to sputum clearance of *M. tuberculosis* (*Mtb*) growth on culture (time to culture conversion, TCC) and a composite unfavorable final treatment outcome of death, relapse, or 6-month treatment failure (**Methods**, Supplementary Fig. 7a). We associated either CNN-predicted and phenotypically measured MICs for first-line drugs with outcome, adjusting for patient confounders (Table 4) and predicted drug exposure[55]. Measured pyrazinamide MICs were excluded as they were only available for 262 patients. Of the 373 participants included in the analysis, 27 patients (7%) had an unfavorable outcome (12 treatment failures, 3 deaths, and 12 relapses). Of the 353 participants with valid TCCs (**Methods**), the median TCC was 9 weeks, and 25% had *Mtb*-positive cultures at 12 weeks post enrollment (Supplementary Fig. 7c).

The Cox models identify recognized associations between TCC or unfavorable outcomes and HIV with low CD4 + T cell count, age, smear grade, and time to culture positivity (TTP) (Fig. 6b, Supplementary Fig. 8, Supplementary Note 9). We also validate a recently developed severity marker, high percentage of lung involvement (PLI) on chest X-ray, which is associated with longer TCC (Supplementary Fig. 8b)[56]. Cavitation from chest X-ray was excluded because it was not available for 26 patients, but we note that PLI has been reported to be a better predictor of unfavorable TB treatment outcomes than cavitation[56]. We also note that lineage 2 *Mtb* isolates, which comprised 49% of the

**Table 3 | Mutations with evidence for association with drug hypersusceptibility**

| Drug | Variant | Effect | Predicted MIC (μg/mL) | $Log_2$(FC) |
|---|---|---|---|---|
| Bedaquiline | mmpL5_p.Gln702* | stop_gained | 0.00 | -3.0 |
| | mmpL5_p.Tyr185* | stop_gained | 0.01 | -2.7 |
| | mmpL5_p.Trp39* | stop_gained | 0.01 | -2.6 |
| | mmpL5_p.Gln60* | stop_gained | 0.01 | -2.6 |
| | mmpL5_p.Tyr586* | stop_gained | 0.01 | -2.5 |
| | mmpL5_p.Gln504* | stop_gained | 0.01 | -2.1 |
| | mmpL5_p.Ala136_Ala139del | inframe_deletion | 0.01 | -1.8 |
| | mmpL5_p.Trp833* | stop_gained | 0.01 | -1.5 |
| | mmpL5_p.Met638_Met641del | inframe_deletion | 0.01 | -1.5 |
| | mmpL5_p.Ile555_Gly559del | inframe_deletion | 0.01 | -1.4 |
| | mmpL5_p.Ala350del | inframe_deletion | 0.01 | -1.3 |
| | mmpL5_p.Tyr300* | stop_gained | 0.01 | -1.3 |
| | mmpL5_p.Asp767Ala | missense_variant | 0.02 | -1.1 |
| Ethambutol | embB_p.Leu370Arg | missense_variant | 0.36 | -1.8 |
| | embA_p.Val1073Ile | missense_variant | 0.62 | -1.0 |
| Ethionamide | ethA_p.Trp391Arg | missense_variant | 0.30 | -1.7 |
| | ethR_p.Met1? | start_lost | 0.36 | -1.5 |
| | ethA_p.Leu397Arg | missense_variant | 0.40 | -1.3 |
| | ethR_p.Trp145* | stop_gained | 0.41 | -1.3 |
| | ethA_p.Leu393Arg | missense_variant | 0.42 | -1.2 |
| | ethR_p.Thr3Ala | missense_variant | 0.49 | -1.0 |
| | inhA_c.-158C > T | upstream_gene_variant | 0.49 | -1.0 |
| Levofloxacin | gyrA_p.Asn282Lys | missense_variant | 0.12 | -1.0 |

23 mutations graded as 3) Uncertain significance in the catalog with predicted $log_2$FC ≤ -1. Mutations are ordered by drug then by increasing predicted MIC. Fold change is computed relative to the CNN-predicted MIC for H37Rv for each drug.

**Table 4 | Patient confounding variables tested in Cox models with log-rank test p-values of categorical variables based on Kaplan-Meier estimates of time to culture conversion**

| Covariate | Variable Type | Log-Rank Test p-value in TCC KM Estimate |
|---|---|---|
| Sex | Binary | 0.07 |
| Age | Continuous | NA |
| BMI | Continuous | NA |
| Smoking | Binary | 0.86 |
| HIV positivity and CD4 + T cell count | Categorical, 3 groups: HIV−; HIV + , <200 CD4 + T cells/$mm^3$; HIV + , ≥ 200 CD4 + T cells/$mm^3$ | HIV + , <200 CD4 + T cells/$mm^3$: 0.39 HIV + , ≥ 200 CD4 + T cells/$mm^3$: $3.9 \times 10^{-4}$ |
| Diabetes | Binary | 0.39 |
| Smear Positivity | Binary | $1.8 \times 10^{-9}$ |
| Time to culture positivity (TTP) | Continuous | NA |
| % Lung Involvement > 20% | Binary | $1.7 \times 10^{-6}$ |
| Lineage 2 | Binary | 0.01 |
| F2 Lineage Mixing Metric | Continuous, 0-0.5 | NA |

For the HIV+ variables, the log-rank test was calculated against the HIV− group. The continuous variables tested in the Cox model are listed in the table, but no log-rank test was performed because variables must be categorical. All variables were measured at baseline, within 2 weeks of enrollment.

dataset, were associated with worse composite outcomes than L1, L3, and L4 (Fig. 6b). Lineage 2 has previously been associated with increased risk of tuberculosis treatment failure[57,58] and relapse[57,59,60]. We binarized the predicted rifampicin MIC at the median value of 0.06 μg/mL in MGIT medium due to the small numerical range of values. The small range of predicted values (0.03 - 0.07 μg/mL) makes hazard ratio estimates for a continuous variable unstable and misleading. Measured MICs do not associate with either outcome, but CNN-predicted rifampicin MIC > 0.06 μg/mL is associated with unfavorable outcomes (p = 0.006) (Fig. 6b, Supplementary Fig. 8).

## Discussion

We built a suite of convolutional neural networks that predict quantitative resistance for MTBC with diagnostic-grade accuracy across 8 first- and second-line anti-tubercular drugs, including the critical drug bedaquiline **(Results Section D)**. The CNNs achieve error rates within one concentration doubling, which is comparable to the experimental error rate of measuring MICs. For binary resistance classification, the CNNs perform comparably to or better than the 2023 WHO catalog classification method, despite being trained on a fraction of the data[37]. CNNs have better accuracy for ethambutol, a drug known for genetic determinants that have subtle effects on MICs. CNNs also offer better sensitivity for isoniazid and levofloxacin and higher specificity for ethionamide. CNNs achieve high essential agreement for predicting sub-critical concentration MICs for 4 first-line drugs in a cohort of 373 patients with rifampicin-susceptible TB in South Africa. In this cohort, MIC distributions differ considerably from those used for training the CNNs **(Results Section J)**.

In this cohort, we also find an association between higher CNN-predicted rifampicin MIC below the critical concentration and unfavorable treatment outcomes, adjusted for patient confounders. However, the MIC range was very narrow, and so the estimated hazard ratio may not generalize to other cohorts with wider MIC distributions. While the direction of this estimated effect aligns with previous findings by Colangeli *et al.*[14], in which similar rifampicin concentrations of 0.016-0.125 μg/mL were studied, and supports the hypothesis that sub-critical concentration elevations in MIC may influence outcomes, the effect should be interpreted cautiously due to the limited numerical variability in this dataset. Measured rifampicin MIC was not associated

https://doi.org/10.1038/s41467-026-72225-x

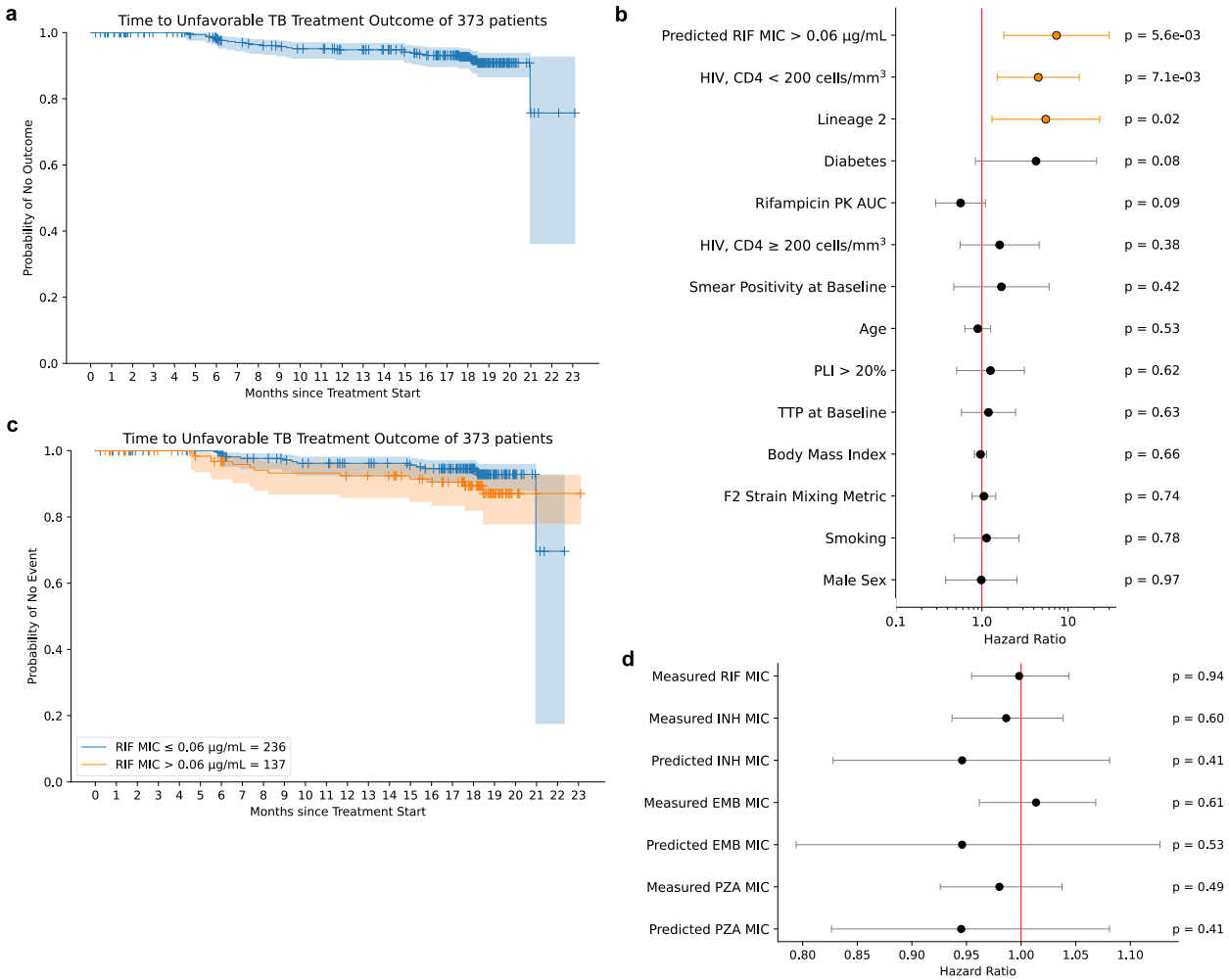



**Fig. 6 | Associations between covariates and composite unfavorable outcomes in multivariate models. a** Kaplan-Meier estimate of event-free survival for 373 participants included in this analysis. **b** Forest plot of hazard ratios for demographic covariates, CNN-predicted rifampicin MIC (binarized at 0.06 μg/mL), and rifampicin drug exposure ("Rifampicin PK AUC," which was predicted from patient demographic data[55].) Hazard ratios for age and Rifampicin PK AUC are per increase in 10. **c** Kaplan-Meier estimate of event-free survival stratified by predicted rifampicin MIC ( > 0.06 μg/mL or ≤ 0.06 μg/mL). **d** Hazard ratios for predicted and measured MICs, except predicted rifampicin MIC, which is in panel (**b**). Hazard ratios for continuous MIC variables are per increase in 0.1 μg/mL. Confidence intervals and two-sided *p* values for the hazard ratios in panels (**b**, **d**) are based on the standard normal approximation (Wald). Standard errors were computed by the Cox regression fitter in lifelines. Variables in panels (**b**, **d**) with $p < 0.05$ are shown in orange, and otherwise in black.

with outcomes, which may be due to noise in MIC testing, especially at low testing concentrations. Time to initial culture conversion has been reported to be a poor proxy of final treatment outcomes, which may be why we detect an association with composite outcomes but not with TCC[61,62]. Unlike the work by Colangeli et al.[14], all participants with isoniazid resistance in our cohort had a favorable outcome, and we did not find an association between isoniazid MIC and unfavorable outcomes. Rifampicin, however, is recognized to be a more important drug in producing a lasting cure than isoniazid, and hence our cohort size may not have had sufficient power to study effects of isoniazid MIC[63]. We note that although HIV+ participants with low CD4 + T cell count had a higher risk of unfavorable outcomes, which is well known in tuberculosis disease[64,65], these participants also have significantly faster time to sputum clearance. This is likely due to the significantly lower bacterial burden at baseline in these participants. Although the bacteria is cleared from sputum, it may persist longer in participants with uncontrolled HIV infection, leading to worse long-term outcomes.

Due to universal limitations in dataset size and sampling, as well as the complexity of biological phenotypes, we hypothesized that ML models would benefit substantially from encoding as much biological information as possible. We show that CNNs cannot learn the different consequences of nonsense and missense mutations on protein function from nucleotide data alone and must be explicitly shown how each mutation affects protein sequence and length. Adding protein sequences and lineage/ancestry information improved model performance and reduced spurious associations. We also show that the core CNN architecture, which allows modeling of higher order interactions between genetic variants, is needed to achieve high performance. Simpler linear regression models that consider variants as independent are consistently less accurate across the 8 drugs we study. Other non-linear models, such as XGBoost models, may provide better accuracy but remain limited by their reliance on tabular representations of independent features. For comparison, we trained an XGBoost model to predict rifampicin MIC. Although it achieved slightly higher essential agreement than regression, it had slightly worse error and was outperformed by the CNN on both metrics, highlighting the importance of capturing local sequence context and positional dependencies among variants.

The prediction accuracy of ML models in biology has often been viewed as being at odds with model interpretability, but in this work, we show that CNNs can combine both state-of-the-art accuracy and

interpretability. CNNs provide evidence to classify many mutations that have been graded uncertain in the WHO MTBC resistance mutation catalog, such as those in low-level activators of ethionamide. The models also identify putative binding pockets in proteins without solved structures, despite not being trained on 3D structures of the proteins. This would be especially important for new antibiotics for which the exact mechanisms are not yet known. By training models on MICs, rather than binary resistance phenotypes, we show that the models can associate genetic variants with drug hypersusceptibility and subtle shifts in MIC.

The CNN models are not without limitations. The models are still limited by the adequacy of training data. For example, due to low rates of bedaquiline resistance, the CNN sensitivity for binary resistance prediction is still low. The bedaquiline CNN is nevertheless state-of-the-art in comparison with the WHO catalog and now offers the prediction of MIC and borderline mutations. CNNs may assume that mutations of different types that occur on the same codon have similar effects, such as the silent variant pncA_c.3 G > A and the missense variant ethR_p.Thr3Ala. Loss of the start codons of *pncA* and *ethR* are associated with pyrazinamide resistance and ethionamide susceptibility, respectively, and the models extrapolate this association to other mutations on the same codons. This problem would be alleviated with richer training data and additional priors that weigh mutations at start codons differently. Ultimately, CNNs cannot learn the biological impact of a mutation unless it or a neighboring mutation in either nucleotide or protein space is associated with the correct phenotype. Encoding additional information on protein stability or drug binding free energy[66] may support models to accurately generalize to unseen or very rare mutations, particularly amino acid substitution mutations.

In this work, we show that machine learning models can reliably and accurately predict quantitative antibiotic resistance for the MTBC with clinical-grade accuracy, and we verify that they learn from relevant genetic variants rather than spurious genetic correlations, setting them apart from previous efforts to predict MIC in the MTBC **(Results Section E)**[21]. We find a significant association between higher predicted rifampicin MIC and unfavorable treatment outcomes among individuals with rifampicin-susceptible *Mtb* infections, providing further evidence that low-level MICs are clinically relevant. There is increasing adoption of targeted sequencing for mutation detection to diagnose antibiotic resistance in the MTBC, and current commercial assays detect mutations in the majority of loci on which the CNN are trained (the CNNs were trained on additional loci to make use of information in non-canonical loci)[67]. CNNs could be deployed in the clinic, along with adequate information on interpretability for clinicians, to provide more comprehensive resistance information to inform treatment and dosing decisions[68–70]. We speculate that the use of these models may now realize the full potential of precision treatment of tuberculosis as they achieve state-of-the-art accuracy for high level resistance and allow for the detection of borderline resistance that has been long elusive due to limited access to MICs. Future work should explore the use of CNNs in clinical trials, either prospectively or in retrospective reanalysis, to study associations with outcome on larger datasets. Further, MIC prediction on the new and repurposed drugs–linezolid, delamanid, and pretomanid–may require different model architectures due to the current extreme rarity of resistant clinical isolates available for training.

## Methods

This research complies with all relevant ethical regulations. The protocol for this project is approved by the South African Medical Research Council Human Research Ethics Committee, the University of Cape Town Human Research Ethics Committee, the Stellenbosch University Human Research Ethics Committee, and the Boston University Institutional Review Board. Written, informed consent to participate was obtained from all participants if ≥ 18 years of age or written individual consent and separate parental consent if < 18 years[54].

### Phenotypic data

MTBC isolates were obtained from the Comprehensive Resistance Prediction for Tuberculosis: An International Consortium (CRyPTIC) project[25], the National Center for Biotechnology Information database, PATRIC[26], and the MIC-ML Consortium (Supplementary Data 1). We trained single-drug models on MICs for eight anti-tubercular drugs–rifampicin, isoniazid, ethambutol, pyrazinamide, ethionamide, levofloxacin, moxifloxacin, and bedaquiline (Table 1, and Supplementary Fig. 1).

When measuring microbial MICs, specific concentrations of each drug are tested by different laboratories (typically 2-fold dilutions). To harmonize phenotypic data tested from different laboratories across testing media, we normalized all MICs to 7H10 medium using the ratio of the critical concentrations in the test medium and 7H10 medium, according to the method in Farhat *et al.* 2019[4]. For pyrazinamide, all MICs were kept in MGIT medium as it was the testing medium for all available isolates, and for bedaquiline, all MICs were normalized to 7H11 medium.

In the MTBC, resistance-conferring mutations arise primarily on the chromosome with no contribution from plasmids and horizontal gene transfer, as in other bacteria. Instead of passing the entire MTBC genome into the models, we selected regions of interest for each drug to reduce computational complexity and overfitting to genetic variants that are difficult to interpret, and aid in targeted sequence-based efforts of diagnosing drug resistance in MTBC[70]. Models were trained on genomic regions known or hypothesized to be relevant to resistance to the 8 drugs, including upstream promoter regions and intergenic regions between contiguous genes (Table 1, Supplementary Table 1). The following quality control steps were performed on a per-drug basis. We excluded isolates with a Group 1 catalog resistance mutation[27] passing the quality control thresholds (**Methods**) and an MIC upper bound less than half of the critical concentration due to potential phenotypic mislabeling. We additionally removed isolates whose MIC range spans the drug's critical concentration from model performance comparisons due to the inability to know if these isolates are drug-resistant or drug-susceptible from training, but we obtained post-training predictions for them. Isolate IDs, lineages, and MICs are in Supplementary Data 1. For all analyses, MICs were log2-transformed (Supplementary Fig. 1). Isolates were split into training (80%), validation (10%), and testing (10%) sets, stratified by binary resistance phenotype and primary lineage. The validation set was used to determine when to stop training the CNNs and the regularization parameter for regression models. The final model performance metrics were computed on the test set. The MIC distributions for all drugs (without augmentation) are not significantly different between the combined train + validation set and the test set by a two-sided Kolmogorov-Smirnov test ($p$-values ≥ 0.99) (Supplementary Fig. 1).

### Phenotypic augmentation for bedaquiline and pyrazinamide

Due to the low number of bedaquiline-resistant isolates, leading to several resistance-conferring variants being absent from the training data (Supplementary Note 2), we augmented the training dataset with 827 bedaquiline-resistant isolates with high-quality binary phenotypes published by the 2nd edition of the WHO mutation catalog[27]. These isolates were given an MIC of >0.25 μg/mL.

Due to the low number of total isolates with pyrazinamide MICs, we trained a separate CNN on binary pyrazinamide resistance phenotypes using the 1021 isolates with MICs, which were binarized at the critical concentration of 100 μg/mL, and 13,086 additional publicly available isolates with high-quality binary phenotypes. This dataset has 20.0% resistant isolates, and the lineage distribution is

more similar to those of the other drug datasets (Supplementary Fig. 4a).

### Variant calling pipeline

Variant calling was performed as follows, with all software installed using conda (v23.10.0). FASTQ files were trimmed with fastp (v0.22.0), dropping trimmed reads shorter than 50 base pairs. Only reads mapping to a custom database built using 155 MTBC genomes with kraken (v2.1.2) were aligned with bwa-mem2 (v2.2.1) to the *Mycobacterium tuberculosis* H37Rv strain complete genome (NCBI RefSeq: NC_000962.3). The custom kraken database is composed of the H37Rv, F11, CTRI-2, and CDC1551 *Mtb* reference genomes and 151 MTBC short-read/long-read hybrid assemblies[71]. After alignment, WGS runs were included if they had a median read depth ≥ 15 and at least 95% of the reference genome covered at a depth of 20x. For samples with multiple available sequencing runs, BAM files passing this alignment criteria were merged before variant calling. Duplicate reads were removed with picard (v2.20.4), followed by variant calling with pilon (v1.24). We excluded isolates with multiple strains, as measured by an F2 metric[72] > 0.1. Variants were annotated with snpEff (v5.2) using the same H37Rv reference sequence (GenBank assembly GCF_000195955.2_ASM19595v2) and the bacterial genetic code. We performed lineage typing using the fast-lineage-caller tool and used the lineages assigned by the Coll 2014 scheme[32].

### Featurizing nucleotide sequences

As input for all models, we passed in a one-hot encoded nucleotide sequence matrix composed of all relevant loci for each drug. A single locus may contain many contiguous genes and their intergenic regions. Non-contiguous loci were passed in as separate channels. Multiple sequence alignments (MSAs) were generated for each contiguous locus using a custom Python script that inserts SNPs and indels from VCF files into the H37Rv reference sequence. Imprecise structural variants (pilon INFO/IMPRECISE = True) were excluded because they were not accurately resolved by the variant caller. We encoded variants with an alternative allele fraction (AF) > 0.75 as the alternative allele and variants with AF ≤ 0.25 as the reference allele. Low-quality SNVs, MNVs, and inframe indels (pilon FILTER ≠ PASS or Amb, pilon QUAL ≤ 10, read depth <5, base quality <20, mapping quality <30, 0.25 < AF ≤ 0.75) were inserted as "N" nucleotides to distinguish them from variants of high confidence. However, we did not encode frameshift indels as missing because the insertion of "N" nucleotides can lead to improper translation to protein sequences. Frameshifts that passed all the QC criteria but had $AF \in (0.25, 0.75]$ were encoded as the alternative allele, and frameshifts that did not pass these QC criteria were left as reference.

The sequence alignment was one-hot encoded into a 5 x L matrix, where L is the length of the aligned locus. The 5 positions represent A, C, G, T, and the gap character, in this order. A = (1,0,0,0,0), C = (0,1,0,0,0), G = (0,0,1,0,0), T = (0,0,0,1,0), '-' = (0,0,0,0,1), and N = (0,0,0,0,0). When there are multiple loci, each locus is a separate channel so that non-contiguous genes are never convolved over together. Shorter loci are padded at the end with gap characters. The final one-hot encoded input matrix for a single CNN sample is 5 x M x P, where M is the longest locus in the set and P is the number of loci. To feed data into regression models, the one-hot encoded matrix for each locus was flattened into a vector of length 5 x $L_i$, where $L_i$ is the length of the $i^{th}$ locus. Vectors for all loci in the set were then concatenated, giving a full length of $\sum_{i=1}^{P} 5 \cdot L_i$. Because they do not contribute to the signal, features with no variation across the training set were removed for computational efficiency.

### Featurizing amino acid sequences

We encoded amino acid biochemical properties, which implicitly encode protein lengths, to add additional biologically relevant information to the models and passed them into an additional convolutional block in the model, with the same layer structure as the nucleotide convolutional block (Fig. 1b). We translated the nucleotide sequences in the MSAs and inspired by EVEscape[28], encoded three features–molecular weight (g/mol), isoelectric point, and hydrophobicity (Eisenberg scale)–as a 3 x K matrix, where K is the length of a given protein. Individual proteins were encoded as separate channels, even if they are translated from the same locus (*i.e.*, rpoB and rpoC proteins are in two separate channels in the amino acid block, while the corresponding gene sequences are concatenated into a single contiguous locus in the nucleotide block). Protein sequences for different isolates were left-aligned (*i.e.*, not a multiple sequence alignment) and then padded to the same length in the same manner as for the nucleotide sequences.

### Lineage SNP inputs

To test the effect of including MTBC lineage information on model performance, we created a binary vector (1 = SNP present, 0 = absent or ambiguous) of 62 lineage SNPs based on the Coll *et al.* SNP scheme for each isolate[32]. The SNPs had to pass the same quality control thresholds previously described for the nucleotide MSA to be considered present. SNPs were concatenated with the flattened output of the convolutions and then passed through the two densely connected layers (Fig. 1a). The distribution of primary lineage designations across all eight single-drug datasets is given in Fig. 1c. Because lineage SNPs do not significantly worsen performance, we included them in all drug models for downstream analyses to reduce overfitting to low-frequency genetic variants[31].

### Model design and training

Nucleotide sequences and amino acid feature matrices were passed into two separate convolutional blocks, each with the same model architecture, as previously described[7]: a shallow network with two sets of two convolutional layers and a max pooling layer. The outputs of the two convolutional blocks were flattened, concatenated together with lineage SNPs, and passed through two densely connected layers of 256 nodes each to obtain a final MIC prediction (Fig. 1c). All layers except the output layer have ReLU activation.

Owing to the labor-intensive nature of measuring microbial MICs, specific concentrations of each drug are tested by different laboratories (typically 2-fold dilutions), making MIC a semi-continuous variable. If an isolate grows at a concentration of N, but not at 2 N, the MIC is recorded as the range (N, 2 N]. Because the true MIC within this interval is unknown, we built a custom loss function for the CNN (Fig. 1b, and Supplementary Fig. 2). The binned loss function computes mean absolute error (MAE) only for MIC predictions that lie outside the concentration interval associated with an isolate. For predictions within the bounds, the error = 0. For predictions outside of these bounds, the loss is computed relative to the closer bound (i.e., the absolute error for a prediction of 0.7 for an isolate with a true MIC in the range [0.2, 0.4] would be 0.3). Our dataset is composed of MICs measured from different labs, which right-censor MICs differently: two isolates with the same true MIC of 100 μg/mL may be recorded as >2 μg/mL and >50 μg/mL because one laboratory did not test higher concentrations. The binned loss function accounts for these differences because it penalizes the predictions only if they are below the lower bounds of 2 and 50, respectively. We selected to minimize the binned mean absolute error because it is more robust to outliers than mean squared error or root mean squared error, especially given that the MIC datasets are considerably smaller than typical machine learning datasets. Additionally, squared error loss functions tend to weight larger mispredictions more than smaller mispredictions, and we wanted to minimize even small mispredictions of small MIC values.

The simplest and most interpretable model for predicting an unbounded continuous value is linear regression. Although regression

models have achieved considerable success identifying genetic mutations that are associated with antibiotic resistance in the MTBC[4,30,36,73–75], regression considers all features to be independent of one another and cannot learn from the locations of and distances between features. Further, the target must be a linear combination of inputs, which is not always the case for complex phenotypes driven by interactions between alleles (*i.e.*, epistasis). Regression models were fit using scikit-learn's (v1.4.2) Ridge regression function. We selected L2 regularization to correct for collinearity between nucleotides due to linkage disequilibrium in bacteria. The strength of the regularization parameter was selected among all powers of 10 from $10^{-5}$ to $10^{5}$ to minimize the validation loss. Because the scikit-learn implementation does not allow passing in multiple targets for each sample, it was trained to minimize mean absolute error using the midpoints of each MIC range.

CNNs were trained using tensorflow (v2.7.0) and Python (v3.9.16). We used 5-fold cross-validation, stratified by binary resistance phenotype, to compare model performance metrics on the hold-out test set. The cross-validated CNNs were trained using early stopping, until the test loss did not decrease by at least 1% for 200 (*i.e.*, the patience period is 200 epochs long). Significance testing was performed using Welch's t-test between the 5 splits of the two given models. We compared binned mean absolute error (referred to simply as "mean absolute error") between models. For sensitivity and specificity, 95% confidence interval bounds were computed on the cross-validation results as $\bar{X} \pm t\frac{s}{\sqrt{N}}$, where N = 5, the number of cross-validation folds; $\bar{X}$ and s are the mean and standard deviation, respectively, of the five folds; and t is the score from the Student's t-distribution with $\alpha = 0.05$ and degrees of freedom = N - 1 = 4. After cross-validation, a final model was trained on the full dataset for each drug using these same training parameters for downstream analyses.

For the binary pyrazinamide models, all training parameters, including the number of training epochs and stopping criterion, were the same as for the quantitative models, but the models were trained using binary cross-entropy as the loss. We selected the threshold for binarizing the predicted scores to maximize the sum of sensitivity and specificity.

### Saliency mapping

We used the DeepLIFT attribution algorithm[76] as implemented in DeepExplain[77] to measure how strongly each categorical feature (nucleotide identity and lineage SNP presence vs. absence) affects model predictions. By comparing each sample to a reference sequence (H37Rv), the algorithm assigns a "saliency score" to each feature that differs from the reference in each training isolate. The score for a single nucleotide position is the sum of scores across the four values associated with the four non-reference alleles at that site, and we computed an aggregate maximum (positive) and minimum (negative) score for each position. Lineage saliency scores are similarly computed with the reference lineage type being the reference strain H37Rv (*i.e.*, the vector of lineage SNPs composed entirely of 0 s). We performed a permutation test to assess the significance of the scores. The phenotype labels for each dataset were randomly shuffled, and a new model was trained using the same previously selected regularization parameter until the loss did not decrease by 1% for 100 epochs. This process was repeated 10 times. For positions with a non-zero maximum or minimum saliency score, *p*-values were computed using the permuted scores. For sites with a 0 score, there is no meaningful *p*-value because the permuted scores are also zero.

We computed correlations between the saliency scores and the average log2-MIC observed in the training dataset for different lineage SNPs. We included only sublineages with at least 2 isolates and with significant saliency scores according to the above permutation test.

### In silico mutagenesis

To isolate the effects of individual genetic variants on predicted MIC, we performed in silico mutagenesis experiments. Individual mutations or lineage SNPs were inserted into the H37Rv reference to observe how the trained CNNs predict MIC on a sequence with a single genetic change. Because the CNN can only make new predictions for input matrices of the same dimensions (longest locus length x one-hot encodings x number of loci) as the data it was trained on, insertions at positions where there were no insertions in the training dataset could not be tested because they increase the size of the multiple sequence alignment. Due to some naming inconsistencies in the WHO mutation catalog, we derived the nucleotide change using the name of a mutation according to the Sequence Ontology nomenclature used in the catalog and verified the accuracy by annotating with snpEff. As in the catalog, we considered loss of function (LoF) mutations for coding regions to be loss of start codons, early stop codons, frameshift mutations, and gene deletions / ablations. For amino acids encoded by multiple codons, the codon with the smallest edit distance from the reference codon was selected to encode it in the in silico sequence. Frameshift variants were not included in this analysis because different nucleotide changes of drastically different sizes could all result in the same frameshift mutation in the WHO catalog naming system.

We inserted the 62 Coll lineage SNPs and all compatible mutations from the 2nd edition of the WHO resistance mutation catalog into H37Rv. We additionally performed site-saturation mutagenesis for some proteins by mutating each residue to the other 19 possible amino acids and the stop codon. For the first amino acid, all changes except for the six alternative start codons in bacteria were considered to be start lost mutations. Because all changes to the first codon are either loss of start mutations or initiator codon mutations that are translated to methionine, the first codon was excluded from the clustering computation because it does not have missense mutations. The in silico mutagenesis procedures yield a predicted MIC for H37Rv. We used this predicted value to compute fold changes for mutations.

### 3D space clustering computation

The Getis-Ord statistic is used to measure spatial association of a user-defined metric and identify hot and coldspots[42]. For the $i^{th}$ object in a group, the Getis-Ord ($G_i^*$) statistic is the z-score of the metric, weighted by the distance between the object and all other objects. It therefore estimates whether the object has a very high (positive) or low (negative) z-score and is also surrounded by objects with extreme z-scores. We computed $G_i^*$ statistics using the average log fold change of predicted MIC ($\log_2$FC) across the 19 possible missense mutations at each residue of a protein, excluding start lost mutations.

A permutation test was then performed with 10,000 replicates to determine in what fraction of permuted structures the $i^{th}$ residue has a $G_i^*$ statistic at least as extreme as the test score. FDR correction was performed using the Benjamini-Hochberg method, and we selected an FDR q-value of 0.05 as the threshold for significant clustering. We considered the $i^{th}$ residue to be a resistance-associated "hotspot" if $G_i^* > 0$ and $\log_2FC_i > 0$ and a susceptibility-associated "coldspot" if $G_i^* < 0$ and $\log_2FC_i < 0$. A residue with a significant Getis-Ord statistic (Benjamini-Hochberg-corrected $q \leq 0.05$) implies it is surrounded by residues with predicted MICs significantly higher or significantly lower than the average for the entire protein. A residue with a high $\log_2$FC may not necessarily have a significant $G_i^*$ statistic if it is not also surrounded by residues with high predicted MICs, and vice versa for a low $\log_2$FC.

The proteins for which we performed 3D clustering are as follows: rpoB (UniProt: P9WGY9, PDB: 5UHB, chain C), gyrA and gyrB (UniProt: P9WG45, P9WG47; PDB: 5BS8), pncA (UniProt: I6XD65, PDB: 3PL1), katG (UniProt: P9WIE5, PDB: 4C51), ethA (UniProt: P9WNF9), and Rv0678 (UniProt: I6Y8F7, PDB: 4NB5).

### Patient outcomes models

To test if MICs are associated with unfavorable TB treatment outcome, we built multivariate Cox proportional hazards models to associate clinical and microbiological data from participants in the TRUST study[54] with two events–sputum culture conversion and unfavorable outcomes. Participants in this study had rifampicin-susceptible pulmonary TB diagnosed by GeneXpert and were all treated with a 6-month regimen of RIPE (rifampicin, isoniazid, pyrazinamide, and ethambutol), followed by 12 months of follow-up. Time zero for each participant is the time of initial screening and treatment initiation. A total of 440 individuals were enrolled at the time this analysis was conducted. 60% are male, 28% have HIV, 38% had a previous TB disease episode, 4% have diabetes, and 54% have a body mass index (BMI) < 18. A total of 373 participants remained after excluding participants who changed treatment regimens due to developing rifampicin resistance or needing a liver-friendly regimen, participants without WGS of their bacterial cultures performed within the first two weeks of enrollment, and participants with incomplete data (Supplementary Fig. 7a). One participant did not have MICs measured, whom we excluded only from the models with measured MICs as covariates. All participants had MTBC infections belonging to lineages 1-4, including mixed infections.

Unfavorable treatment outcomes were defined as treatment failure, TB disease relapse, or death. Treatment failure was assessed after 6 months of treatment, either by a positive culture or symptomatically. Relapse was determined if a participant was sputum- or smear-positive at the conclusion of the 12-month follow-up period. The median duration of enrollment of the 373 patients was 545 days, and the longest was 835 days. We censored 5 of the 8 patients whose deaths were attributed to non-TB causes, specifically cancer (N = 3), COVID-19 (N = 1), and accidental deaths (N = 1), leaving 3 TB-associated deaths. All other patients were right-censored at their study completion date or their last known clinic appointment if they were lost to follow-up (N = 35) or were still enrolled in the study (N = 12) at the time of analysis.

Sputum was sampled weekly for the first 12 weeks of treatment. Time to negative sputum culture conversion (TCC) was defined as the number of weeks between the time of treatment initiation and the first of two consecutive *Mtb*-negative sputum cultures not followed by another positive culture. To be included in the TCC analysis, participants had to have at least 3 uncontaminated sputum samples and a positive *Mtb* culture within the first 5 weeks of initial screening, leaving 353/373 patients remaining for the TCC analysis. We imputed missing cultures after this using the mice package (v3.16.0) in R (v4.4.2)[78]. To impute each culture, we used age, sex, BMI, HIV/CD4 status, smoked substance use (methamphetamine, methaqualone, or cannabis), alcohol use, isoniazid resistance, time to culture positivity at baseline, and all smear positivity and culture results taken before and at the same week as the culture to be imputed, similar to as done previously[78]. TCC was computed on each of 30 imputed datasets, and model results were pooled across 30 imputations using Rubin's rules[79].

We selected 11 relevant clinical covariates using domain knowledge and past evidence for association with poor TB treatment outcomes (Table 4)[80–82]. Percent lung involvement and drug exposure measured by the area under the pharmacokinetic ROC curve[55] were predicted from chest x-ray images and patient metadata, respectively, while the other independent variables were determined from patient health records. Based on past work with this and other data[56], we binarized percent lung involvement at 20% to generate a variable "high lung involvement." Cavitation was excluded due to a high rate of missingness. We additionally added a binary variable for *Mtb* lineage (L2 vs. all other lineages due to the low prevalence of L1 and L3 isolates) and included the F2 strain mixing metric at baseline[72].

To satisfy the Cox proportional hazards assumption, we stratified by smear positivity, high lung involvement, and HIV/CD4 status in the TCC model only (Fig. 5d-f). The proportional hazards assumption ($p > 0.01$, Schoenfeld residuals test) was met for all covariates in the unfavorable outcomes model. To these models, we added measured or CNN-predicted MICs ($\log_2$-transformed) for the four first-line drugs to determine if the MICs were significantly associated with outcomes, conditioned on the presence of other relevant patient characteristics. We also controlled for predicted drug exposure in each model with MICs[55].

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The Illumina sequencing data used in this study have been deposited in SRA under various accession codes (Supplementary Data 1). Supplementary Data 1 includes accession codes for data from the CRyPTIC[25], PATRIC[26], and World Health Organization Resistance Mutation Catalog datasets[27]. The MIC data from the MIC-ML consortium are available under restricted access because they are not yet public, access can be obtained by contacting the original data collectors (Supplementary Data 2). The raw clinical cohort data are protected and are not available due to data privacy laws. Partial processed clinical cohort data are available at (https://github.com/sanju99/MtbQuantCNN/tree/main/analysis/TRUST/processed_data).

## Code availability

All code is available at (https://github.com/sanju99/MtbQuantCNN) and can be cited from (https://doi.org/10.5281/zenodo.19269138).

## References

1. Su, M., Satola, S. W. & Read, T. D. Genome-based prediction of bacterial antibiotic resistance. *J. Clin. Microbiol.* **57**, e01405-18 (2019).
2. Banerjee, R. & Patel, R. Molecular diagnostics for genotypic detection of antibiotic resistance: current landscape and future directions. *JAC Antimicrob. Resist.* **5**, dlad018 (2023).
3. Dookie, N., Rambaran, S., Padayatchi, N., Mahomed, S. & Naidoo, K. Evolution of drug resistance in Mycobacterium tuberculosis: a review on the molecular determinants of resistance and implications for personalized care. *J. Antimicrob. Chemother.* **73**, 1138–1151 (2018).
4. Farhat, M. R., Freschi, L., Calderon, R. & Murray, M. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat. Comm.* **10**, 2128 (2019).
5. Chakravorty, S. et al. The new Xpert MTB/RIF Ultra: Improving detection of Mycobacterium tuberculosis and resistance to rifampin in an assay suitable for point-of-care testing. *MBio* **8**, e00812-17 (2017).
6. Dippenaar, A. et al. Diagnostic accuracy of the FluoroType MTB and MTBDR VER 2.0 assays for the centralized high-throughput detection of Mycobacterium tuberculosis complex DNA and isoniazid and rifampicin resistance. *Clin. Microbiol. Infect.* **27**, 1351.e1–1351.e4 (2021).
7. Green, A. G., Yoon, C. H., Chen, M. L., Farhat, M. R. A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis. *Nat. Commun.* **13**, 3817 (2022).
8. *Technical Report on Critical Concentrations for Drug Susceptibility Testing of Medicines Used in the Treatment of Drug-Resistant Tuberculosis*. https://www.who.int/publications/i/item/WHO-CDS-TB-2018.5 (2018).
9. Ängeby, K., Juréen, P., Kahlmeter, G., Hoffner, S. E. & Schön, T. Challenging a dogma: antimicrobial susceptibility testing

breakpoints for Mycobacterium tuberculosis. *Bull. World Health Organ.* **90**, 693–698 (2012).
10. WHO announces updated critical concentrations for susceptibility testing to rifampicin. in *World Health Organization (WHO)* (2021).
11. Colman, R. E. et al. Detection of low-level mixed-population drug resistance in Mycobacterium tuberculosis using high fidelity amplicon sequencing. *PLoS One* **10**, e0126626 (2015).
12. Safi, H. et al. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-arabinose biosynthetic and utilization pathway genes. *Nat. Genet.* **45**, 1190–1197 (2013).
13. Sirgel, F. A., Wiid, I. J. F. & van Helden, P. D. Measuring minimum inhibitory concentrations in mycobacteria. *Methods Mol. Biol.* **465**, 173–186 (2009).
14. Colangeli, R. et al. Bacterial factors that predict relapse after tuberculosis therapy. *N. Engl. J. Med.* **379**, 823–833 (2018).
15. Kowalska-Krochmal, B. & Dudek-Wicher, R. The minimum inhibitory concentration of antibiotics: Methods, interpretation, clinical relevance. *Pathogens* **10**, 165 (2021).
16. Gröschel, M. I. et al. GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Med* **13**, 138 (2021).
17. Kouchaki, S. et al. Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics* **35**, 2276–2282 (2019).
18. Kuang, X., Wang, F., Hernandez, K. M., Zhang, Z. & Grossman, R. L. Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN. *Sci. Rep.* **12**, 2427 (2022).
19. Sharma, A., Machado, E., Lima, K. V. B., Suffys, P. N. & Conceição, E. C. Tuberculosis drug resistance profiling based on machine learning: A literature review. *Braz. J. Infect. Dis.* **26**, 102332 (2022).
20. Suster, C. J. E., Pham, D., Kok, J. & Sintchenko, V. Emerging applications of artificial intelligence in pathogen genomics. *Front. Bacteriol.* **3**, 1326958 (2024).
21. CRyPTIC consortium Quantitative drug susceptibility testing for Mycobacterium tuberculosis using unassembled sequencing data and machine learning. *PLoS Comput. Biol.* **20**, e1012260 (2024).
22. Alharbi, W. S. & Rashid, M. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum. Genom.* **16**, 26 (2022).
23. Al-Ajlan, A. & El Allali, A. CNN-MGP: Convolutional neural networks for metagenomics gene prediction. *Interdiscip. Sci.* **11**, 628–635 (2019).
24. Hu, W., Li, M., Xiao, H. & Guan, L. Essential genes identification model based on sequence feature map and graph convolutional neural network. *BMC Genom.* **25**, 47 (2024).
25. The CRyPTIC Consortium A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics. *PLoS Biol.* **20**, e3001721 (2022).
26. Gillespie, J. J. et al. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.* **79**, 4286–4298 (2011).
27. *Catalogue of Mutations in Mycobacterium Tuberculosis Complex and Their Association with Drug Resistance, Second Edition*, (2023).
28. Thadani, N. N. et al. Learning from prepandemic data to forecast viral escape. *Nature* **622**, 818–825 (2023).
29. Goig, G. A. et al. Ecology, global diversity and evolutionary mechanisms in the Mycobacterium tuberculosis complex. *Nat. Rev. Microbiol.* **23**, 602–614 (2025).
30. Oppong, Y. E. A. et al. Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. *BMC Genom.* **20**, 252 (2019).
31. Earle, S. G. et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1**, 16041 (2016).
32. Coll, F. et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5**, 4812 (2014).
33. Humphries, R. M. et al. CLSI Methods Development and Standardization Working Group best practices for evaluation of antimicrobial susceptibility tests. *J. Clin. Microbiol.* **56**, (2018).
34. Kelley, C. L., Rouse, D. A. & Morris, S. L. Analysis of ahpC gene mutations in isoniazid-resistant clinical isolates of Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **41**, 2057–2058 (1997).
35. Sreevatsan, S., Pan, X., Zhang, Y., Deretic, V. & Musser, J. M. Analysis of the oxyR-ahpC region in isoniazid-resistant and -susceptible Mycobacterium tuberculosis complex organisms recovered from diseased humans and animals in diverse localities. *Antimicrob. Agents Chemother.* **41**, 600–606 (1997).
36. Consortium, C. R.yP. T. I. C. Quantitative measurement of antibiotic resistance in Mycobacterium tuberculosis reveals genetic determinants of resistance and susceptibility in a target gene approach. *Nat. Commun.* **15**, 488 (2024).
37. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance, 2nd ed. https://www.who.int/publications/i/item/9789240082410 (2023).
38. Köser, C. U., Georghiou, S. B., Schön, T. & Salfinger, M. On the consequences of poorly defined breakpoints for rifampin susceptibility testing of Mycobacterium tuberculosis complex. *J. Clin. Microbiol.* **59**, e02328-20 (2021).
39. Lale Ngema, S. et al. Isoniazid resistance-conferring mutations are associated with highly variable phenotypic resistance. *J. Clin. Tuberc. Other Mycobact. Dis.* **33**, 100387 (2023).
40. *World Health Organization. Technical Report on Critical Concentrations for Drug Susceptibility Testing of Isoniazid and the Rifamycins (rifampicin, Rifabutin and Rifapentine). Geneva: The Organization*. (2021).
41. García-Marín, A. M. et al. Role of the first WHO mutation catalogue in the diagnosis of antibiotic resistance in Mycobacterium tuberculosis in the Valencia Region, Spain: a retrospective genomic analysis. *Lancet Microbe* **5**, e43–e51 (2024).
42. Getis, A. & Ord, J. K. The analysis of spatial association by use of distance statistics. in *Advances in Spatial Science* 127–145 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010).
43. Maruri, F. et al. A systematic review of gyrase mutations associated with fluoroquinolone-resistant Mycobacterium tuberculosis and a proposed gyrase numbering system. *J. Antimicrob. Chemother.* **67**, 819–831 (2012).
44. Briffotaux, J., Huang, W., Wang, X. & Gicquel, B. MmpS5/MmpL5 as an efflux pump in Mycobacterium species. *Tuberculosis (Edinb.)* **107**, 13–19 (2017).
45. Radhakrishnan, A. et al. Crystal structure of the transcriptional regulator Rv0678 of Mycobacterium tuberculosis. *J. Biol. Chem.* **289**, 16526–16540 (2014).
46. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023).
47. Casali, N. et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat. Genet.* **46**, 279–286 (2014).
48. Hicks, N. D., Carey, A. F., Yang, J., Zhao, Y. & Fortune, S. M. Bacterial genome-wide association identifies novel factors that contribute to ethionamide and prothionamide susceptibility in Mycobacterium tuberculosis. *MBio* **10**, e00616-19 (2019).
49. Grant, S. S. et al. Baeyer-Villiger monooxygenases EthA and MymA are required for activation of replicating and non-replicating Mycobacterium tuberculosis inhibitors. *Cell Chem. Biol.* **23**, 666–677 (2016).

50. Vilchèze, C. et al. Mycothiol biosynthesis is essential for ethionamide susceptibility in Mycobacterium tuberculosis. *Mol. Microbiol.* **69**, 1316–1329 (2008).
51. Xu, X. et al. Precise null deletion mutations of the mycothiol synthesis genes reveal their role in isoniazid and ethionamide resistance in Mycobacterium smegmatis. *Antimicrob. Agents Chemother.* **55**, 3133–3139 (2011).
52. Ang, M. L. T. et al. EthA/R-independent killing of Mycobacterium tuberculosis by ethionamide. *Front. Microbiol.* **8**, 710 (2017).
53. Vilchèze, C. & Jacobs, W. R., Jr. Resistance to isoniazid and ethionamide in Mycobacterium tuberculosis: genes, mutations, and causalities. *Microbiol. Spectr.* **2**, MGM2–0014–2013 (2014).
54. Myers, B. et al. Impact of alcohol consumption on tuberculosis treatment outcomes: a prospective longitudinal cohort study protocol. *BMC Infect. Dis.* **18**, 488 (2018).
55. Wijk, M. et al. The impact of alcohol and illicit substance use on the pharmacokinetics of first-line TB drugs. *J. Antimicrob. Chemother.* **79**, 2022–2030 (2024).
56. Ghanem, M. et al. Tuberculosis Disease Severity Assessment Using Clinical Variables and Radiology Enabled by Artificial Intelligence. *J. Infect. Dis.*https://doi.org/10.1093/infdis/jiag149 (2026).
57. Lan, N. T. N. et al. Mycobacterium tuberculosis Beijing genotype and risk for treatment failure and relapse, Vietnam. *Emerg. Infect. Dis.* **9**, 1633–1635 (2003).
58. Parwati, I. et al. Mycobacterium tuberculosis Beijing genotype is an independent risk factor for tuberculosis treatment failure in Indonesia. *J. Infect. Dis.* **201**, 553–557 (2010).
59. Sun, Y.-J., Lee, A. S. G., Wong, S.-Y. & Paton, N. I. Association of Mycobacterium tuberculosis Beijing genotype with tuberculosis relapse in Singapore. *Epidemiol. Infect.* **134**, 329–332 (2006).
60. Huyen, M. N. T. et al. Tuberculosis relapse in Vietnam is significantly associated with Mycobacterium tuberculosis Beijing genotype infections. *J. Infect. Dis.* **207**, 1516–1524 (2013).
61. Kurbatova, E. V. et al. Sputum culture conversion as a prognostic marker for end-of-treatment outcome in patients with multidrug-resistant tuberculosis: a secondary analysis of data from two observational cohort studies. *Lancet Respir. Med.* **3**, 201–209 (2015).
62. Alene, K. A. et al. Comparison of the validity of smear and culture conversion as a prognostic marker of treatment outcome in patients with multidrug-resistant tuberculosis. *PLoS One* **13**, e0197880 (2018).
63. Jindani, A., Doré, C. J. & Mitchison, D. A. Bactericidal and sterilizing activities of antituberculosis drugs during the first 14 days. *Am. J. Respir. Crit. Care Med* **167**, 1348–1354 (2003).
64. Tornheim, J. A. & Dooley, K. E. Tuberculosis associated with HIV infection. *Microbiol. Spectr.* **5**, 1128 (2017).
65. Bruchfeld, J., Correia-Neves, M. & Källenius, G. Tuberculosis and HIV coinfection. *Cold Spring Harb. Perspect. Med.* **5**, a017871 (2015).
66. Brankin, A. E. & Fowler, P. W. Predicting antibiotic resistance in complex protein targets using alchemical free energy methods. *J. Comput. Chem.* **43**, 1771–1782 (2022).
67. GenoScreen (2022). *Deeplex® Myc-TB: From Clinical Samples to Drug Resistance Profile* https://www.genoscreen.fr/images/genoscreen-services/deeplex/technical_note_20221221_RUO-avec_compression_1.pdf.
68. Chang, V. K. et al. Risk-stratified treatment for drug-susceptible pulmonary tuberculosis. *Nat. Commun.* **15**, 9400 (2024).
69. Zhou, M. et al. The efficacy and safety of high-dose isoniazid-containing therapy for multidrug-resistant tuberculosis: a systematic review and meta-analysis. *Front. Pharmacol.* **14**, 1331371 (2023).
70. WHO launches new guidance on the use of targeted next-generation sequencing tests for the diagnosis of drug-resistant TB and a new sequencing portal. https://www.who.int/news/item/20-03-2024-who-launches-new-guidance-on-the-use-of-targeted-next-generation-sequencing-tests-for-the-diagnosis-of-drug-resistant-tb-and-a-new-sequencing-portal. (2024).
71. Marin, M. G. et al. Pitfalls of bacterial pan-genome analysis approaches: a case study of Mycobacterium tuberculosis and two less clonal bacterial species. *Bioinformatics* **41**, btaf219 (2025).
72. Wyllie, D. H. et al. Identifying mixed Mycobacterium tuberculosis infection and laboratory cross-contamination during mycobacterial sequencing programs. *J. Clin. Microbiol.* **56**, (2018).
73. Naz, S. et al. GWAS and functional studies suggest a role for altered DNA repair in the evolution of drug resistance in Mycobacterium tuberculosis. *Elife* **12**, e75860 (2023).
74. Lagutkin, D. et al. Genome-Wide Study of Drug Resistant Mycobacterium tuberculosis and Its Intra-Host Evolution during Treatment. *Microorganisms* **10**, 1440 (2022).
75. The CRyPTIC Consortium The CRyPTIC Consortium. Genome-wide association studies of global Mycobacterium tuberculosis resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. *PLoS Biol.* **20**, e3001755 (2022).
76. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. *arXiv [cs.CV]* (2017).
77. Ancona, M., Ceolini, E., Öztireli, C. & Gross, M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1711.06104 (2017).
78. Malatesta, S. et al. Methods for handling missing data in serially sampled sputum specimens for mycobacterial culture conversion calculation. *BMC Med. Res. Methodol.* **22**, 297 (2022).
79. Rubin, D. B. et al. *Multiple Imputation for Nonresponse in Surveys*. (John Wiley & Sons, Nashville, TN, 2009).
80. Prakash Babu, S. et al. Factors associated with unfavorable treatment outcomes among persons with pulmonary tuberculosis: A multicentric prospective cohort study from India. *Clin. Infect. Dis.* **79**, 1034–1038 (2024).
81. Limenh, L. W. et al. Tuberculosis treatment outcomes and associated factors among tuberculosis patients treated at healthcare facilities of Motta Town, Northwest Ethiopia: a five-year retrospective study. *Sci. Rep.* **14**, 7695 (2024).
82. Chaves Torres, N. M., Quijano Rodríguez, J. J., Porras Andrade, P. S., Arriaga, M. B. & Netto, E. M. Factors predictive of the success of tuberculosis treatment: a systematic review with meta-analysis. *PLoS One* **14**, e0226507 (2019).
83. Shell, S. S. et al. Leaderless transcripts and small proteins are common features of the mycobacterial translational landscape. *PLoS Genet* **11**, e1005641 (2015).
84. Kapopoulou, A., Lew, J. M. & Cole, S. T. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb.)* **91**, 8–13 (2011).
85. Gagneux, S. Ecology and evolution of Mycobacterium tuberculosis. *Nat. Rev. Microbiol.* **16**, 202–213 (2018).
86. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nat. Med.* **2**, 662–667 (1996).


## Acknowledgements

We thank Marie Wijk and Paolo Denti for providing pharmacokinetic predictions for the patient outcomes analysis. We thank the study participants and study staff for making this work possible. Computational resources and support were provided by the Orchestra High Performance Compute Cluster at Harvard Medical School, which is funded by the NIH (NCRR 1S10RR028832-01). S.G.K. was supported by a National Institutes of Health Training Grant GM132089 and a National Science Foundation Graduate Research Fellowship DGE2140743. A.G.G. was

supported by a National Institutes of Health NLM Training Grant T15LM007092 and NIH/NIAIDF32AI161793. M.R.F. was supported by a National Institute of Allergy and Infectious Diseases / National Institutes of Health grant R01AI155765. RMW was supported by the South African Medical Research Council.

## Author contributions

S.G.K. implemented all code and ran all analyses, except for the Getis-Ord statistic computation, which was implemented by A.G.G. S.G.K. and M.R.F. interpreted the data and results and wrote the manuscript. B.C.M. and N.R. prepared whole-genome sequencing libraries and performed MIC testing. S.K.G., S. Malatesta, and N.C. cleaned patient data and shared results for the survival analyses. S. Mulaudzi cleaned and shared phenotypic data from the MIC-ML consortium. All authors provided feedback on the manuscript. M.R.F., K.R.J., and R.M.W. supervised the research.

## Competing interests

All authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-72225-x.

**Correspondence** and requests for materials should be addressed to Maha R. Farhat.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.



## MIC-ML Consortium

**Keertan Dheda[6,7,8], Shaheed Vally Omar[9], Hafid Soualhine[10], Nazir Ismail[11], Barun Mathema[12], Neel Gandhi[13], Helen McIlleron[14], Satoshi Mitarai[15], Kyungjong Kim[16], Helen Cox[17,18], Stuart Meier[6], Robin M. Warren[3], Elizabeth Streicher[3], Sacha Laurent[19], Timothy Rodwell[19], Carl-Michael Nathanson[20], Barry Kreiswirth[21], Elise DeVos[22], Daniela Cirillo[23], Vincent E. Escuyer[24], Elena Martinez[25,26,27], Vitali Sintchenko[25,26,27], Anzaan Dippenaar[28], Tim Heupink[28], Annelies Van Rie[28] & James C. M. Brust[29]**

[6]Centre for Lung Infection and Immunity, Division of Pulmonology, Department of Medicine and UCT Lung Institute, University of Cape Town, Cape Town, South Africa. [7]South African MRC/UCT Centre for the Study of Antimicrobial Resistance, University of Cape Town, Cape Town, South Africa. [8]Faculty of Infectious and Tropical Diseases, Department of Infection Biology, London School of Hygiene & Tropical Medicine, London, United Kingdom. [9]Centre for Tuberculosis, National and WHO Supranational TB Reference Laboratory, National Institute for Communicable Diseases a division of the National Health Laboratory Service, Johannesburg, South Africa. [10]National Microbiology laboratory, Public Health Agency of Canada, Winnipeg, Canada. [11]Department of Clinical Microbiology and Infectious Diseases, University of the Witwatersrand, Johannesburg, South Africa. [12]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, USA. [13]Department of Medicine, Emory University School of Medicine, Departments of Epidemiology and Global Health, Emory University Rollins School of Public Health, Emory University Rollins School of Public Health, Atlanta, USA. [14]Department of Medicine, University of Cape Town, Cape Town, South Africa. [15]Department of Mycobacterium Reference and Research, Research Institute of Tuberculosis, Japan Anti-Tuberculosis Association, Tokyo, Japan. [16]Department of R&D, The Korean Institute of Tuberculosis, Cheongju, South Korea. [17]Division of Medical Microbiology, Department of Pathology, University of Cape Town, Cape Town, South Africa. [18]Institute of Infectious Disease and Molecular Medicine and Wellcome Centre for Infectious Disease Research in Africa, University of Cape Town, Cape Town, South Africa. [19]Foundation for Innovative New Diagnostics (FIND), Geneva, Switzerland. [20]Global Tuberculosis Programme, World Health Organization (WHO), Geneva, Switzerland. [21]Hackensack Meridian Health Center for Discovery and Innovation, Nutley, USA. [22]Institute of Tropical Medicine, Antwerp, Belgium. [23]IRCCS San Raffaele Scientific Institute, Milano, Italy. [24]New York State Department of Health, Wadsworth Center, Albany, USA. [25]Sydney Infectious Diseases Institute, The University of Sydney, Sydney, Australia. [26]Centre for Infectious Diseases and Microbiology-Public Health, Westmead Hospital, Western Sydney Local Health District, Sydney, Australia. [27]NSW Mycobacterium Reference Laboratory, Centre for Infectious Diseases and Microbiology-Laboratory Services, Institute of Clinical Pathology and Medical Research, NSW Health Pathology, Sydney, Australia. [28]Tuberculosis Omics Research Consortium, Department of Family Medicine and Population Health, Global Health Institute, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium. [29]Division of General Internal Medicine and Division of Infectious Diseases, Montefiore Medical Center, Albert Einstein College of Medicine, New York, USA.