Title: A convolutional neural network highlights mutations relevant to antimicrobial resistance in *Mycobacterium tuberculosis*

Authors: Anna G. Green[1*] and Chang H. Yoon[1*], Michael L. Chen[1,4], Luca Freschi[1], Matthias I. Gröschel[1], Isaac Kohane[1], Andrew Beam[1,2*], Maha Farhat[1,3*]

Affiliations: [1]Department of Biomedical Informatics, Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA

[2] Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA

[3] Division of Pulmonary & Critical Care, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114, USA

[4]Stanford University School of Medicine, 291 Campus Dr, Stanford, CA 94305, USA

[*]Equal contribution. Other authors listed alphabetically.

Corresponding authors: Prof. Maha Farhat, Maha_Farhat@hms.harvard.edu, Prof. Andrew Beam, andrew_Beam@hms.harvard.edu

Abstract

Long diagnostic wait times hinder international efforts to address multi-drug resistance in *M. tuberculosis*. Pathogen whole genome sequencing, coupled with statistical and machine learning models, offers a promising solution. However, generalizability and clinical adoption have been limited in part by a lack of interpretability and verifiability, especially in deep learning methods. Here, we present a deep convolutional neural network (CNN) that predicts the antibiotic resistance phenotypes of *M. tuberculosis* isolates. The CNN performs with state-of-the-art levels of predictive accuracy. Evaluation of salient sequence features permits biologically meaningful interpretation and validation of the CNN's predictions, with promising repercussions for functional variant discovery, clinical applicability, and translation to phenotype prediction in other organisms.

1. Introduction

Tuberculosis is a leading cause of death worldwide from an infectious pathogen, with more than 1.5 million people succumbing to the disease annually(*1*). Rising rates of antibiotic-resistant *Mycobacterium tuberculosis*, the causative agent of tuberculosis, continue to rise, pose a threat to public health(*2*). A major challenge in combatting antibiotic-resistant tuberculosis is the timely selection of appropriate treatments for each patient, particularly when growth-based drug susceptibility testing takes weeks(*1*).

Molecular diagnostic tests for *M. tuberculosis* antimicrobial resistance reduce the time to result to hours or days, but only target a small number of loci relevant to a few antibiotics, and cannot detect most rare genetic variants(*3*). Although whole genome sequencing-related diagnostic tests offer the promise of resolving some of these deficiencies, statistical association techniques have seen limited success, hindered by their inability to assess newly observed variants and epistatic effects(*3–7*). More complex models such as deep learning provide promising flexibility but are often uninterpretable, making them difficult to audit for safety purposes (*8*, *9*). Moreover, interrogating black box models offers the opportunity for hypothesis generation which can be later validated, potentially improving scientific understanding of the underlying phenomenon (*10*).

A recent "wide-and-deep" neural network applied to *M. tuberculosis* genomic data outperformed previous methods to predict antimicrobial resistance to 10 antibiotics(*11*); however, like most deep learning methods, the logic behind its predictions was indiscernible. Although more interpretable rule-based classifiers of antimicrobial resistance in *M. tuberculosis* have been developed(*12*, *13*), these rely on predetermined single-nucleotide polymorphisms or *k*-mers, hindering their flexibility to generalize to newly observed mutations, and universally ignore genomic context. Deep convolutional neural networks (CNNs), which greatly reduce the number of required parameters compared to traditional neural networks, could be used to consider multiple complete genomic loci with the ultimate goal of incorporating the whole genome. This would allow the model to assess mutations in their genetic context by capturing the order and distance between resistance mutations of the same locus, allowing a better incorporation of rare or newly observed variation. Deep CNNs, when paired with attribution methods that highlight the most salient features informing the model predictions, are a promising means of harnessing the predictive power of deep neural networks in genomics for biological discovery and interpretation(*14*). CNNs also have the added advantage of minimizing the preprocessing needed of genomic variant data. The extent to which we may trust these highlighted features remains the subject of ongoing scientific exploration(*8*, *15*, *16*).

Here, we show that CNNs perform *en par* with the state-of-the-art in predicting antimicrobial resistance in *M. tuberculosis* and provide biological interpretability through motif representation captured in saliency mapping. We train two models: one designed for accuracy that incorporates genetic and phenotypic information about all drugs; and a second designed for interpretability that forces the model to only consider putatively causal regions for a particular drug. Our models are trained on the entire genetic sequence of 18 regions of the genome known or predicted to influence antibiotic resistance, using data collected from over 20,000 *M. tuberculosis* strains spanning the four major global lineages. Across each locus,

we calculate genomic positions that most influence the prediction of resistance for each drug, validating our method by recapitulating known positions and providing predictions of new positions potentially involved in drug resistance. Given the growing movement towards greater interpretability in machine learning methods(*16*, *17*), we expect this model to have implications for hypothesis generation about molecular mechanisms of antimicrobial resistance through genotype-phenotype association.

2. Results

Training dataset characteristics

We train and cross-validate our models using 10,201 *M. tuberculosis* isolates from the ReSeqTB and the WHO Supranational Reference Laboratory Network **(**sources detailed in the **Materials and Methods)**. Each isolate is phenotyped for resistance to at least one of thirteen antitubercular drugs: the four first-line drugs isoniazid, rifampicin, ethambutol, and pyrazinamide, and nine additional second-line drugs (**Table 1**). All drugs are represented by at least 250 phenotyped isolates.

Model design

We build two models to predict antibiotic resistance phenotypes from genome sequences. The first is a multi-drug convolutional neural network (MD-CNN), designed to predict resistance phenotypes to all 13 drugs at once. The model inputs are the full sequences of 18 loci in the *M. tuberculosis* genome, selected based on known or putative roles in antibiotic resistance **(Table 2).** We chose the final MD-CNN architecture using an iterative process **(Figure 1, Supplementary Figure 1).** As superior performance of multi-task over single-task models has been demonstrated with convolutional neural networks in computer vision(*18–20*), the MD-CNN is designed to optimize performance by combining all genetic information and relating it to the full resistance antibiogram. We compare the MD-CNN with 13 single-drug convolutional neural networks (SD-CNN), each of which has a single-task, single-label architecture, in which only loci with previously known causal associations for any given drug are incorporated **(Supplementary Figure 2)**. We benchmark both types of CNNs against an existing state-of-the-art multi-drug wide-and-deep neural network (MD-WDNN)(*11*), and a logistic regression with L2 regularization penalty.

Benchmarking CNN models against state-of-the-art

We used 5-fold cross-validation to compare the performance of the four architectures (MD-CNN, the SD-CNN, L2 regression, and WDNN(*11*)) on the training dataset (N=10,201 isolates, **Supplementary Table 1**).

The mean MD-CNN AUC of 0.912 for second-line drugs is significantly higher than the mean 0.860 for L2 regression (Welch's t-test with Benjamini-Hochberg FDR $q<0.05$), but the mean AUCs for first-line drugs (0.948 *vs.* 0.923) are not significantly different (Benjamini-Hochberg $q=0.055$). The mean SD-CNN AUCs of 0.938 (first-line drugs) and 0.877 (second-line drugs) are not significantly different than for L2 regression (first-line $q=0.20$, second-line $q=0.16$). However, L2 regression demonstrates much wider confidence intervals than the CNN

models (median 0.037 versus 0.010, IQR 0.035 versus 0.014), indicating a lack of reliability as the performance depends on the particulars of the cross-validation split (**Figure 2)**.

Against the state-of-the-art WDNN, the AUCs, sensitivities, and specificities of the MD-CNN are comparable: the MD-CNN's mean AUC is 0.948 (vs. 0.959 for the MD-WDNN, q=0.15) for first-line drugs, and 0.912 (vs. 0.924 for the MD-WDNN, q=0.30) for second-line drugs. The SD-CNN is less accurate than the MD-WDNN for both first-line (Benjamini-Hochberg q=0.006) and second-line drugs (q = 0.005, **Supplementary Table 1, Figure 2).**

The SD-CNN (mean AUC of 0.938 for first-line drugs; mean AUC of 0.877 for second-line drugs) performs comparably to the MD-CNN for first-line drugs (q=0.19), and is less accurate than the MD-CNN for second-line drugs (q=0.009).

CNN models generalize well on hold-out test data

We test the generalizability and real-world applicability of our CNN models on a hold-out dataset of 12,848 isolates which were curated on a rolling basis during our study (**Table 1b, Materials and Methods**). Rolling curation provides a more realistic test of generalizability to newly produced datasets. Due to rolling curation and source difference, the test dataset exhibits different proportions of resistance to the 13 drugs (e.g. isoniazid resistance in 28% vs. 43% in the training dataset). We assessed generalizability of the models using phenotype data for 11 drugs in the hold-out test dataset, since it contained low resistance counts for ciprofloxacin and ethionamide.

We find that the MD-CNN generalizes well to never-before-seen data for first-line antibiotic resistance prediction, achieving mean AUCs of 0.965 (95% confidence interval [C.I.] 0.948 - 0.982) on both training and hold-out test sets for first-line drugs **(Figure 3)**. However, generalization for second-line drugs is mixed: for the drugs streptomycin, amikacin, ofloxacin, and moxifloxacin, the model generalizes well, achieving mean AUCs of 0.939 (CI 0.928 - 0.949) on the test data (compared with 0.939 (CI 0.929 - 0.949) on the training data). For the second-line drugs capreomycin, kanamycin, and levofloxacin, the model generalization was reduced, achieving mean AUCs of 0.831 (CI 0.824 - 0.838) on the test data (compared with 0.955 (CI 0.931 – 0.978) on the training data). We find that the SD-CNN generalizes well on first-line drug resistance for hold-out test data, with a mean AUC of 0.956 (CI 0.929 – 0.974). The SD-CNN also generalizes well for second-line drugs, with a mean AUC of 0.862 (CI 0.830 – 0.894).

We test the hypothesis that missed resistance (false negatives) is due to mutations affecting phenotype found outside of the 18 incorporated loci. To achieve this, we compute the number of mutations in the incorporated loci that separate each test isolate from the nearest isolate(s) in the training set and the corresponding phenotype of the nearest isolates **(Methods)**. We find that many of the false negatives have a genomically identical yet sensitive isolate in the training set, ranging from a minimum of 34% for pyrazinamide to a maximum of 86% for kanamycin, and suggesting that additional mutations outside of the examined loci may influence the resistance phenotype.

MD-CNN achieves accuracy by learning dependency structure of drug resistances

Because the inputs to the CNN models are the complete sequence of 18 genetic loci involved in drug resistance, we are able to assess the contribution of every site, in its

neighboring genetic context, to the prediction of antibiotic resistance phenotype. We do this by calculating an importance score for each nucleotide site in each input sequence using DeepLIFT(*21*). For any input, DeepLIFT calculates the change in predicted resistance relative to a reference input, and then backpropagates that difference through all neurons in the network to attribute the change in output to changes in the input variable. We use the pan-susceptible H37Rv genome as a reference(*22*). We take the highest magnitude (positive or negative) importance score for each nucleotide across all isolates in the training set (**Methods).**

We find evidence that the MD-CNN achieves high performance by relying on drug-drug resistance correlations. Due to the global standard therapeutic regimen for tuberculosis, resistances to first-line drugs almost always evolve before resistances to second-line drugs, and frequently in a particular order(*23*) **(Figure 4A-B)**. When considering the top 0.01% (N=17) of positions with the highest DeepLIFT importance scores for each drug, we observe that an average of 85.0% are known to confer resistance to any drug(*24*), but only a mean of 24.0% are known to confer resistance to the particular drug being investigated. For example, the top three hits for the antibiotic kanamycin are, in order, a causal hit to the *rrs* gene, an ethambutol-resistance causing hit to the *embB* gene, and a fluoroquinolone-resistance-causing hit to the *gyrA* gene (**Extended Data 1).** To probe this further, we introduce mutations that confer resistance to the first-line drugs rifampicin and isoniazid into a pan-susceptible genomic sequence background, *in silico*, and this increased the MD-CNN predicted resistance probability of pyrazinamide, streptomycin, amikacin, moxifloxacin and ofloxacin resistance (**Figure 5A**). The MD-CNN model generalized well for all five of these drugs: AUC of 0.939 for these drugs versus 0.831 for the remaining second-line drugs. Taken together, these observations show that the MD-CNN benefits from the correlation structure of antibiotic resistance.

SD-CNN saliencies highlight known and new potential predictors of resistance

We assess whether the DeepLIFT saliency scores for the SD-CNN models are able to capture known causal, resistance-conferring variants by cross-referencing the WHO catalog of established resistance-conferring mutations(*24*). We find that of the 0.1% of sites with the largest absolute DeepLIFT saliencies in each model, a large proportion are in the WHO catalog of known resistance-conferring positions (ranging from 37.5% for streptomycin to 100% for capreomycin, **Methods, Supplementary Table 2).** In total, we identify 38 variants in the top 0.1% of sites that are not previously known to cause resistance, or classified by the WHO as of "uncertain significance". Variants associated with the *M. tuberculosis* population structure comprise a smaller proportion, ranging from 0% to 8% of the top 0.1% of hits for each locus (**Methods, Supplementary Table 3**). We examine the distribution of saliency scores closely for two drugs with well understood resistance mechanisms: rifampicin and isoniazid; and for pyrazinamide a drug for which elucidating resistance mechanisms has been more challenging.

Rifampicin: Positions in the *rpoB* gene known to cause rifampicin resistance(*24*) constitute 86% of the top 0.1% and 55% of the top 1% of saliency scores **(Supplementary Figure 4).** Four of the five highest-scoring variants that have not been previously identified as resistance-causing are located in three-dimensional proximity (minimum atom distance < 8Å) to resistance-conferring variants in the RpoB protein structure, demonstrating the biological plausibility for these newly identified sites to confer resistance **(Supplementary Figure 4)**.

Isoniazid: The common causal site KatG S315 has the highest maximum saliency in the isoniazid SD-CNN **(Figure 5A)**. We observe several high saliency peaks in the promoter region of the *ahpC* gene, which are currently designated as "uncertain significance" to isoniazid resistance by the WHO(*25*). We observe three saliency peaks in the InhA protein, the mycolic acid biosynthesis enzyme targeted by isoniazid. One peak was at the known resistance-conferring mutation S94, and two at positions I21 and I194, of uncertain significance in the WHO catalogue. All three of these positions are close in 3D structure (minimum atom distance <8Å) to the bound isoniazid molecule(*26*). **(Figure 6B)**

Pyrazinamide: Of the top 1% of high saliency positions, 62% are known to be resistance-conferring, and an additional 23% are in *pncA,* but not previously known to cause resistance. The top three of these unknown *pncA* mutations are physically adjacent to known resistance-conferring mutations (**Figure 5 C,D).** The top 1% of salient positions also includes positions in *clpC1,* a gene recently implicated in pyrazinamide resistance, but mutations within are not yet recognized to be useful for resistance prediction(*27*, *28*) **(Extended Data 1).**

3. Discussion

In summary, we find that the convolutional neural networks offer similar predictive accuracy to the state-of-art MD-WDNN while also being able to discover new loci implicated in resistance, and to visualize them in their genomic context. Another major advantage of the CNNs, is that they require significantly less pre-processing and curation because they directly analyze alignments of genomic loci, allowing the models to consider not only single nucleotide polymorphisms but also sequence features such as insertions and deletions or more complex variation. They also circumvent challenges arising from differing variant naming conventions, and in reconciling variant features across datasets and time.

We find the MD-CNN's AUCs to be similar to those of the drug-specific SD-CNNs for first-line drugs, and are significantly higher for second-line drugs. CNNs generalize well to the distinct, hold-out, test isolates for these first-line antibiotics, a promising aspect if they are to be deployed in clinical practice. By contrast, there are more mixed results and generally lower hold-out test AUCs for second-line drugs. For both first- and second-line drugs, we observe that false negative isolates are often genetically identical at the considered loci to their drug-sensitive counterparts in the training dataset, indicating that additional genetic information is needed to accurately predict the phenotype for certain isolates.

Although deep neural networks are generally deemed to be less interpretable than traditional statistical methods, we are able to apply two distinct methods to interpret the network's inner logic: first, assessing model predictions using *in silico* mutagenesis; and second, assessing DeepLIFT importance scores for every input site. By computationally introducing resistance-conferring mutations into known susceptible sequences, we discover that the MD-CNN's predictions for second-line drugs relies on the correlation structure of drug resistance which is present in both the training and test set. Using DeepLIFT, we highlight which sequence features are informing model predictions: for example, our model confirmed the importance of known, resistance-conferring mutations, such as in the *rpoB,* and *katG* genes.

In addition to highlighting known resistance-conferring mutations, our model discovers 38 resistance variants previously unknown or of "uncertain significance" based on the WHO

catalogue(*24*). Including these mutations in resistance prediction may be useful for clinical diagnosis of antibiotic resistance – for example, 6% of isoniazid resistant strains contain at least one newly discovered mutation, and 2.4% contain only newly discovered mutations and no canonical resistance variants. The interpretable, nucleotide-level saliency scores permit the protein contextualization of mutations and offers the prospect of modeling how certain mutations would impact protein structure, and drug binding. This can allow for *in silico* prioritization of putative mutations for further experimental validation.

Limitations of this study include: first, the genomic variants highlighted by saliency analysis and protein contextualization cannot be confirmed to be causative without further *in silico* and *in vitro* corroboration, although further validation in independent data will support a causal role. Second, traditional laboratory-based susceptibility testing can have high variance, especially for second-line drugs, introducing a potential source of error. Third, there is insufficient phenotypic data for certain anti-TB drugs (e.g. the novel agents bedaquiline and pretomanid, and second-line agents like ethionamide). Fourth, the non-causal mutation correlations observed in the MD-CNN boosted performance, but both the training and test data were enriched for multi-drug resistance. Further assessment of generalizability to a clinical setting with a low background prevalence of multi-drug-resistant *M. tuberculosis* is needed. Finally, additional computational resources would allow the inclusion of more loci of interest, likely augmenting the performance of the MD-CNN and SD-CNNs.

We believe this to be the first study to demonstrate the feasibility of interpretable, convolutional neural networks for prediction of antibiotic resistance in *M. tuberculosis*. Greater interpretability, reliability and accuracy make this model more clinically applicable than existing benchmarks and other deep learning approaches. Saliency mapping and protein contextualization also offer the possibility of creating hypotheses on mechanisms of anti-TB drug resistance to focus further research. Along with increasingly accessible WGS-capable infrastructure globally, machine-learning-based diagnostics may support faster initialization of appropriate treatment for MDR-TB, reducing morbidity and mortality, and improving health economic endpoints(*1*, *29*).

Acknowledgements

We thank members of the Farhat lab for discussion and input. We are grateful to Dr. Peter Koo, Dr. Avika Dixit, and Greg Raskind for discussions regarding importance score calculation, validation analyses on the MD-WDNN, and CNN codebase proofreading, respectively. 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). AGG was supported by a National Institutes of Health NLM Training Grant T15LM007092 and NIH/NIAID F32AI161793. CHY was supported by the US-UK Fulbright Commission (USA/UK), the BUNAC Educational Scholarship Trust (UK), the Gavin and Ann Kellaway Research Fellowship (Auckland Medical Research Foundation, New Zealand), and the Royal Australasian College of Physicians Rowden White Fellowship (Australasia). MIG was supported by the German Research Foundation (GR5643/1-1). MF is supported by NIH/NIAID R01AI155765.

Code Availability: Implementation of all models and data analysis can be found at: https://github.com/aggreen/MTB-CNN

Materials and methods

Sequence data

The training, cross-validation, and test datasets consist of a combined 23,049 *M. tuberculosis* isolates for which whole genome sequence data and antibiotic resistance phenotype data are available. The sequencing data are obtained through the National Center for Biotechnology Information database, PATRIC, and published literature,: 10,201 strains are in the "train" dataset (for training and cross-validation) (*6*, *30–42*), 7,537 are in the hold-out "test_1" dataset (for hold-out testing) (*31*, *43–47*), and the remaining 5,312 are in the hold-out "test_GenTB" dataset (for hold-out testing) (*31*, *43–47*).

We process sequences in the train and test_1 datasets using a previously validated pipeline as described by Ezewudo et al. (2018), with modifications as elaborated by Freschi et al. (2020)(*42*, *48*). Briefly, reads are trimmed and filtered using PRINSEQ(*49*), contaminated isolates are removed using Kraken(*50*), and aligned to the reference genome H37Rv using BWA-MEM(*22*, *51*). Duplicate reads are removed using Picard, and we dropped isolates with less than 95% coverage of the reference genome at 10x coverage.

For the "test_GenTB" dataset, we prepare the sequencing data in accordance with the protocol in Groschel *et al.*(*52*) , a different variant of the Ezewudo et al. pipeline.

With regard to curated genetic variants, the predictor sets of features for the multi-drug wide and deep neural network (MD-WDNN, see *Machine learning models* below) are processed as described by Chen et al. (2019)(*11*). Conversely, for the single-drug and multi-drug convolutional neural networks (SD-CNN and MD-CNN, see *Machine learning models* below), only the FASTA files for the loci of interest are necessary.

Antimicrobial resistance phenotype data

Culture-based antimicrobial drug susceptibility to 2-to-13 anti-TB drugs are available for all 23,049 isolates in the combined training, cross-validation, and test dataset, collated with quality control criteria described by Farhat et al. (2016)(*3*). Phenotypes (drug susceptibility test results) for isolates in the training and cross-validation dataset are from the ReSeqTB data portal, the PATRIC database, and manual curation of phenotypic data available in the literature(*6*, *30–42*). Phenotypes for the test dataset isolates are from data available in the literature(*31*, *43–47*). Each isolate's phenotype is classified as resistant, susceptible, or unavailable, with respect to a combination of 13 possible first-line (rifampicin, isoniazid, pyrazinamide, ethambutol) and second-line drugs (streptomycin, ciprofloxacin, levofloxacin, moxifloxacin, ofloxacin, capreomycin, amikacin, kanamycin, ethionamide). (**Table 1**). In the hold-out test dataset, ethionamide and ciprofloxacin were excluded due to data missingness (0/2 resistant to ciprofloxacin; 12/25 resistant to ethionamide).

Selecting input loci

The loci of the isolate sequences are selected from genes known or suspected to cause resistance based on previous models and experiments (**Table 1**). In order to incorporate regulatory sequences from the immediate genetic neighborhood, regions upstream from genes of interest are included. Loci were aligned to the H37Rv reference genome for comparison of coordinates and genome annotations are based on H37Rv coordinates from Mycobrowser(*53*).

Machine learning models

The Multi-Drug (multi-task) Wide-and-Deep Neural Network (MD-WDNN) is described by Chen et al. (2019), and involves three hidden layers (256 ReLU), dropout, and batch normalization(*11*)

The Multi-Drug Convolutional Neural Network (MD-CNN) comprises two convolution layers (with filter size 12 nucleotides in length), one max-pooling layer, two convolution layers, one max-pooling layer, followed by two fully-connected hidden layers each with 256 rectified linear units (ReLU) (Table 1). This architecture is selected based on its performance, as defined by area under the receiver operator characteristic curve (AUC), compared to other architectures with fewer convolutional layers and differential filter sizes **(Supplementary Figure 1)**. Neither random nor cartesian grid search of optimal hyperparameters is conducted.

The MD-CNN is trained for 250 epochs via stochastic gradient descent and the Adam optimizer (learning rate of e^{-9}). We select an optimal number of epochs based on minimizing validation loss (**Supplementary Figure 5**). The training is performed simultaneously using the resistance phenotype for all 13 drugs, hence the 13 nodes in the final output layer (Table 1), the output of each node corresponding to the sigmoid probability of the strain being resistant to the respective drug.

The MD-CNN's loss function is adapted from the masked, class-weighted binary cross-entropy function described by Chen et al. (2019)(*11*). This function addresses the dataset imbalance (missing resistance phenotypes for a varying number of drugs in any given isolate) by upweighting the sparser of the susceptible and resistant classes for each drug, and masking outputs where resistance status was completely missing.

The Single-Drug Convolutional Neural Networks (SD-CNNs) are thirteen individually trained convolutional neural networks, each trained to predict for only one drug, hence the output layer having size 1 instead of 13. Each SD-CNN is given only the input loci relevant to its particular antibiotic, resulting in different input sizes depending on the longest locus for each drug. The architecture for the SD-CNNs is otherwise identical to that of the MD-CNN. The SD-CNNs are initially trained for 150 epochs using stochastic gradient descent and the Adam optimizer (learning rate of e^{-9}) and an optimal number of epochs for each SD-CNN is selected to minimize the validation loss **(Supplementary Table 4)**.

Logistic regression benchmark

We build a logistic regression benchmark to evaluate the performance of our neural network models. For each of the 18 input loci used in the MD-CNN and SD-CNNs, we select all sites with a minor allele frequency of at least 0.1%, resulting in 3,011 sites across 23,049 genomes. Sites are then encoded using a major/minor allele encoding.

Using the same train/test partitioning as for the neural network models, we use GridSearchCV in Scikit-learn v.0.23.2(*54*) to select the optimal L2 penalty weight for a LogisticRegression classifier with balanced class weights. Hyperparameter search is performed for each drug independently, testing the values C=[0.0001, 0.001, 0.01, 0.1, 1]. After selecting the optimal L2 weight, we use five-fold cross-validation on the training set to assess the AUC,

specificity, and sensitivity, selecting a model threshold that maximized the sum of specificity and sensitivity.

Training and model evaluation

Five-fold cross-validation is performed five times to obtain the performance metrics – area under the receiver operator characteristic curve (AUC), sensitivity, specificity, and probability threshold (to maximize the sum of sensitivity and specificity) – and the 95% confidence intervals of the AUC values between the models.

Model performance on the hold-out test sets is evaluated using the probability threshold selected during training.

Computational details

The MD-CNN is developed and implemented using TensorFlow 2.3.0 in Python 3.7.9 with CUDA 10.1(*55*–*57*). Model training is performed on an NVIDIA GeForce GTX Titan X graphics processing unit (GPU).

Analysis of mis-predicted isolates

For each SD-CNN model, we compute the genetic distance (number of different sites) between all isolates in the training and test sets. Only the loci included in each SD-CNN model are incorporated in the calculation.

Importance Score calculation

Importance scores are calculated using DeepLIFT v. 0.6.12.0, using the recommended defaults for genomics: "rescale" rule applied to convolutional layers, and "reveal-cancel" rule applied to fully connected layers. We use the H37Rv reference genome, which is sensitive to all antibiotics, as the baseline(*22*).

Importance scores for each isolate sequence are calculated relative to the H37Rv baseline. For our analysis of positions influencing antibiotic resistance prediction, we take the maximum of the absolute value of the scores at each position across all resistant isolates.

Lineage variant analysis

We define lineage variants as those found in the Coll *et al.* or Freschi *et al.* barcode of lineage-defining variants(*58*, *59*). We further annotate any position in our 18 loci as lineage associated if that position has an identical distribution of major/minor alleles to any position in the Freschi *et al* barcode, excluding the position *1,137,518* which defines lineage 7 (not present in our dataset).

Tables

Drug	Resistant (n)	Susceptible (n)	Total (n)	Resistant proportion
ISONIAZID	4232	5723	9955	0.425
RIFAMPICIN	3472	6428	9900	0.351
ETHAMBUTOL	2273	6390	8663	0.262
PYRAZINAMIDE	1505	5393	6898	0.218
STREPTOMYCIN	2643	4362	7005	0.377
AMIKACIN	773	2632	3405	0.227
CAPREOMYCIN	737	2838	3575	0.206
KANAMYCIN	796	2502	3298	0.241
CIPROFLOXACIN	118	388	506	0.233
OFLOXACIN	912	2246	3158	0.289
MOXIFLOXACIN	398	1941	2339	0.170
LEVOFLOXACIN	66	189	255	0.259
ETHIONAMIDE	791	1647	2438	0.324
Total isolates			10201	

Table 1a: training & cross-validation isolates

Drug	Resistant (n)	Susceptible (n)	Total (n)	Resistant proportion
ISONIAZID	3384	8870	12254	0.276
RIFAMPICIN	3007	9708	12715	0.236
ETHAMBUTOL	1498	7853	9351	0.160
PYRAZINAMIDE	1211	7490	8701	0.139
STREPTOMYCIN	382	1756	2138	0.179
AMIKACIN	93	1481	1574	0.059
CAPREOMYCIN	61	1652	1713	0.036
KANAMYCIN	83	2202	2285	0.036
OFLOXACIN	230	2897	3127	0.074
MOXIFLOXACIN	103	2495	2598	0.040
LEVOFLOXACIN	85	49	134	0.634
Total isolates			12848	

Table 1b: test dataset isolates

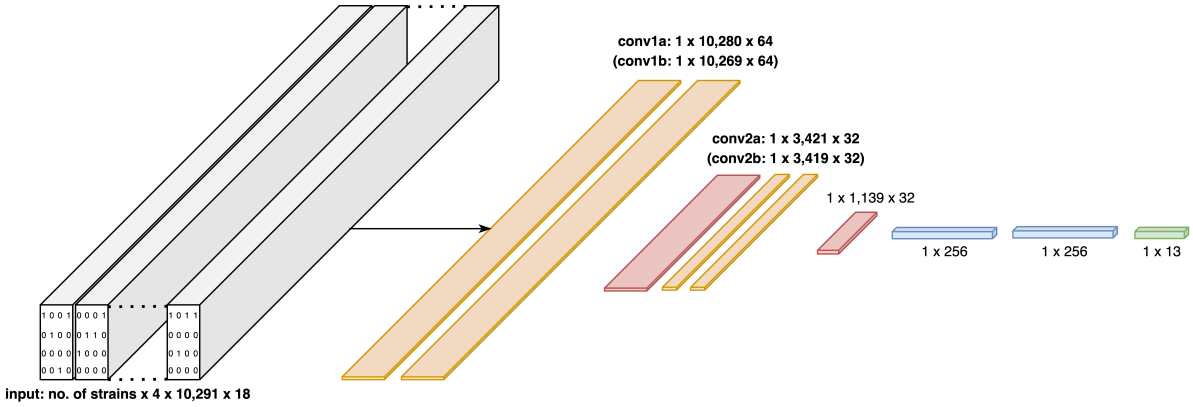
Tables 1a & 1b : Phenotypic summary of the 23,049 isolates used to train and cross-validate (1a), and test (1b) the models: the numbers of resistant isolates, susceptible isolates, the total tested (sum of the numbers of resistant and susceptible isolates), and the resistant proportion, with respect to each of the 13 anti-TB drugs (training and cross-validation) or 11 anti-TB drugs (test). Ciprofloxacin and ethionamide were excluded from the test dataset due to small numbers (0/2 resistant to ciprofloxacin; 12/25 resistant to ethionamide).

Locus	Start	End	Drug(s)	Length (in H37Rv)
acpM-kasA	2517695	2519365	Isoniazid	1670
gid	4407528	4408334	Streptomycin	806
rpsA	1833378	1834987	Pyrazinamide	1609
clpC	4036731	4040937	Pyrazinamide	4206
embCAB	4239663	4249810	Ethambutol	10147
aftB-ubiA	4266953	4269833	Ethambutol	2880
rrs-rrl	1471576	1477013	Streptomycin, Amikacin, Capreomycin, Kanamycin	5437
ethAR	4326004	4328199	Ethionamide	2195
oxyR-ahpC	2725477	2726780	Isoniazid	1303
tlyA	1917755	1918746	Capreomycin	991
katG	2153235	2156706	Isoniazid	3471
rpsL	781311	781934	Streptomycin	623
rpoBC	759609	767320	Rifampicin	7711
fabG1-inhA	1672457	1675011	Isoniazid, Ethionamide	2554
eis	2713783	2716314	Kanamycin, Amikacin	2531
gyrBA	4997	9818	Ciprofloxacin, Levofloxacin, Moxifloxacin, Ofloxacin	4821
panD	4043041	4045210	Pyrazinamide	2169
pncA	2287883	2289599	Pyrazinamide	1716

Table 2: Loci included in the MD-CNN and SD-CNN models. The 18 loci included in the MD-CNN and their start and end coordinates (in H37Rv numbering). Each locus was designated as putatively involved in resistance to at least one drug. To construct the 13 SD-CNN models, the relevant loci for each drug were combined – for example, the isoniazid (INH) model contained the *acpM-kasA, oxyR-ahpC, katG, and fabG1-inhA* loci.

Figures

Figure 1: schematic diagram and table of the multi-drug convolutional neural network (MD-CNN). In the output layer, each of the 13 nodes is composed of a sigmoid function to compute a probability of resistance for their respective anti-TB drug (13 anti-TB drugs in total). The input consisted of '10,201' isolates (TB strains) for which there was resistance phenotype data for at least 2 anti-TB drugs; '5' for one-hot encoding of each nucleotide (5 dimensions, one for each nucleotide – adenine, thymine, guanine, cytosine plus gaps); '10,291' being the number of nucleotides of the longest locus (*embC-embA-embB*); '18' loci of interest were incorporated as detailed in 'Materials and methods'.



convolution + ReLU
max pooling
fully-connected + ReLU
sigmoid

Layer	Operation	Number of Filters	Filter Size	Stride	Output Dimensions
Input	-	-	-	-	10,201 x 5 x 10,291 x 18
1D convolution	Convolution ReLU	64	5 x 12	1 x 1	10,201 x 1 x 10,280 x 64
1D convolution	Convolution ReLU	64	1 x 12	1 x 1	10,201 x 1 x 10,269 x 64
Pooling	Max pooling	1	1 x 3	1 x 1	10,201 x 1 x 3,423 x 64
1D convolution	Convolution ReLU	32	1 x 3	1 x 1	10,201 x 1 x 3,421 x 32
1D convolution	Convolution ReLU	32	1 x 3	1 x 1	10,201 x 1 x 3,419 x 32
Pooling	Max pooling	1	1 x 3	1 x 1	10,201 x 1 x 1,139 x 32
Inner product (two times)	Fully connected ReLU	-	-	-	256
Output	-	-	-	-	13

Figure 2: MD-CNN performs comparably to state-of-art WDNN for both first- and second-line drugs. Results of five-fold cross validation on the training dataset for the four models: WDNN, logistic regression + L2 benchmark, SD-CNN, and MD-CNN. (A) mean AUC and 95% confidence intervals, pooled for first and second line drugs. (B) mean AUC and 95% confidence intervals for each drug. The WDNN was not initially trained on levofloxacin or ethionamide and thus was not evaluated for these drugs.

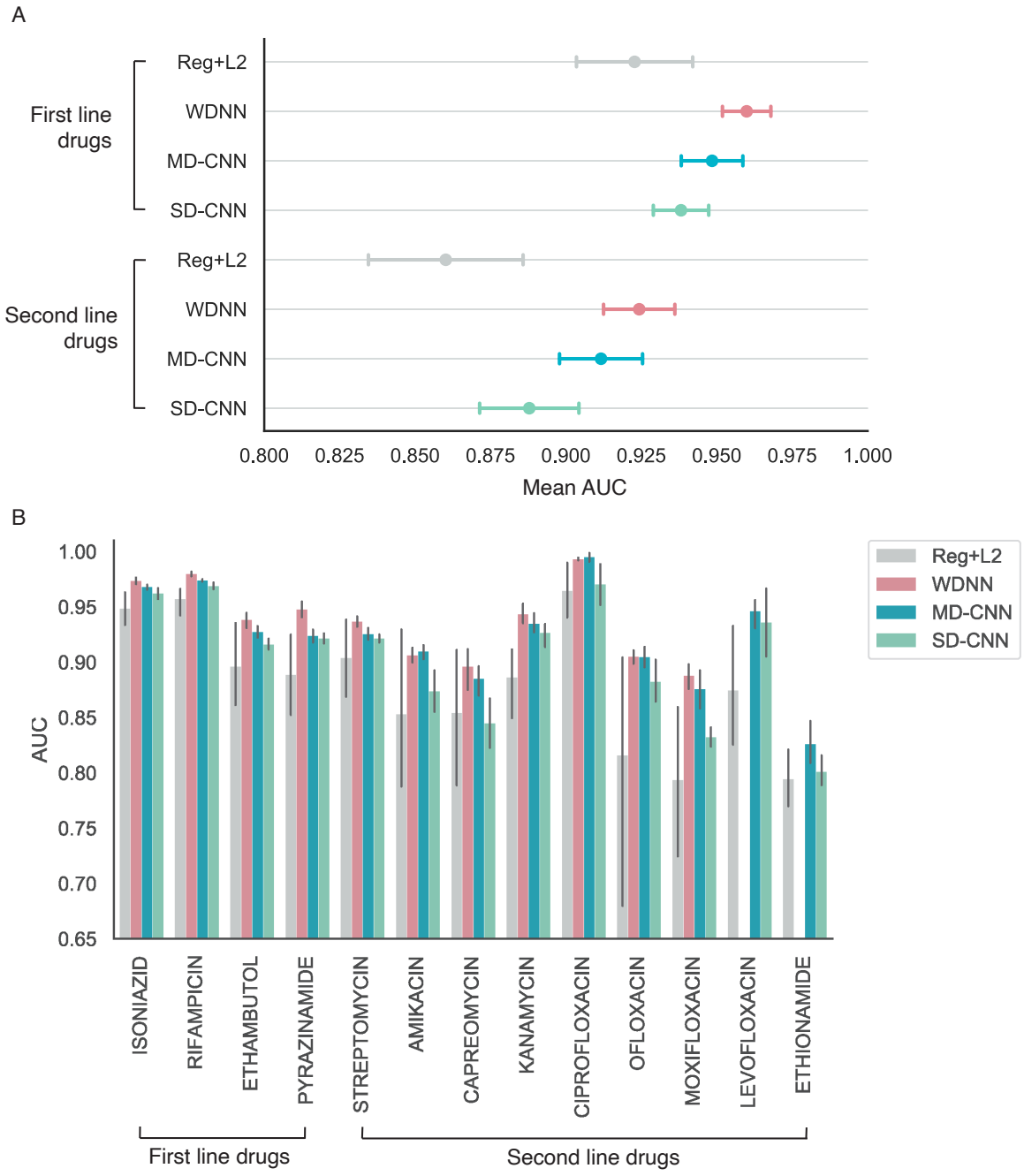


Figure 3: MD-CNN and SD-CNN model generalize well on hold-out test data. Performance of CNN models trained on the entire training dataset evaluated on either the training dataset or the hold-out test dataset. (A) Mean AUC and 95% confidence intervals (calculated across drugs) for first- and second-line drugs, pooled. (B) Mean AUC for each drug with confidence intervals generated by 100x bootstrapping with 80% of isolates. Ciprofloxacin and ethionamide were not assessed due to low number of resistant isolates.

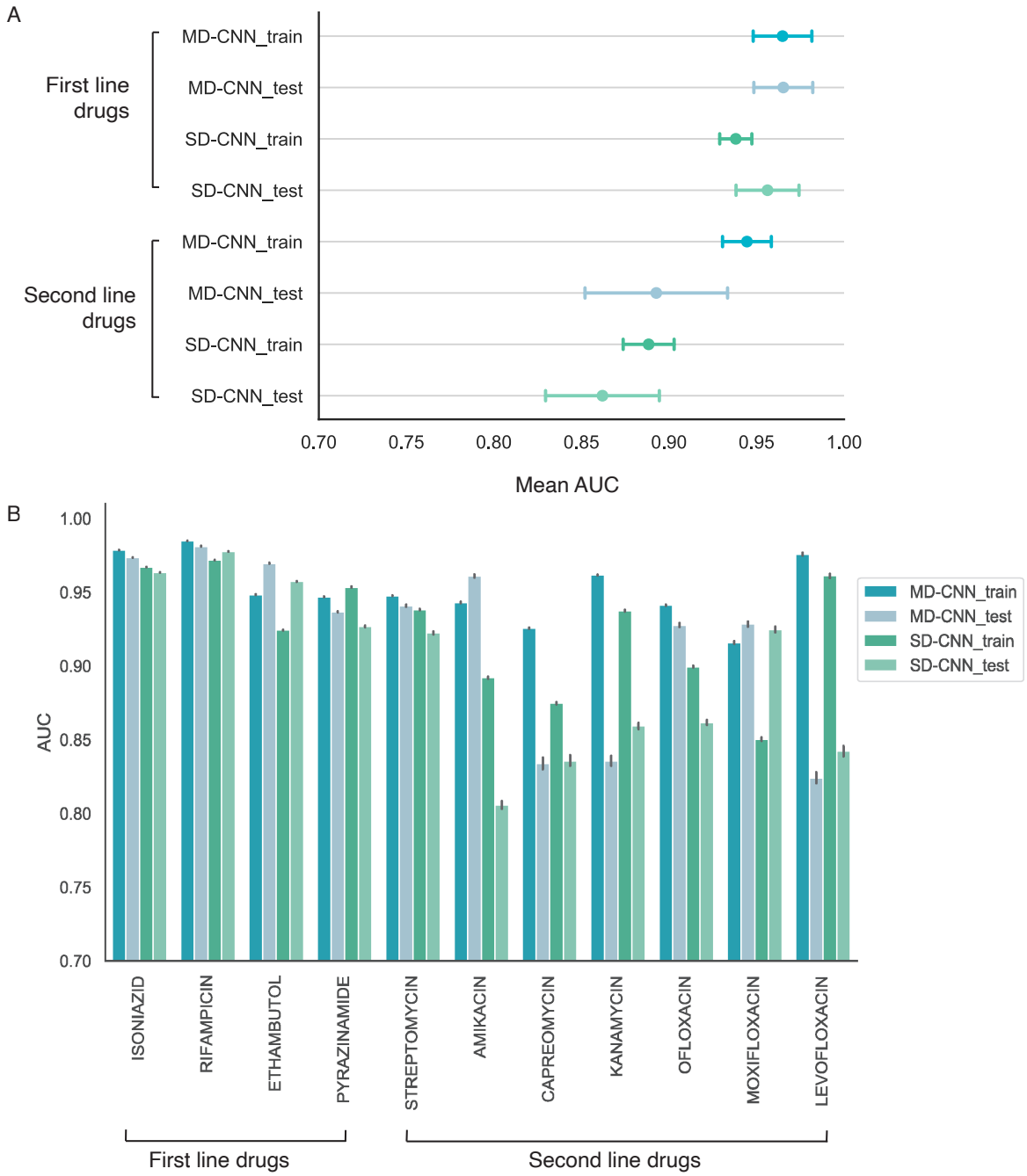


Figure 4: MD-CNN learns dependency structure of antibiotic resistance. (A) Introduction of single resistance-conferring mutations into pan-susceptible wild-type background (H37Rv) is sufficient to cause MD-CNN model to predict false positive resistances. A single isoniazid-resistance conferring mutations (2155168G, *katG* S315T) or one isoniazid- and one rifampicin-resistance conferring mutation (2155168G and 761155T, *rpoB* S450L) were introduced *in silico* into the wild-type background sequence and resistances were predicted using the MD-CNN model. **(B)** Dependency heatmaps of drug resistance for training isolates. The horizontal axis represents the drugs to which isolates exhibited resistance. Based on this condition of resistance, the proportion of resistance to other drugs (vertical axis) was computed.

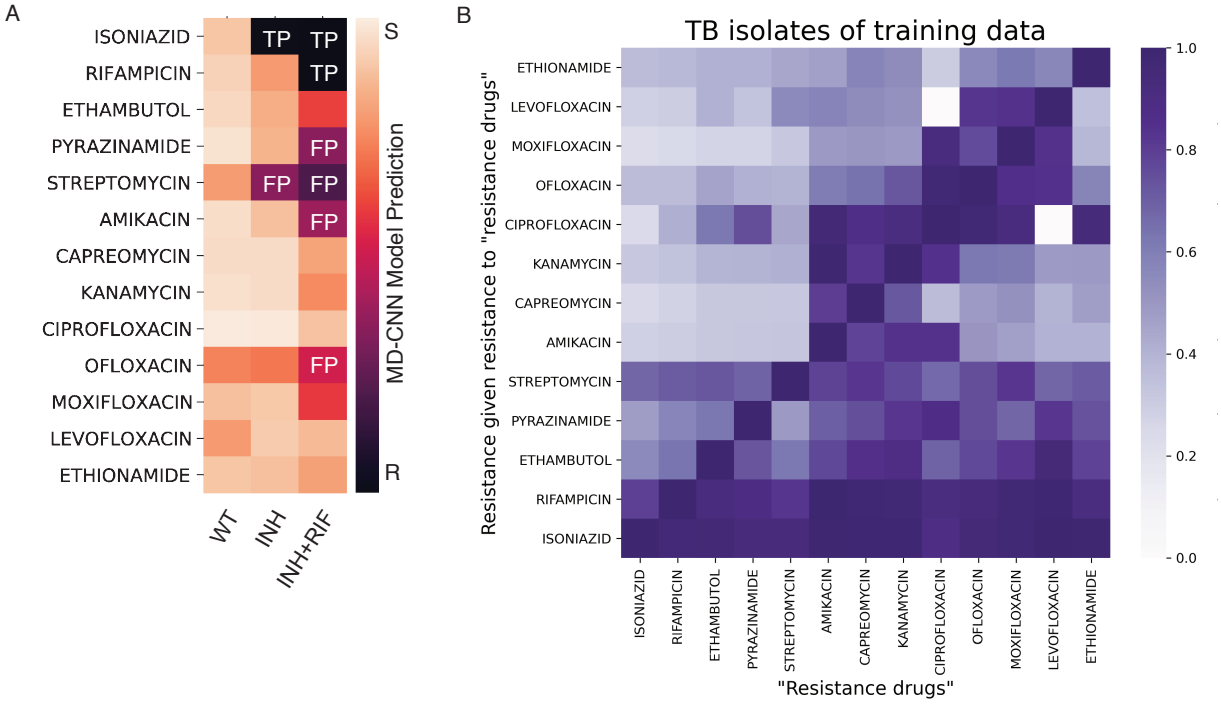
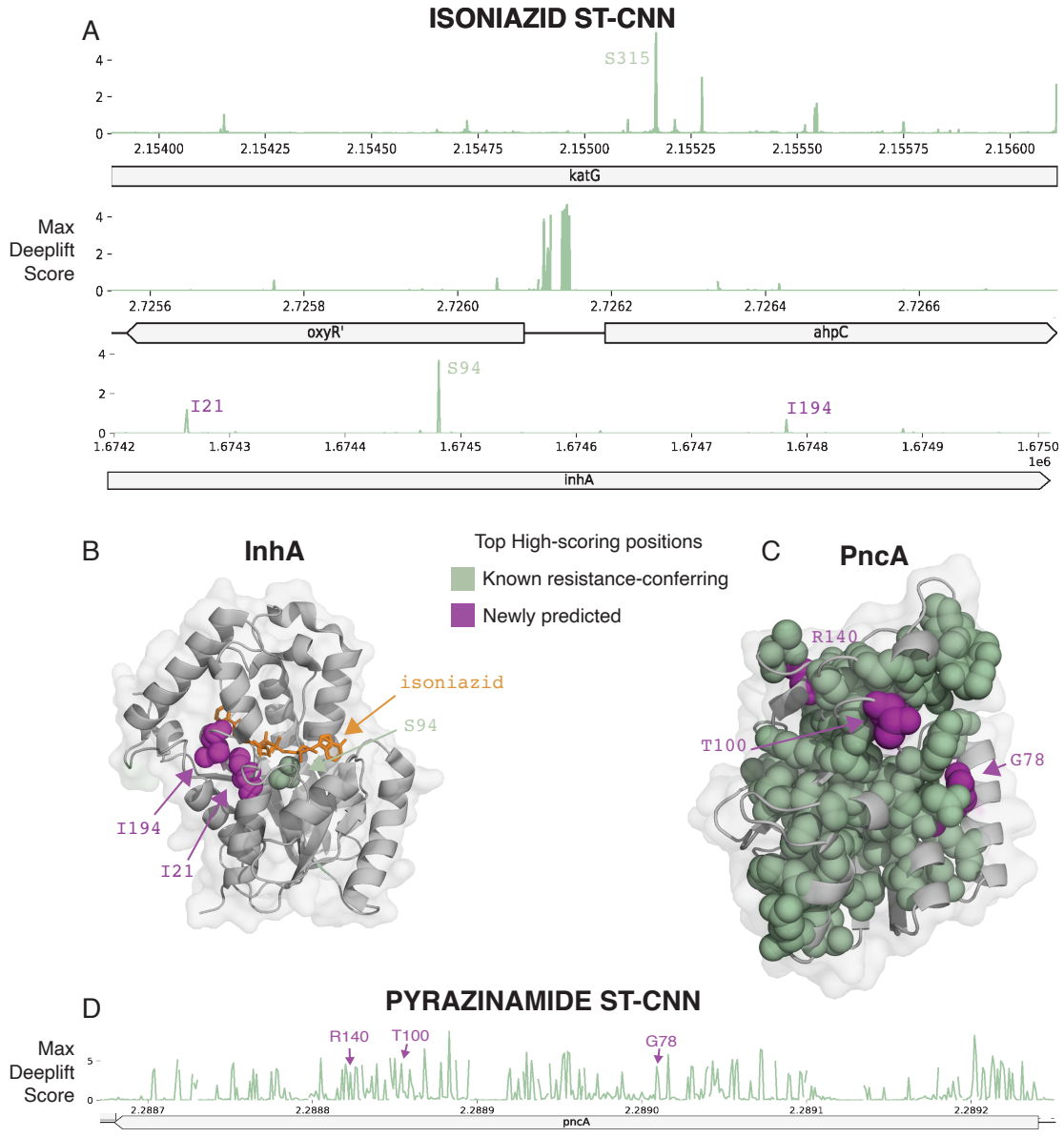


Figure 5: SD-CNN importance scores highlight known and plausible new resistance-conferring loci. Variants not known to cause resistance according to the WHO(*24*) are shown in purple. (A**)** Maximum of absolute value DeepLIFT importance scores for the isoniazid SD-CNN across all isoniazid-resistant loci. (B) High-importance variants in the InhA protein mapped to its crystal structure(*60*). (C) High-importance variants in the PncA protein mapped to its crystal structure(*61*). (D) Maximum of absolute value DeepLIFT importance scores for the pyrazinamide SD-CNN in the *pncA* locus.



References

1. WHO, "Global tuberculosis report 2018" (World Health Organization, 2018), (available at https://apps.who.int/iris/bitstream/handle/10665/274453/9789241565646-eng.pdf?sequence=1&isAllowed=y).

2. C. Lange, D. Chesov, J. Heyckendorf, C. C. Leung, Z. Udwadia, K. Dheda, Drug-resistant tuberculosis: An update on disease burden, diagnosis and treatment. *Respirology*. **23**, 656–673 (2018).

3. M. R. Farhat, R. Sultana, O. Iartchouk, S. Bozeman, J. Galagan, P. Sisk, C. Stolte, H. Nebenzahl-Guimaraes, K. Jacobson, A. Sloutsky, D. Kaur, J. Posey, B. N. Kreiswirth, N. Kurepina, L. Rigouts, E. M. Streicher, T. C. Victor, R. M. Warren, D. van Soolingen, M. Murray, Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. *Am. J. Respir. Crit. Care Med.* **194**, 621–630 (2016).

4. C. Allix-Beguec, I. Arandjelovic, L. Bi, P. Beckert, M. Bonnet, P. Bradley, A. M. Cabibbe, I. Cancino-Munoz, M. J. Caulfield, A. Chaiprasert, D. M. Cirillo, D. A. Clifton, I. Comas, D. W. Crook, M. R. De Filippo, H. de Neeling, R. Diel, F. A. Drobniewski, K. Faksri, M. R. Farhat, J. Fleming, P. Fowler, T. A. Fowler, Q. Gao, J. Gardy, D. Gascoyne-Binzi, A. L. Gibertoni-Cruz, A. Gil-Brusola, T. Golubchik, X. Gonzalo, L. Grandjean, G. He, J. L. Guthrie, S. Hoosdally, M. Hunt, Z. Iqbal, N. Ismail, J. Johnston, F. M. Khanzada, C. C. Khor, T. A. Kohl, C. Kong, S. Lipworth, Q. Liu, G. Maphalala, E. Martinez, V. Mathys, M. Merker, P. Miotto, N. Mistry, D. A. J. Moore, M. Murray, S. Niemann, S. V. Omar, R. T. Ong, T. E. A. Peto, J. E. Posey, T. Prammananan, A. Pym, C. Rodrigues, M. Rodrigues, T. Rodwell, G. M. Rossolini, E. Sanchez Padilla, M. Schito, X. Shen, J. Shendure, V. Sintchenko, A. Sloutsky, E. G. Smith, M. Snyder, K. Soetaert, A. M. Starks, P. Supply, P. Suriyapol, S. Tahseen, P. Tang, Y. Y. Teo, T. N. T. Thuong, G. Thwaites, E. Tortoli, D. van Soolingen, A. S. Walker, T. M. Walker, M. Wilcox, D. J. Wilson, D. Wyllie, Y. Yang, H. Zhang, Y. Zhao, B. Zhu, Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N. Engl. J. Med.* **379**, 1403–1415 (2018).

5. M. Hunt, P. Bradley, S. G. Lapierre, S. Heys, M. Thomsit, M. B. Hall, K. M. Malone, P. Wintringer, T. M. Walker, D. M. Cirillo, I. Comas, M. R. Farhat, P. Fowler, J. Gardy, N. Ismail, T. A. Kohl, V. Mathys, M. Merker, S. Niemann, S. V. Omar, V. Sintchenko, G. Smith, D. van Soolingen, P. Supply, S. Tahseen, M. Wilcox, I. Arandjelovic, T. E. A. Peto, D. W. Crook, Z. Iqbal, Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Res*. **4**, 191 (2019).

6. T. M. Walker, T. A. Kohl, S. V. Omar, J. Hedge, C. Del Ojo Elias, P. Bradley, Z. Iqbal, S. Feuerriegel, K. E. Niehaus, D. J. Wilson, D. A. Clifton, G. Kapatai, C. L. C. Ip, R. Bowden, F. A. Drobniewski, C. Allix-Beguec, C. Gaudin, J. Parkhill, R. Diel, P. Supply, D. W. Crook, E. G. Smith, A. S. Walker, N. Ismail, S. Niemann, T. E. A. Peto, Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect. Dis.* **15**, 1193–1202 (2015).

7. Y. Yang, K. E. Niehaus, T. M. Walker, Z. Iqbal, A. S. Walker, D. J. Wilson, T. E. A. Peto, D. W. Crook, E. G. Smith, T. Zhu, D. A. Clifton, Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics*. **34**, 1666–1671 (2018).

8. M. Ghassemi, L. Oakden-Rayner, A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. **3**, e745–e750 (2021).

9. I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, New York, NY, USA, 2020), *FAT* '20*, pp. 33–44.

10. M. W. Libbrecht, W. S. Noble, Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).

11. M. L. Chen, A. Doddi, J. Royer, L. Freschi, M. Schito, M. Ezewudo, I. S. Kohane, A. Beam, M. Farhat, Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. *EBioMedicine*. **43**, 356–369 (2019).

12. H. Zabeti, N. Dexter, A. H. Safari, N. Sedaghat, M. Libbrecht, L. Chindelevitch, INGOT-DR: an interpretable classifier for predicting drug resistance in M. tuberculosis. *Algorithms Mol. Biol.* **16**, 17 (2021).

13. A. Drouin, G. Letarte, F. Raymond, M. Marchand, J. Corbeil, F. Laviolette, Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci. Rep.* **9**, 1–13 (2019).

14. P. K. Koo, S. R. Eddy, Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Comput. Biol.* **15**, e1007560 (2019).

15. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity Checks for Saliency Maps. *arXiv [cs.CV]* (2018), (available at http://arxiv.org/abs/1810.03292).

16. P. K. Koo, S. Qian, G. Kaplun, V. Volf, D. Kalimeris, Robust Neural Networks are More Interpretable for Genomics. *Cold Spring Harbor Laboratory* (2019), p. 657437.

17. C. H. Yoon, R. Torrance, N. Scheinerman, Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *J. Med. Ethics* (2021), doi:10.1136/medethics-2020-107102.

18. A. Dobrescu, M. V. Giuffrida, S. A. Tsaftaris, Doing More With Less: A Multitask Deep Learning Approach in Plant Phenotyping. *Front. Plant Sci.* **11**, 141 (2020).

19. C. Zhang, Z. Zhang, in *IEEE Winter Conference on Applications of Computer Vision* (2014), pp. 1036–1041.

20. R. Caruana, Multitask Learning. *Mach. Learn.* **28**, 41–75 (1997).

21. A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, 3145–3153 (2017).

22. S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton,

R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, B. G. Barrell, Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*. **393**, 537–544 (1998).

23. Y. Ektefaie, A. Dixit, L. Freschi, M. R. Farhat, Globally diverse Mycobacterium tuberculosis resistance acquisition: a retrospective geographical and temporal analysis of whole genome sequences. *Lancet Microbe*. **2**, e96–e104 (2021).

24. W. H. (hq) Global Tuberculosis Programme, "WHO consolidated guidelines on tuberculosis. Module 3: Diagnosis - Rapid diagnostics for tuberculosis detection 2021 update" (2021).

25. T. M. Wilson, D. M. Collins, ahpC, a gene involved in isoniazid resistance of the Mycobacterium tuberculosis complex. *Mol. Microbiol.* **19**, 1025–1034 (1996).

26. C. Vilchèze, F. Wang, M. Arai, M. H. Hazbón, R. Colangeli, L. Kremer, T. R. Weisbrod, D. Alland, J. C. Sacchettini, W. R. Jacobs Jr, Transfer of a point mutation in Mycobacterium tuberculosis inhA resolves the target of isoniazid. *Nat. Med.* **12**, 1027–1029 (2006).

27. E. A. Lamont, N. A. Dillon, A. D. Baughn, The Bewildering Antitubercular Action of Pyrazinamide. *Microbiol. Mol. Biol. Rev.* **84** (2020), doi:10.1128/MMBR.00070-19.

28. P. Gopal, J. P. Sarathy, M. Yee, P. Ragunathan, J. Shin, S. Bhushan, J. Zhu, T. Akopian, O. Kandror, T. K. Lim, M. Gengenbacher, Q. Lin, E. J. Rubin, G. Grüber, T. Dick, Pyrazinamide triggers degradation of its target aspartate decarboxylase. *Nat. Commun.* **11**, 1661 (2020).

29. Y. Chen, Z. Yuan, X. Shen, J. Wu, Z. Wu, B. Xu, Time to Multidrug-Resistant Tuberculosis Treatment Initiation in Association with Treatment Outcomes in Shanghai, China. *Antimicrob. Agents Chemother.* **62**, e02259-17 (2018).

30. A. R. Wattam, J. J. Davis, R. Assaf, S. Boisvert, T. Brettin, C. Bun, N. Conrad, E. M. Dietrich, T. Disz, J. L. Gabbard, S. Gerdes, C. S. Henry, R. W. Kenyon, D. Machi, C. Mao, E. K. Nordberg, G. J. Olsen, D. E. Murphy-Olson, R. Olson, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, V. Vonstein, A. Warren, F. Xia, H. Yoo, R. L. Stevens, Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* **45**, D535-d542 (2017).

31. F. Coll, J. Phelan, G. A. Hill-Cawthorne, M. B. Nair, K. Mallard, S. Ali, A. M. Abdallah, S. Alghamdi, M. Alsomali, A. O. Ahmed, S. Portelli, Y. Oppong, A. Alves, T. B. Bessa, S. Campino, M. Caws, A. Chatterjee, A. C. Crampin, K. Dheda, N. Furnham, J. R. Glynn, L. Grandjean, D. Minh Ha, R. Hasan, Z. Hasan, M. L. Hibberd, M. Joloba, E. C. Jones-López, T. Matsumoto, A. Miranda, D. J. Moore, N. Mocillo, S. Panaiotov, J. Parkhill, C. Penha, J. Perdigão, I. Portugal, Z. Rchiad, J. Robledo, P. Sheen, N. T. Shesha, F. A. Sirgel, C. Sola, E. Oliveira Sousa, E. M. Streicher, P. Van Helden, M. Viveiros, R. M. Warren, R. McNerney, A. Pain, T. G. Clark, Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **50**, 307–316 (2018).

32. T. M. Walker, C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, J. Parkhill, D. Harris, A. S. Walker, R. Bowden,

P. Monk, E. G. Smith, T. E. Peto, Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).

33. H. Zhang, D. Li, L. Zhao, J. Fleming, N. Lin, T. Wang, Z. Liu, C. Li, N. Galwey, J. Deng, Y. Zhou, Y. Zhu, Y. Gao, T. Wang, S. Wang, Y. Huang, M. Wang, Q. Zhong, L. Zhou, T. Chen, J. Zhou, R. Yang, G. Zhu, H. Hang, J. Zhang, F. Li, K. Wan, J. Wang, X. E. Zhang, L. Bi, Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat. Genet.* **45**, 1255–1260 (2013).

34. K. A. Cohen, T. Abeel, A. Manson McGuire, C. A. Desjardins, V. Munsamy, T. P. Shea, B. J. Walker, N. Bantubani, D. V. Almeida, L. Alvarado, S. B. Chapman, N. R. Mvelase, E. Y. Duffy, M. G. Fitzgerald, P. Govender, S. Gujja, S. Hamilton, C. Howarth, J. D. Larimer, K. Maharaj, M. D. Pearson, M. E. Priest, Q. Zeng, N. Padayatchi, J. Grosset, S. K. Young, J. Wortman, K. P. Mlisana, M. R. O'Donnell, B. W. Birren, W. R. Bishai, A. S. Pym, A. M. Earl, Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. *PLoS Med.* **12**, e1001880 (2015).

35. Y. Blouin, Y. Hauck, C. Soler, M. Fabre, R. Vong, C. Dehan, G. Cazajous, P.-L. Massoure, P. Kraemer, A. Jenkins, E. Garnotel, C. Pourcel, G. Vergnaud, Significance of the Identification in the Horn of Africa of an Exceptionally Deep Branching Mycobacterium tuberculosis Clade. *PLoS One*. **7**, e52841 (2012).

36. T. G. Clark, K. Mallard, F. Coll, M. Preston, S. Assefa, D. Harris, S. Ogwang, F. Mumbowa, B. Kirenga, D. M. O'Sullivan, A. Okwera, K. D. Eisenach, M. Joloba, S. D. Bentley, J. J. Ellner, J. Parkhill, E. C. Jones-López, R. McNerney, Elucidating Emergence and Transmission of Multidrug-Resistant Tuberculosis in Treatment Experienced Patients by Whole Genome Sequencing. *PLoS One*. **8**, e83012 (2013).

37. J. M. Bryant, A. C. Schurch, H. van Deutekom, S. R. Harris, J. L. de Beer, V. de Jager, K. Kremer, S. A. van Hijum, R. J. Siezen, M. Borgdorff, S. D. Bentley, J. Parkhill, D. van Soolingen, Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infect. Dis.* **13**, 110 (2013).

38. A. Chatterjee, K. Nilgiriwala, D. Saranath, C. Rodrigues, N. Mistry, Whole genome sequencing of clinical strains of Mycobacterium tuberculosis from Mumbai, India: A potential tool for determining drug-resistance and strain lineage. *Kekkaku*. **107**, 63–72 (2017).

39. M. Merker, C. Blin, S. Mona, N. Duforet-Frebourg, S. Lecher, E. Willery, M. G. Blum, S. Rusch-Gerdes, I. Mokrousov, E. Aleksic, C. Allix-Beguec, A. Antierens, E. Augustynowicz-Kopec, M. Ballif, F. Barletta, H. P. Beck, C. E. Barry 3rd, M. Bonnet, E. Borroni, I. Campos-Herrero, D. Cirillo, H. Cox, S. Crowe, V. Crudu, R. Diel, F. Drobniewski, M. Fauville-Dufaux, S. Gagneux, S. Ghebremichael, M. Hanekom, S. Hoffner, W. W. Jiao, S. Kalon, T. A. Kohl, I. Kontsevaya, T. Lillebaek, S. Maeda, V. Nikolayevskyy, M. Rasmussen, N. Rastogi, S. Samper, E. Sanchez-Padilla, B. Savic, I. C. Shamputa, A. Shen, L. H. Sng, P. Stakenas, K. Toit, F. Varaine, D. Vukovic, C. Wahl, R. Warren, P. Supply, S. Niemann, T. Wirth, Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nat. Genet.* **47**, 242–249 (2015).

40. J. L. Gardy, J. C. Johnston, S. J. H. Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. M. Jones, F. S. L. Brinkman, R. C. Brunham, P. Tang, Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).

41. J. J. Davis, A. R. Wattam, R. K. Aziz, T. Brettin, R. Butler, R. M. Butler, P. Chlenski, N. Conrad, A. Dickerman, E. M. Dietrich, J. L. Gabbard, S. Gerdes, A. Guard, R. W. Kenyon, D. Machi, C. Mao, D. Murphy-Olson, M. Nguyen, E. K. Nordberg, G. J. Olsen, R. D. Olson, J. C. Overbeek, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, C. Thomas, M. VanOeffelen, V. Vonstein, A. S. Warren, F. Xia, D. Xie, H. Yoo, R. Stevens, The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).

42. M. Ezewudo, A. Borens, A. Chiner-Oms, P. Miotto, L. Chindelevitch, A. M. Starks, D. Hanna, R. Liwski, M. Zignol, C. Gilpin, S. Niemann, T. A. Kohl, R. M. Warren, D. Crook, S. Gagneux, S. Hoffner, C. Rodrigues, I. Comas, D. M. Engelthaler, D. Alland, L. Rigouts, C. Lange, K. Dheda, R. Hasan, R. McNerney, D. M. Cirillo, M. Schito, T. C. Rodwell, J. Posey, Integrating standardized whole genome sequence analysis with a global Mycobacterium tuberculosis antibiotic resistance knowledgebase. *Sci. Rep.* **8**, 15382 (2018).

43. M. Zignol, A. M. Cabibbe, A. S. Dean, P. Glaziou, N. Alikhanova, C. Ama, S. Andres, A. Barbova, A. Borbe-Reyes, D. P. Chin, D. M. Cirillo, C. Colvin, A. Dadu, A. Dreyer, M. Driesen, C. Gilpin, R. Hasan, Z. Hasan, S. Hoffner, A. Hussain, N. Ismail, S. M. M. Kamal, F. M. Khanzada, M. Kimerling, T. A. Kohl, M. Mansjö, P. Miotto, Y. D. Mukadi, L. Mvusi, S. Niemann, S. V. Omar, L. Rigouts, M. Schito, I. Sela, M. Seyfaddinova, G. Skenders, A. Skrahina, S. Tahseen, W. A. Wells, A. Zhurilo, K. Weyer, K. Floyd, M. C. Raviglione, Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. *Lancet Infect. Dis.* **18**, 675–683 (2018).

44. K. R. Wollenberg, C. A. Desjardins, A. Zalutskaya, V. Slodovnikova, A. J. Oler, M. Quiñones, T. Abeel, S. B. Chapman, M. Tartakovsky, A. Gabrielian, S. Hoffner, A. Skrahin, B. W. Birren, A. Rosenthal, A. Skrahina, A. M. Earl, Whole-Genome Sequencing of Mycobacterium tuberculosis Provides Insight into the Evolution and Genetic Composition of Drug-Resistant Tuberculosis in Belarus. *J. Clin. Microbiol.* **55**, 457–469 (2017).

45. J. E. Phelan, D. R. Lim, S. Mitarai, P. F. de Sessions, M. A. A. Tujan, L. T. Reyes, I. A. P. Medado, A. G. Palparan, A. N. M. Naim, S. Jie, E. Segubre-Mercado, B. Simoes, S. Campino, J. C. Hafalla, Y. Murase, Y. Morishige, M. L. Hibberd, S. Kato, M. C. G. Ama, T. G. Clark, Mycobacterium tuberculosis whole genome sequencing provides insights into the Manila strain and drug-resistance mutations in the Philippines. *Sci. Rep.* **9**, 9305 (2019).

46. N. D. Hicks, J. Yang, X. Zhang, B. Zhao, Y. H. Grad, L. Liu, X. Ou, Z. Chang, H. Xia, Y. Zhou, S. Wang, J. Dong, L. Sun, Y. Zhu, Y. Zhao, Q. Jin, S. M. Fortune, Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. *Nat Microbiol.* **3**, 1032–1042 (2018).

47. K. Dheda, J. D. Limberis, E. Pietersen, J. Phelan, A. Esmail, M. Lesosky, K. P. Fennelly, J. Te Riele, B. Mastrapa, E. M. Streicher, T. Dolby, A. M. Abdallah, F. Ben-Rached, J. Simpson, L. Smith, T. Gumbo, P. van Helden, F. A. Sirgel, R. McNerney, G. Theron, A. Pain, T. G. Clark, R. M. Warren, Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. *Lancet Respir Med*. **5**, 269–281 (2017).

48. L. Freschi, R. Vargas, A. Husain, S. M. M. Kamal, A. Skrahina, S. Tahseen, N. Ismail, A. Barbova, S. Niemann, D. M. Cirillo, A. S. Dean, M. Zignol, M. R. Farhat, Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. *Nat. Commun.* **12**, 1–11 (2021).

49. R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. **27**, 863–864 (2011).

50. D. E. Wood, S. L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

51. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013), (available at http://arxiv.org/abs/1303.3997).

52. M. I. Gröschel, M. Owens, L. Freschi, R. Vargas Jr, M. G. Marin, J. Phelan, Z. Iqbal, A. Dixit, M. R. Farhat, GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Med.* **13**, 138 (2021).

53. A. Kapopoulou, J. M. Lew, S. T. Cole, The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis* . **91**, 8–13 (2011).

54. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

55. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Others, in *12th ${USENIX} symposium on operating systems design and implementation ({OSDI}$ 16)* (usenix.org, 2016), pp. 265–283.

56. G. Van Rossum, F. L. Drake, *Python 3 Reference Manual: (Python Documentation Manual Part 2)* (CreateSpace Independent Publishing Platform, 2009).

57. J. Nickolls, I. Buck, M. Garland, K. Skadron, Scalable Parallel Programming with CUDA: Is CUDA the parallel programming model that application developers have been waiting for? *Queueing Syst.* **6**, 40–53 (2008).

58. F. Coll, R. McNerney, J. A. Guerra-Assunção, J. R. Glynn, J. Perdigão, M. Viveiros, I. Portugal, A. Pain, N. Martin, T. G. Clark, A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5**, 4812 (2014).

59. L. Freschi, R. Vargas, A. Hussain, S. M. Mostofa Kamal, A. Skrahina, S. Tahseen, N. Ismail, A. Barbova, S. Niemann, D. M. Cirillo, A. S. Dean, M. Zignol, M. R. Farhat,

Population structure, biogeography and transmissibility of Mycobacterium tuberculosis (2020), p. 2020.09.29.293274, , doi:10.1101/2020.09.29.293274.

60. U. H. Manjunatha, S. P. S Rao, R. R. Kondreddi, C. G. Noble, L. R. Camacho, B. H. Tan, S. H. Ng, P. S. Ng, N. L. Ma, S. B. Lakshminarayana, M. Herve, S. W. Barnes, W. Yu, K. Kuhen, F. Blasco, D. Beer, J. R. Walker, P. J. Tonge, R. Glynne, P. W. Smith, T. T. Diagana, Direct inhibitors of InhA are active against Mycobacterium tuberculosis. *Sci. Transl. Med.* **7**, 269ra3 (2015).

61. S. Petrella, N. Gelus-Ziental, A. Maudry, C. Laurans, R. Boudjelloul, W. Sougakoff, Crystal structure of the pyrazinamidase of Mycobacterium tuberculosis: insights into natural and acquired resistance to pyrazinamide. *PLoS One*. **6**, e15785 (2011).