DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Identifying Antibiotic Resistance in Mycobacterium Tuberculosis With Machine Learning: A Quick and Accurate Alternative to Conventional Diagnostics

Citation
Chen, Michael L. 2020. Identifying Antibiotic Resistance in Mycobacterium Tuberculosis With Machine Learning: A Quick and Accurate Alternative to Conventional Diagnostics. Bachelor's thesis, Harvard College.

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364665

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Identifying antibiotic resistance in *Mycobacterium tuberculosis* with machine learning: a quick and accurate alternative to conventional diagnostics

A THESIS PRESENTED
BY
MICHAEL L. CHEN
TO
THE DEPARTMENT OF APPLIED MATHEMATICS

IN PARTIAL FULFILLMENT OF THE HONORS REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECT OF
APPLIED MATHEMATICS

HARVARD COLLEGE
CAMBRIDGE, MASSACHUSETTS
APRIL 2020

©2020 – Michael L. Chen
all rights reserved.

Thesis advisors: Drs. Isaac Kohane and Andrew Beam Michael L. Chen

Identifying antibiotic resistance in *Mycobacterium tuberculosis* with machine learning: a quick and accurate alternative to conventional diagnostics

Abstract

The diagnosis of multidrug resistant and extensively drug resistant tuberculosis is a global health priority. Whole genome sequencing of clinical *Mycobacterium tuberculosis* isolates promises to circumvent the long wait times and limited scope of conventional phenotypic antimicrobial susceptibility testing, but gaps remain in predicting phenotype accurately from genotypic data especially for certain drugs. My primary aim was to implement and explore statistical methods and deep learning algorithms using a rich dataset to build a high performing and fast predicting model to detect anti-tuberculosis drug resistance.

I collected targeted or whole genome sequencing and conventional drug resistance phenotyping data from 3,601 *Mycobacterium tuberculosis* strains enriched for resistance to first- and second-line drugs. I investigated the utility of (1) rare variants and variants known to be determinants of resistance for at least one drug and (2) statistical methods and deep learning architectures in predicting phenotypic drug resistance to 10 anti-tuberculosis drugs. Performance was validated on an independent validation set, as well as compared to a convolutional neural network approach on an expanded set of 10,198 *Mycobacterium tuberculosis* strains.

The highest performing machine and statistical learning methods included both rare variants and those known to be causal of resistance for at least one drug. Both simpler L2 penalized regression and a multidrug wide and deep neural network (MD-WDNN) had high predictive performance. The average AUCs for the highest performing model, the MD-WDNN, were 0.979 for first-line drugs and 0.936 for second-line drugs during repeated cross-validation. On an independent validation set, the highest performing model showed average AUCs, sensitivities, and specificities, respectively, of 0.937, 87.9%, and 92.7% for first-line drugs and 0.891, 82.0% and 90.1% for second-line drugs. The method has higher predictive performance compared to previously reported machine learning models during cross-validation, with higher AUCs for 8 of 10 drugs. High performance remained on the expanded set of 10,198 strains, and the extension to a convolutional neural network approach showed promising results with interpretable saliency map visualizations.

Overall, the machine learning models described in this work significantly improve the accuracy of antibiotic resistance prediction and hold promise in bringing sequencing technologies closer to the bedside.

Contents

List of Figures

List of Tables

I dedicate this work to my family: Lillian, John, Allison, and William for their continual advice, love, and support.

Acknowledgments

First and foremost, my thesis would not have been possible without my amazing research mentors. I am eternally grateful for my thesis advisors, Dr. Isaac Kohane and Dr. Andrew Beam, who took me on as a first-year undergraduate student. I am incredibly appreciative of Dr. Maha Farhat, who has been a fervent supporter of my research endeavors from the very start. I am thankful for the computational foundation and unwavering guidance that you have provided me throughout this project spanning four years. I cannot even begin to imagine what this research experience would have been like without the three of you. Thank you for all that you do.

I also would like to thank the members of the Kohane, Beam, and Farhat labs who have helped me in many capacities. I thank the reviewers and editorial team of *EBioMedicine*, who have provided feedback and shaped my work in the journal article in which some of this thesis work has been published. A big shout-out to my friends who have seen me throughout many stages of this work. Finally, to my family: words cannot express my gratitude. Thank you.

1

Introduction

1.1 Motivation

Tuberculosis remains a global health threat as the tenth leading cause of death worldwide in 2018[1]. Antibiotics serve as the main method for treating tuberculosis, and the use of antibiotics within healthcare has grown significantly since 2005 and is projected to continue to grow dramatically over the next 10 years[2]. The development of antibiotic resistance to one or multiple drugs poses a significant barrier to providing effective care to patients with tuberculosis. Clinical isolates of tuberculosis that are rifampicin-resistant or multidrug-resistant, defined as being resistant to rifampicin and isoniazid (the two leading antibiotics used to treat tuberculosis), account for 18% of the current tuberculosis cases and 3.4% of the new tuberculosis cases in 2018[1]. Furthermore, clinical tuberculosis isolates that are extensively drug-resistant, defined as being resistant to rifampicin, isoniazid, one second-line injectable drug, and one fluoroquinolone, account for approximately 6.2% of the multidrug-resistant cases[1]. The rates of favorable treatment outcome for tuberculosis patients with multidrug-resistant or extensively drug-resistant strains are significantly lower at 56% and 39%, respectively, because of the challenges in treating these strains clinically[1]. The World Health Organization has deemed multidrug-resistant tuberculosis as a global public health crisis[1].

The accessibility of antibiotic resistance detection for tuberculosis is one major barrier to treating tuberculosis patients effectively. The current gold standard for identifying antibiotic resistance is through culture-based antimicrobial susceptibility testing[1]. There are three major issues with reliance on culture-based testing. First, the laboratory resources required for culture-based testing are available in fewer than half of the countries that have a high burden of multidrug-resistant tuberculosis[3]. Second, culture-based testing can take up to 12 weeks before the susceptibility results are available due to the slow growth of *Mycobacterium tuberculosis*[4]. Third, conducting the cultures is a biohazard, potentially resulting in laboratory personnel and health care workers getting infected by tuberculosis themselves. The ratio of healthcare workers to the general adult population who reported getting infected by tuberculosis is high for many low-income countries, reported at ratios of between 2 and 6, indicating a need for safer alternatives to laboratory testing[1].

Molecular diagnostic tests are one alternative to conventional cultures, but the narrow scope of molecular diagnostics presents a number of key challenges in effectively identifying antibiotic resistance. First, the molecular tests often rely on one or few genetic loci, which results in a low sensitivity for these tests[5]. Second, molecular tests rely on common loci, but there has been evidence that rare loci can contribute to antibiotic resistance. For example, the acquisition of resistance by rare loci has been shown for pyrazinamide, one of the four first-line antibiotics used in treating tuberculosis[6]. Third, approved tubercular molecular tests by the World Health Organization are limited to five antibiotics, which does not include two key first-line agents that are used in the standard starting regimen to treat tuberculosis. These molecular tests also omit 5 other antibiotics that are used to treat tuberculosis. Fourth, molecular tests do not process information from gene-gene interactions despite allelic exchange experiments illustrating that gene-gene interactions contribute to resistance in rifampicin, ethambutol, and fluoroquinolones[7].

The development of whole genome sequencing has shown potential as a robust alternative to culture-based and molecular testing. Whole genome sequencing information for tuberculosis is becoming increasingly accessible, with sequencing technologies like MinION serving as a portable and affordable means for sequencing[8]. Whole genome sequencing captures common and rare variants within the tubercular genome, thus broadening the genetic scope of the test. One past study has used the whole genome sequencing of tuberculosis combined with an analytical direct association approach[9]. This simple method of finding direct correlations between a genetic locus and antibiotic resistance for multiple drugs showed good performance for rifampicin and isoniazid due the fact that resistance to these drugs is conferred through the large effect of few variants. However, the performance in other first-line drugs, second-line injectables, and fluoroquinolones was considerably lower, especially because many tuberculosis isolates contained "indeterminate" variants that were not within the set of variants used for training the direct association model. In addition, phenotypic information is more sparse for second-line drugs and fluoroquinolones, making this direct association approach difficult for drugs with limited data. The inverse relationship between order of resistance (i.e. number of antibiotics to which one tuberculosis isolate is resistant) and predictive performance of the direct association model illustrates the need for a more sophisticated approach to combat multidrug-resistant and extensively drug-resistant tuberculosis.

1.2 Approach

I hypothesize that the limited predictive performance with the direct association method can be improved by using a large dataset containing many multidrug-resistant strains. I aim to build a model that will improve predictive performance for a set of 11 antibiotics that are currently used to treat tuberculosis. The 11 antibiotics include the four first-line drugs (rifampicin, isoniazid, ethambutol, and pyrazinamide), streptomycin, three second-line injectable drugs (capreomycin, amikacin, and kanamycin), and three fluoroquinolones (ciprofloxacin, moxifloxacin, and ofloxacin).

In Chapter 2, I describe the data collection and data processing methods used in further analyses. I describe two main datasets: the variant-style dataset (used in Chapters 3, 4, and 5), and the sequence-style dataset (used in Chapter 5).

In Chapter 3, I describe the process of building a computational model on the variant-style dataset to predict antibiotic resistance to a panel of 11 anti-tubercular drugs. Specifically, I address the following key points in accurately identifying antibiotic resistance: 1) incorporate an expanded set of variants compared to molecular diagnostic tests, 2) incorporate the effect of rare variants that may appear in few isolates, 3) learn additive effects and gene-gene interactions that contribute to antibiotic resistance, and 4) allow for a shared architecture amongst antibiotics through a multi-task structure, allowing drugs with limited phenotypic data to "learn" even for tuberculosis isolates where the resistance phenotype is unavailable. I present an ablation analysis, in which I vary one of the dimensions above and keep the other dimensions constant, for the dimensions above. The ablation analysis is performed using sequencing data processed into the form of tabular variant data. I also investigate the effect of different statistical and deep learning models throughout the ablation analysis.

In Chapter 4, I present a validation of the highest performing models on an independent validation set. I interpret which genetic variants within the chosen model are the most important for predictive performance and visualize the model representation of antibiotic resistance.

In Chapter 5, I present an alternative approach and the justification for this approach based on analyzing the nucleotide sequences of reconstructed tuberculosis genome segments. This analysis, which is done on an expanded dataset of tuberculosis isolates, is compared to the previously chosen model in Chapter 4 re-trained on the expanded dataset. I also present an interpretation of the model and its results through saliency maps.

In summary, the following is an assessment of statistical methods and deep learning models with the goal of building a quick, accessible, and clinically-relevant algorithm for detecting antibiotic resistance from targeted and whole genome sequencing tuberculosis data.

2

Data

2.1 Variant-style data

The variant-style dataset is used for the modeling and further analyses conducted in Chapters 3, 4, and 5.

2.1.1 Sequencing collection

The total training set used for the analysis with variant data included 3,601 tuberculosis isolates. Of the 3,601 isolates, 1,379 tuberculosis isolates were sequenced using molecular inversion probes that targeted 28 genes within the tubercular genome that are known to confer resistance to at least one antibiotic. The gene list, descriptions of each gene, drug resistance association(s), and coordinates for each gene are available in Appendix Table A.8. For each of the 28 genes, the entire gene and 100 base-pairs flanking the gene were sequenced. The remaining 2,222 isolates had whole genome sequencing information for the tubercular genome. The sequencing data was made available by the ReSeqTB data platform, which provides genetic and phenotypic information based on WHO-endorsed assays [10]. For further analyses with this dataset, the variants used to predict antibiotic resistance were limited to 32 candidate genes: 28 that were available in targeted sequencing, and 4 that were not available but were previously deemed to be causative of resistance. Variants in these regions, *eis* and *rpsA*, were added into the set of predictors. For those isolates with missing genotype data, I coded an intermediate status of 0.5 for the missing variant information.

The independent validation set used to validate performance accuracy of the highest performing models included 792 tuberculosis isolates. The data did not contain any overlap with isolates from the training set, and all the mutational features of the validation set were unique from the training set. The data for these isolates were curated from distinct sources from the following references: Lieberman et al. (2016), Chatterjee et al. (2017), Gardy et al. (2011), and Zhang et al. (2013) [11,12,13,14].

The sequencing information was converted to a tabular list of variants using a custom bioinformatics pipeline to analyze the raw filter reads.

2.1.2 Determining predictors

I split the predictors used in the analysis into three categories for the purposes of ablation analyses, in which I evaluated performance for three different subsets of data. Data Category 1 (the smallest subset) included predictors that were previously implicated in the acquisition of resistance for that particular drug. For example, the antibiotic pyrazinamide has only two genes associated with resistance (*rpsA* and *pncA*), so only variants associated with these genes are selected. Data Category 2 included predictors that were implicated in the acquisition of resistance for any of the 11 drugs tested. Data Category 3 includes all variants within the second category but also includes "derived" features, in which I fold "rare" variants into higher-level categories based on the nature of the variant. The process for creating these derived categories is described below.

All candidate variants to be used as predictors included single nucleotide polymorphisms (SNPs), insertions, and deletions within the 32 genes' promoters, intergenic, and coding regions. Within the 3,601 tuberculosis isolates, there were 6,342 unique variants that fit the criteria. In order to distinguish between "rare" and "common" mutations for the purposes of ablation, I defined "common" mutations as those variants that are present in at least 30 of the 3,601 isolates within the training dataset. Of the 6,342 variants, 166 variants were deemed "common" mutations. Of the remaining "rare" mutations, I grouped the mutations by the gene locus (coding, intergenic, or promoter regions). For the coding region variants, I further split the group by mutation type (SNP, frameshift insertion or deletion, and non-frameshift insertion or deletion). For the non-coding region variants, I split the groups into SNPs and insertions/deletions. The final "derived" categories that were contained in at least 30 isolates within the training set included 56 predictors. I combined these 56 predictors with the "common" mutations of Data Category 2 into the final largest predictor set, Data Category 3, which included a final total of 222 predictors (166 "common" mutations and 56 "derived" categories).

In summary, Data Category 1 includes a number of predictors that depends on the antibiotic (and its associated resistance-determining genes). Data Category 2 includes 166 "common" mutations, and Data Category 3 includes 222 "common" and "derived" predictors.

2.1.3 Phenotype data

All the resistance phenotype information used in modeling was curated from culture-based, WHO-approved antibiotic susceptibility testing to at least two antibiotics. The eleven antibiotics of interest used to treat tuberculosis are rifampicin, isoniazid, ethambutol, pyrazinamide, streptomycin, amikacin, capreomycin, kanamycin, ciprofloxacin, moxifloxacin, and ofloxacin. For each isolate-drug pair, the phenotype was one of three options: resistant (coded as 0), susceptible (coded as 1), or unavailable (coded as -1). The performance and phenotypic data is not reported for ciprofloxacin because of limited susceptibility testing within the validation set, thus limiting generalizability in findings. All results are reported on the remainin 10 anti-tuberculosis drugs.

Resistance to isoniazid, one of the first-line drugs, was tested in the highest proportion of the training set at 3,564 isolates. Rifampicin resistance phenotyping was done in 3,542 isolates, and all antibiotics except for ofloxacin had resistance phenotyping in at least 1,204 isolates. Ofloxacin had the lowest number of isolates with susceptibility testing at 739 isolates. The final training dataset that I used was enriched for drug resistance, in which a proportion of between 19.9% and 47.0% of isolates were resistant for each of the 11 drugs. The number of resistant and susceptible isolates per drug is available in Appendix Table A.1.

Within the validation set of 792 isolates, ten of the drugs had between 198 and 736 isolates with available phenotypic data. Ciprofloxacin only had 2 isolates with phenotyping available, so the predictive performance could not be validated on ciprofloxacin. Thus, predictive performance for ciprofloxacin resistance is not reported throughout. The phenotypic information for the independent validation set is available in Appendix Table A.2.

2.2 Sequence-style data

The sequence-style dataset is used for the modeling and further analyses conducted in Chapter 5. This dataset is an expanded dataset compared to the variant-style data and is used for testing a different type of model, a convolutional neural network, that can incorporate spatial information of the tuberculosis genome.

2.2.1 Sequencing collection

The dataset included 10,198 tuberculosis isolates. All of the isolates underwent whole genome sequencing for the full tubercular genome. The data was compiled from a number of public sources, including PATRIC[15] and the ReSeqTB platform[10]. The models trained on the sequence-style data were evaluated using the validation fold performance from cross-validation.

2.2.2 Determining predictors

I created subsets of smaller sequences from the whole genome sequences based on the regions of interest for predicting antibiotic resistance for four antibiotics: two first-line drugs (rifampicin and pyrazinamide) and two second-line injectables (capreomycin and kanamycin). For each drug, I used sequencing regions that included the intergenic, regulatory, and coding regions of the genes of interest. For rifampicin, I used regions of the *rpoB-rpoC genes*; for capreomycin, I used regions of the *rrs-rrl* and *tlyA* genes; for kanamycin, I used regions of the *eis* and *rrs-rrl* genes; and for pyrazinamide, I used regions of the *pncA* and *rpsA* genes. For antibiotics with more than one locus, the two regions were concatenated together along the same dimension. The final length of the input sequences were as follows: rifampicin (7,816 base pairs), pyrazinamide (3,565 base pairs), capreomycin (6,646 base pairs), and kanamycin (8,212 base pairs).

Because the convolutional neural network approach requires sequences of fixed length, the sequences were formatted by inserting a null character to denote insertions and deletions relative to the longest strain sequence with all insertions. For the insertions within a strain, all other strains with the dataset were coded to have null characters in the positions of insertions. For deletions within a strain, that particular strain was coded with null characters in the position of deletions. Thus, each position in the sequence contained one of 5 characters: A, C, T, G, (the four nitrogenous bases), or the null character. The final isolates' genotypes were represented as a fixed length one-hot encoded vector, $N x M x 5$, where N represents the number of isolates, M represents the number of base pairs within the genetic regions of interest, and 5 represents the encoding of the four bases and the null character.

All isolates used in the sequence-style analysis for the convolutional neural network approach were also formatted into the variant-style data described in the section above. This data was used for a comparison of the convolutional neural network to the prior wide and deep neural network re-fit through cross-validation on the larger set of data. This enables a direct comparison of the two approaches (variant-style and sequence-style) within the same dataset of tuberculosis isolates.

2.2.3 Phenotype data

Likewise to the phenotypic data for the variant-style dataset, the resistance phenotype information is available from antibiotic susceptibility testing. For the sequence-style genotype analysis, the four antibiotics tested are rifampicin, pyrazinamide, capreomycin, and kanamycin. All phenotypic information is resistant, susceptible, or not available.

Of the four antibiotics, rifampicin had resistance information available in the highest percentage of the 10,198 isolates (97.1%). Kanamycin had the smallest amount of resistance information available at 32.3% of the isolates. The four drugs had between 7.2% and 34% of isolates resistant. The number of susceptible and resistant isolates for each drug is available in Appendix Table A.3.

3

MD-WDNN and Variant-Style Analysis

3.1 *M. tuberculosis* lineage diversity

The geographic diversity within the dataset is an important consideration to assess the generalizability of the findings to new and distinct populations of *Mycobacterium tuberculosis*. To evaluate the geographic diversity within the 3,601 isolates, I used hierarchical clustering to determine the main lineages within the dataset.

Based on a previous study by Walker et al. (2015)[9], I identified a total of 33 variants that determine the lineages of each isolate. The list of variants is available in Appendix Table A.4. Each isolate was represented as a single vector each of length 33 (one entry for each lineage-defining variant). Each entry was one of two options: 0 or 1, depending on the absence or presence of that variant within a particular tuberculosis isolate.

I conducted the hierarchical clustering process as follows. The dataset originally had dimensions of *N x L*, where *N* represents the number of isolates and *L* is 33, representing the number of lineage-defining variants. I computed the genetic-lineage similarity between each pair of isolates with a Euclidean distance metric, resulting in an *N x N* distance matrix. On the distance matrix, I applied Ward's method of hierarchical clustering, which joins clusters using a bottom-up approach based on the within-cluster squared sum of distances. The final groupings were mapped back to the recognized tuberculosis lineage classifications by matching the expected mutational patterns from the previous study in Walker et al. (2015)[9].

Figure 3.1 shows the results of the lineage analysis. The distance matrix is represented as a heatmap, where the darker colors correspond to a closer relationship between the isolates. The isolates are grouped into five well-defined clusters, which correspond to the following *Mycobacterium tuberculosis* lineages: East Asian (Green), Euro-American with Latin America Mediterranean sub-lineage (Purple), Euro-American with other sub-lineages (Orange), Central Asian (Yellow), and Indo-Oceanic & *M. africanum* (Blue). All 5 lineages were well represented: 632 isolates were from the Euro-American Latin America Mediterranean sub-lineage, 1501 from other Euro-American sub-lineages, 331 from the Indo-Oceanic or Mycobacterium africanum, 643 from the Central Asian lineage, and 494 from the East Asian lineage.

Compared with the training data, the independent validation dataset was geographically distinct and contained a higher proportion of East Asian lineage, 351 isolates (44%), but a lower proportion

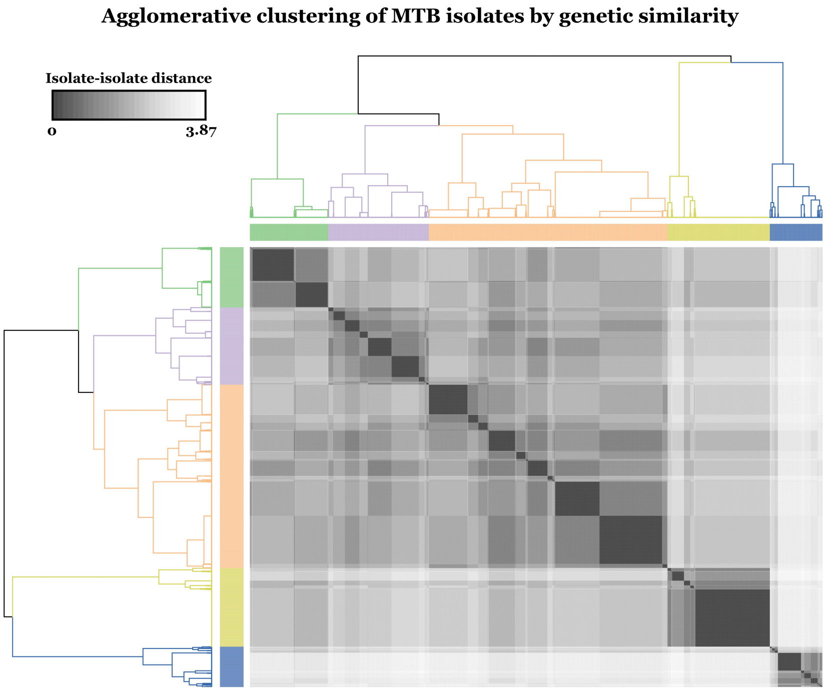


Figure 3.1: Agglomerative clustering and lineage diversity heatmap of 3,601 MTB isolates by genetic similarity. Using the Euclidean distances between isolates represented by their lineage-defining variants, I used hierarchical clustering to construct a dendrogram. The 5 clusters correspond to 5 known lineages: East Asian (Green), Euro-American with Latin America Mediterranean sub-lineage (Purple), Euro-American with other sub-lineages (Orange), Central Asian (Yellow), and Indo-Oceanic & *M. africanum* (Blue).

of other lineages. The Euro-American Latin America Mediterranean sub-lineage had 63 isolates, other Euro-American lineages had 253 isolates, the Central Asian lineage had 32 isolates, and all other lineages had 93 isolates.

3.2 Model training process

Towards the goal of building a comprehensive diagnostic tool to identify multidrug-resistant tuberculosis with respect to a large panel of antibiotics, I assessed a number of machine learning models trained on different feature sets as described in the following sections. To assess each model, I performed ten-fold cross-validation within the training set, repeated 5 times, with performance reported on the validation fold. The average over the 50 different validation folds was reported as the cross-validation performance. All single drug (single-task) models were stratified by class label to match the class imbalances within each drugs' phenotype distribution.

The main performance metric to evaluate each model during the cross-validation procedure was the area under the receiver-operator curve (AUC) with 95% confidence intervals based on the 50 validation folds. Due to the class-imbalance for some of the antibiotics within the dataset, I also measured and reported the average precision (AP) score, which is a summary metric for the precision-recall curve.

I trained all models on a NVIDIA GeForce GTX Titan X graphics processing unit (GPU).

3.3 MD-WDNN architecture

Neural networks have shown promise as a predictive tool in several areas of biology and biomedicine. In the current analysis, I evaluated several neural network architectures, which includes variants of the wide and deep neural network (WDNN) model[16]. The multitask (multidrug) wide and deep neural network (MD-WDNN) is the most complex model within the set of WDNN models tested.

A schematic representation of the MD-WDNN architecture is shown in Figure 3.2. There are a few mathematical features of the MD-WDNN model that suggest it would be a good fit for antibiotic resistance identification. First, the MD-WDNN is a multitask model that predicts antibiotic resistance for all 11 drugs simultaneously. The multitask feature allows drugs with less phenotypic data to borrow pathway information from other drugs with a higher amount of phenotyped isolates. The network is thus able to learn from isolates even for which phenotypic information is not available for a particular drug. Each of the 11 nodes in the final layer represented one drug (top of the schematic) and its output value for each node was the probability that the MTB isolate was resistant to the corresponding drug.

Second, I used a "wide and deep" approach, which combines two models: logistic regression and a deep multilayer perceptron. Logistic regression defines the "log odds" as the linear combination of predictors, X_n, as below, where N represents 222, the number of variants, and w_n represents the weights corresponding to each of the predictors:

$$\ln(\frac{\hat{p}}{1-\hat{p}}) = w_0 + \sum_{n=1}^{N} w_n * X_n$$

Within the context of genomic data, the "wide" logistic regression model can be thought of as modeling the additive portion of the genotype-phenotype relationship.

The "deep" multilayer perceptron portion can capture non-linear relationships within the input genetic variants. The multilayer peceptron contains 3 hidden layers that are connected sequentially via two main operations, described below. W_n represents the weights of the nth layer, X_n represents the values of the nth layer, b_n is the bias, and z is an intermediate matrix:

$$z_{n+1} = W_{n+1}X_n + b_{n+1}$$

$$X_{n+1} = \text{ReLU}(z_{n+1})$$

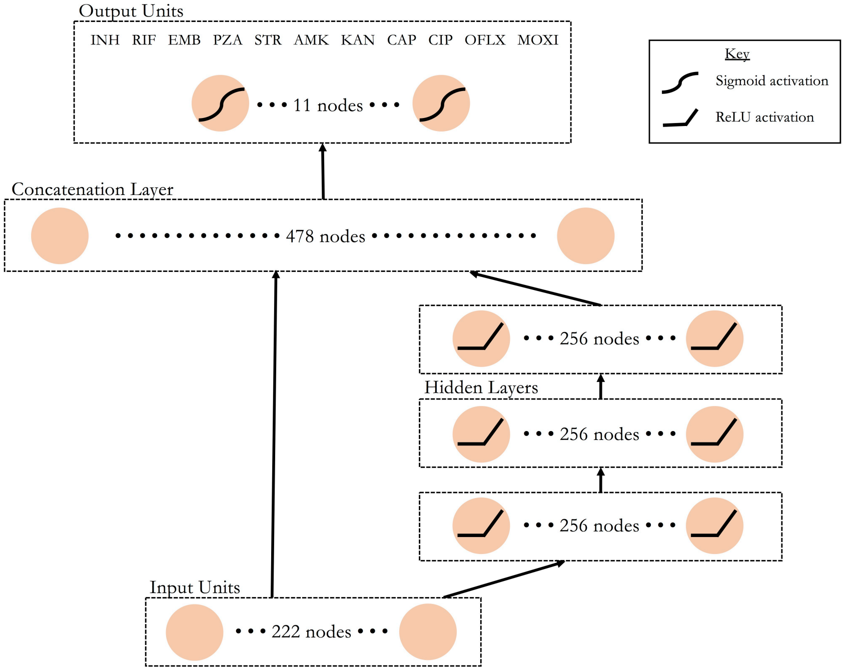


Figure 3.2: Schematic of the multidrug wide and deep neural network architecture. Data flows from bottom to top through the wide (left) and deep (right) paths of the neural network. Nonlinear transformations, where applied, are depicted on the corresponding nodes. Each of the 11 nodes in the output layer represents resistance status predictions in all MTB isolates for one of the 11 anti-tuberculosis drugs.

The first equation is analogous to logistic regression without the final sigmoid activation, as nodes in the subsequent hidden layer are a linear combination of nodes from the previous layer. However, the second equation involves a non-linear activation function called a rectified linear unit (ReLU), allowing for non-linear relationships to be modeled within the multilayer perceptron. Within genomic data, the "deep" portion can thus capture epistatic effects, which are thought to be important in the acquisition of antibiotic resistance[7]. The "wide" and "deep" portions of the MD-WDNN are trained simultaneously and merged in a final classification layer before the output prediction of antibiotic resistance, allowing the MD-WDNN to leverage both the logistic regression and multilayer perceptron approaches.

Third, I built a custom loss function variant of traditional binary cross-entropy. Because there were some isolates without full phenotypic information for the full panel of drugs, the loss function does not penalize the model during training for isolate-drug pairs without resistance information. Due to imbalance between the susceptible and resistant classes within each drug, I adjusted the loss function to upweight the sparser class according to the susceptible-resistant ratio within each drug. The final loss function was a class-weight binary cross entropy that masked outputs where the resistance status was missing.

The procedure for calculating the loss is outlined below, where α is the proportion of isolates resistant to a particular drug, y_{true} is the true phenotypic label (0 = resistant and 1 = susceptible), and y_{pred} is the prediction probability by the network (a probability closer to 1 means predicting susceptible).

First, set the loss equal to zero for any isolate i and drug d pair that does not have known resistance status:

$$\text{Loss}^{id} = 0, \qquad \text{for unknown } y_{true}^{id}.$$

Second, calculate the following loss for each of the remaining isolate-drug pairs:

$$\text{Loss}^{id} = -\alpha^{d}(y^{id}_{true})\log(y^{id}_{pred}) - (1-\alpha^{d})(1-y^{id}_{true})\log(1-y^{id}_{pred})$$

Third, sum over the drugs to get the loss associated with each isolate (or batch of isolates):

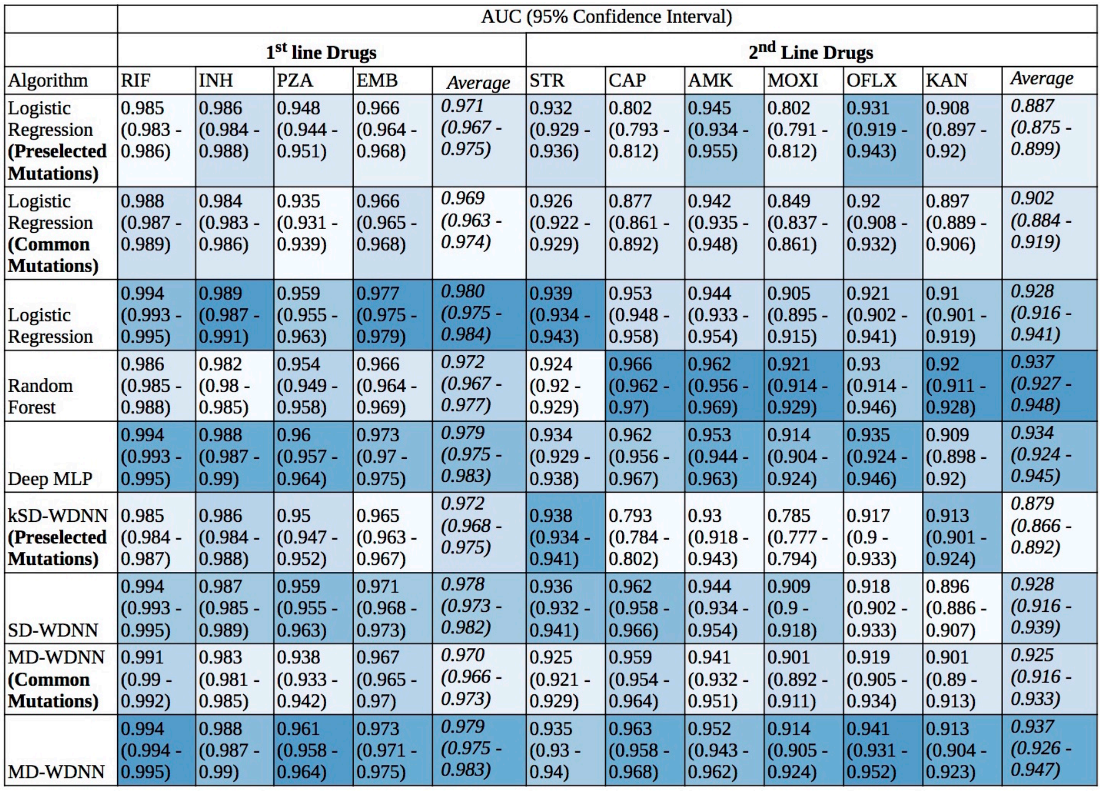
$$\text{Loss}_{i} = \sum_{d} Loss_{id}$$

The final value is used to calculate the final loss and is used to update network weights through backpropagation.

I determined the hyperparameters for the MD-WDNN using a Bayesian optimization routine as implemented by Spearmint and described in Snoek et al. (2012)[17]. The final MD-WDNN had three hidden layers each with 256 rectified linear units (ReLU)[18], dropout[19], batch normalization[20], and L2 regularization. Dropout and L2 regularization are used to prevent overfitting of the models to the training data. I applied L2 regularization to the wide model, the hidden layers of the deep model, and the output sigmoid layer. I trained the network via stochastic gradient descent using the Adam optimizer for 100 epochs with randomized initial starting weights as determined by the Xavier uniform initialization scheme. All hyperparameters are available in Appendix Table A.5.

3.4 Architecture ablation analysis and baseline models

In order to investigate the empirical performance of the MD-WDNN, I performed a comparative ablation analysis following the model training procedure outlined above. The ablation analysis included the following models: the MD-WDNN, a single-task (single drug) wide and deep neural network (SD-WDNN), and a deep multilayer perceptron (MLP). The SD-WDNN has the same architecture as the MD-WDNN with the only difference being that resistance is predicted to one



	AUC (95% Confidence Interval)											
	1^{st} line Drugs					**2^{nd} Line Drugs**						
Algorithm	RIF	INH	PZA	EMB	*Average*	STR	CAP	AMK	MOXI	OFLX	KAN	*Average*
Logistic Regression **(Preselected Mutations)**	0.985 (0.983 - 0.986)	0.986 (0.984 - 0.988)	0.948 (0.944 - 0.951)	0.966 (0.964 - 0.968)	*0.971 (0.967 - 0.975)*	0.932 (0.929 - 0.936)	0.802 (0.793 - 0.812)	0.945 (0.934 - 0.955)	0.802 (0.791 - 0.812)	0.931 (0.919 - 0.943)	0.908 (0.897 - 0.92)	*0.887 (0.875 - 0.899)*
Logistic Regression **(Common Mutations)**	0.988 (0.987 - 0.989)	0.984 (0.983 - 0.986)	0.935 (0.931 - 0.939)	0.966 (0.965 - 0.968)	*0.969 (0.963 - 0.974)*	0.926 (0.922 - 0.929)	0.877 (0.861 - 0.892)	0.942 (0.935 - 0.948)	0.849 (0.837 - 0.861)	0.92 (0.908 - 0.932)	0.897 (0.889 - 0.906)	*0.902 (0.884 - 0.919)*
Logistic Regression	0.994 (0.993 - 0.995)	0.989 (0.987 - 0.991)	0.959 (0.955 - 0.963)	0.977 (0.975 - 0.979)	*0.980 (0.975 - 0.984)*	0.939 (0.934 - 0.943)	0.953 (0.948 - 0.958)	0.944 (0.933 - 0.954)	0.905 (0.895 - 0.915)	0.921 (0.902 - 0.941)	0.91 (0.901 - 0.919)	*0.928 (0.916 - 0.941)*
Random Forest	0.986 (0.985 - 0.988)	0.982 (0.98 - 0.985)	0.954 (0.949 - 0.958)	0.966 (0.964 - 0.969)	*0.972 (0.967 - 0.977)*	0.924 (0.92 - 0.929)	0.966 (0.962 - 0.97)	0.962 (0.956 - 0.969)	0.921 (0.914 - 0.929)	0.93 (0.914 - 0.946)	0.92 (0.911 - 0.928)	*0.937 (0.927 - 0.948)*
Deep MLP	0.994 (0.993 - 0.995)	0.988 (0.987 - 0.99)	0.96 (0.957 - 0.964)	0.973 (0.97 - 0.975)	*0.979 (0.975 - 0.983)*	0.934 (0.929 - 0.938)	0.962 (0.956 - 0.967)	0.953 (0.944 - 0.963)	0.914 (0.904 - 0.924)	0.935 (0.924 - 0.946)	0.909 (0.898 - 0.92)	*0.934 (0.924 - 0.945)*
kSD-WDNN **(Preselected Mutations)**	0.985 (0.984 - 0.987)	0.986 (0.984 - 0.988)	0.95 (0.947 - 0.952)	0.965 (0.963 - 0.967)	*0.972 (0.968 - 0.975)*	0.938 (0.934 - 0.941)	0.793 (0.784 - 0.802)	0.93 (0.918 - 0.943)	0.785 (0.777 - 0.794)	0.917 (0.9 - 0.933)	0.913 (0.901 - 0.924)	*0.879 (0.866 - 0.892)*
SD-WDNN	0.994 (0.993 - 0.995)	0.987 (0.985 - 0.989)	0.959 (0.955 - 0.963)	0.971 (0.968 - 0.973)	*0.978 (0.973 - 0.982)*	0.936 (0.932 - 0.941)	0.962 (0.958 - 0.966)	0.944 (0.934 - 0.954)	0.909 (0.9 - 0.918)	0.918 (0.902 - 0.933)	0.896 (0.886 - 0.907)	*0.928 (0.916 - 0.939)*
MD-WDNN **(Common Mutations)**	0.991 (0.99 - 0.992)	0.983 (0.981 - 0.985)	0.938 (0.933 - 0.942)	0.967 (0.965 - 0.97)	*0.970 (0.966 - 0.973)*	0.925 (0.921 - 0.929)	0.959 (0.954 - 0.964)	0.941 (0.932 - 0.951)	0.901 (0.892 - 0.911)	0.919 (0.905 - 0.934)	0.901 (0.89 - 0.913)	*0.925 (0.916 - 0.933)*
MD-WDNN	0.994 (0.994 - 0.995)	0.988 (0.987 - 0.99)	0.961 (0.958 - 0.964)	0.973 (0.971 - 0.975)	*0.979 (0.975 - 0.983)*	0.935 (0.93 - 0.94)	0.963 (0.958 - 0.968)	0.952 (0.943 - 0.962)	0.914 (0.905 - 0.924)	0.941 (0.931 - 0.952)	0.913 (0.904 - 0.923)	*0.937 (0.926 - 0.947)*

Table 3.1: Tuberculosis drug resistance prediction AUROC performance of the models examined using repeated cross-validation. A table of predictive performance across all nine models during repeated cross-validation. The MD-WDNN, SD-WDNN, deep MLP, random forest, and logistic regression models were trained on the full set of predictors. The MD-WDNN (Common Mutations) and logistic regression (Common Mutations) models were trained on mutations not including the derived categories. The kSD-WDNN (Preselected mutations) and logistic regression (Preselected mutations) models were trained on preselected mutations known to be determinants of resistance for each drug. Performance is shown in average AUC and 95% confidence interval across all cross-validation folds. The cells are colored by rank of the model for each drug, colored from lightest to darkest corresponding with lowest to highest AUC value.

antibiotic at a time rather than in a multitask structure. The MLP matches the MD-WDNN except that it does not contain the "wide" logistic regression portion of the model. I included other simpler models for comparison, such as a single drug L2-regularized logistic regression and a single drug random forest classifier. For this portion of the analysis, all models were trained on the full predictor set of features of "common" variants and derived categories of "rare" variants.

The results are shown in Table 3.1. I found the performances across the different neural net model architectures were not significantly different in the data when trained on the full feature set (Table 3.1). The random forest model had inferior performance to either L2 regularized logistic regression or any of the neural net models for three of the four first line drugs, and as a result, I did not examine this model further. The most complex neural net model, the MD-WDNN, showed the highest average performance across both first and second line drugs with an AUC of 0.953, and the highest performing simple model, L2 regularized logistic regression, showed only slightly lower performance, with an average AUC of 0.949.

To directly compare the effect of building a single model for all drugs vs. individual models for each drug (e.g. multi-task vs single-task), I compared the performance of the SD-WDNN to the MD-WDNN. The predictive performance of the MD-WDNN and the SD-WDNN during repeated cross-validation are shown in Figure 3.3. The average AUC for the SD-WDNN was 0.978 for first-line drugs and 0.928 for second-line drugs; the multidrug architecture of the MD-WDNN resulted in a higher average AUC for both first-line drugs (AUC = 0.979) and second-line drugs (AUC = 0.936), although these differences were not significant. The largest gains were observed for the drugs kanamycin and ofloxacin, with AUC differences of 0.023 and 0.017, respectively.

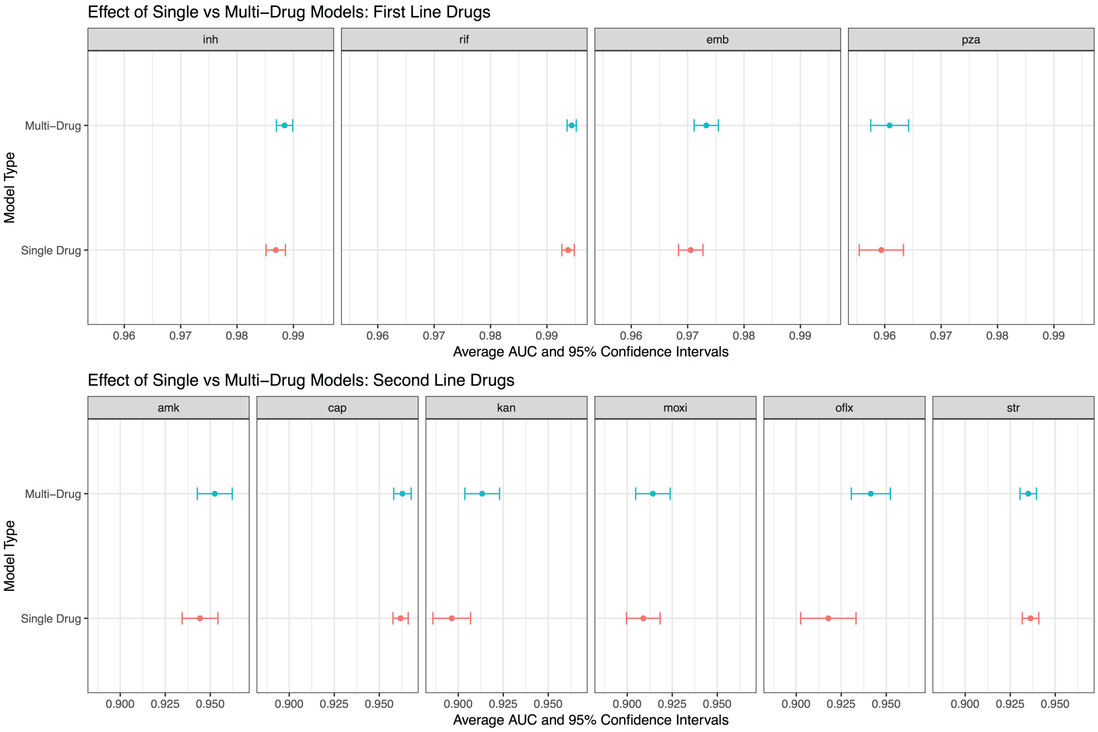


Figure 3.3: Comparison of tuberculosis drug resistance predictive performance between single drug and multidrug models. Area under the ROC curve classification performance and 95% confidence intervals during repeated cross-validation for the MD-WDNN predicting resistance for all drugs simultaneously and for the SD-WDNN.

3.5 Effect of rare, common, and drug-specific variants on performance

I investigated the effect of different feature sets (Data Category 1, 2, and 3) on the two highest-performing models selected from the previous section, i.e. L2-regularized logistic regression and the MD-WDNN. As described in section 2.1, Data Category 1 includes variants implicated in resistance only for the particular drug at hand, which requires a different predictor set for each antibiotic. A multidrug architecture cannot be used when training on Data Category 1, so I use a single-drug WDNN, which I term the kSD-WDNN (short for known-mutation SD-WDNN). For Data Category 2 and 3, which have "common" mutations and "common and derived" mutations, respectively, the multidrug architecture can be used, so the MD-WDNN is used in this analysis. L2-regularized logistic regression is a single drug classifier, so the same classification structure is used across all three Data Categories.

Figure 3.4 shows the performance results on the three feature sets. The largest step up in AUC across any of the models and feature sets was observed between the models trained using genetic regions known to be causative of resistance for each particular drug and the models trained on the full predictor set of variants known to be determinants of resistance to at least one drug (Figure 3.4, Table 3.1). For the second-line drugs, the average AUC was 0.887 for L2 regularized logistic regression using the preselected variants vs. 0.929 for L2 regression using the full predictor set. The importance of using rare genetic variation in predicting resistance is highlighted by the loss of performance seen with the WDNN or L2-regularized logistic regression built without the derived variables (Figure 3.4). This performance gap was most notable for the drugs pyrazinamide, capreomycin, and moxifloxacin.

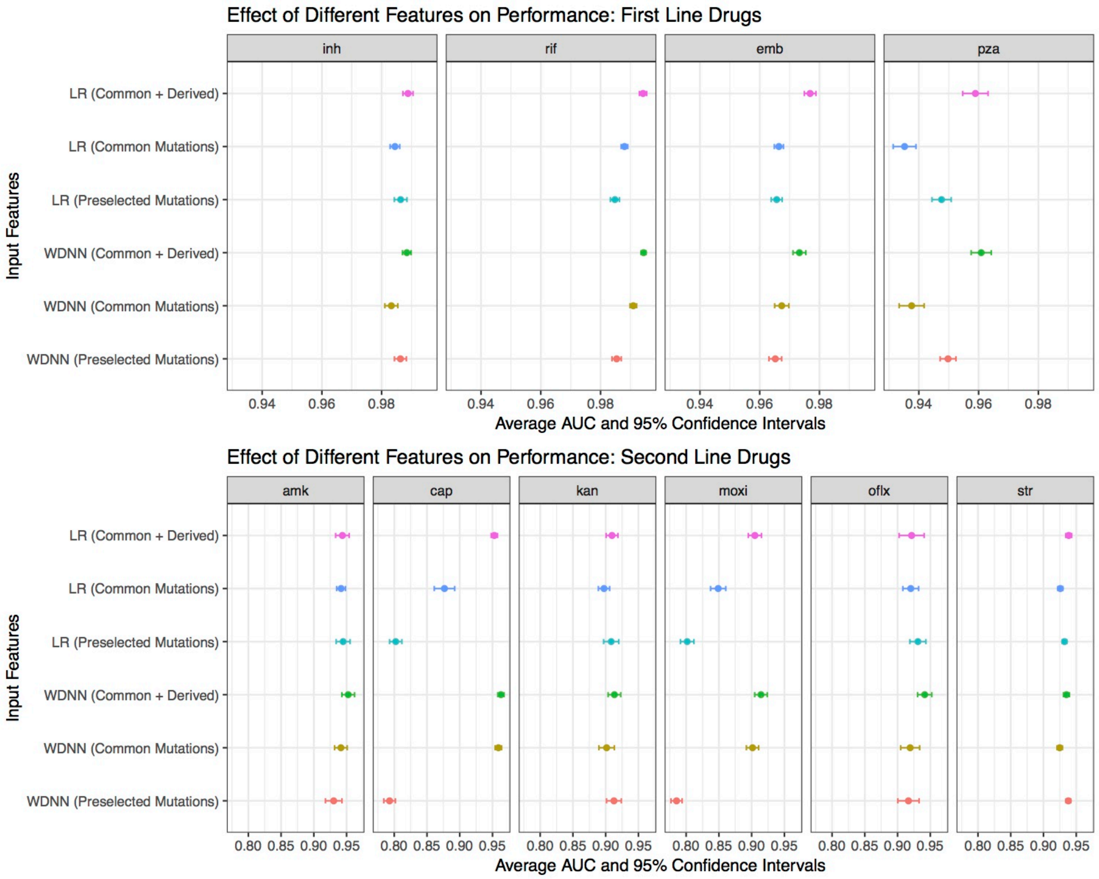


Figure 3.4: Comparison of tuberculosis drug resistance predictive performance based on input feature set. Area under the ROC curve classification performance and 95% confidence intervals during repeated cross-validation is reported. The WDNN and logistic regression models are trained on all features (common and derived mutations), on just the common mutations, and on variants occurring in genes known to be resistant determinants for each drug (Preselected Mutations).

3.6 Overall comparison

I demonstrated here that the highest-performing models included the MD-WDNN and L2-regularized logistic regression trained on the full feature set of "common" and "derived" mutation categories. The highest performing model was the MD-WDNN with an average AUC of 0.953 across the full panel of antibiotics, and L2-regularized logistic regression had an average AUC of 0.949, which was not significantly different from the MD-WDNN's performance. Both models exceeded the performance of the previously published direct-association approach[9], as well as my home department's previously published random forest model[6].

4

Independent Validation and Interpretation

4.1 Independent validation

Given the high performance of the MD-WDNN and L2-regularized logistic regression, I proceeded to validate the models on an independent validation set. For each model on the full predictor set of "common" and "derived" features, I reported ROC curves, AUC values, and two pairs of sen-

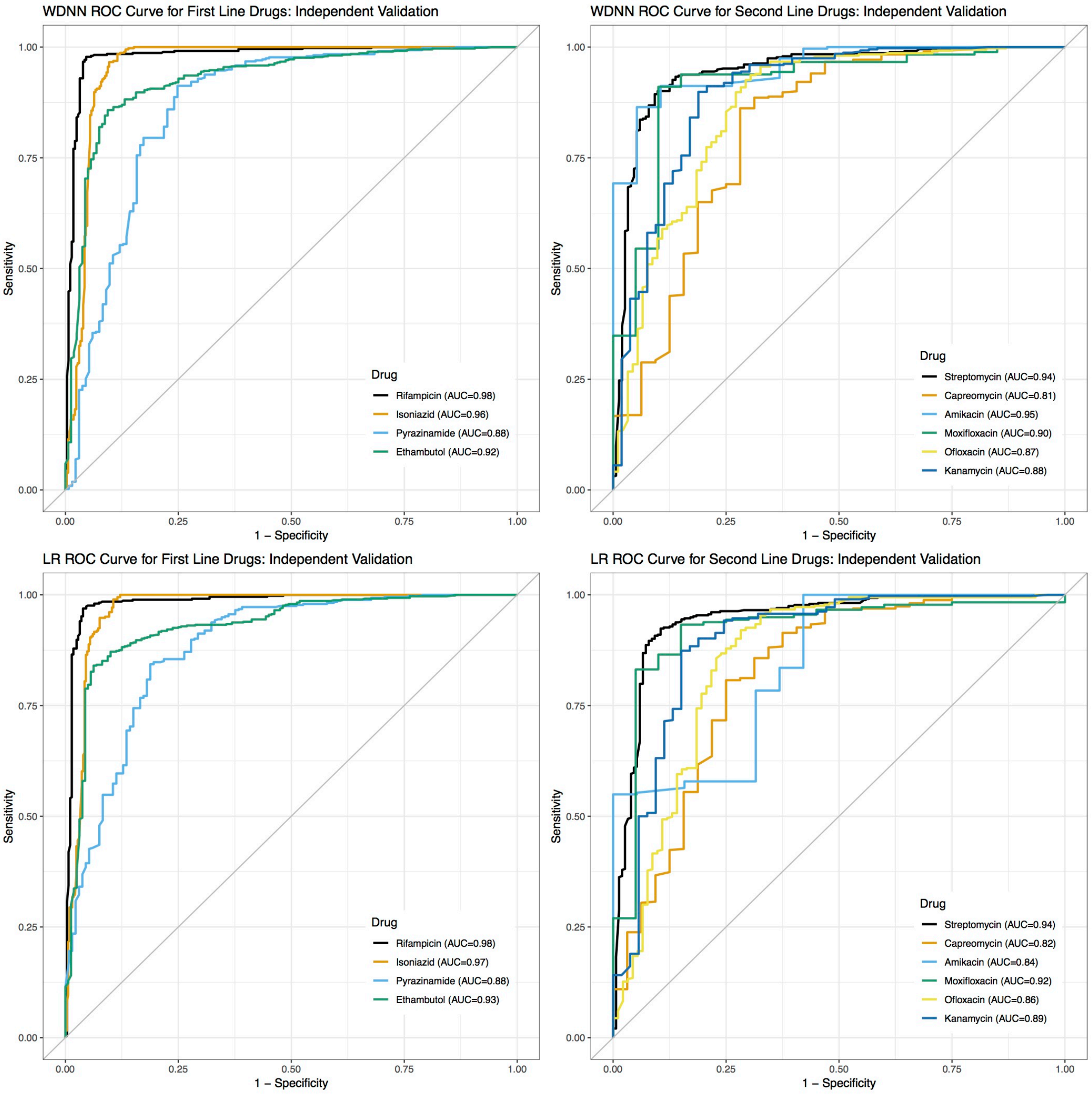
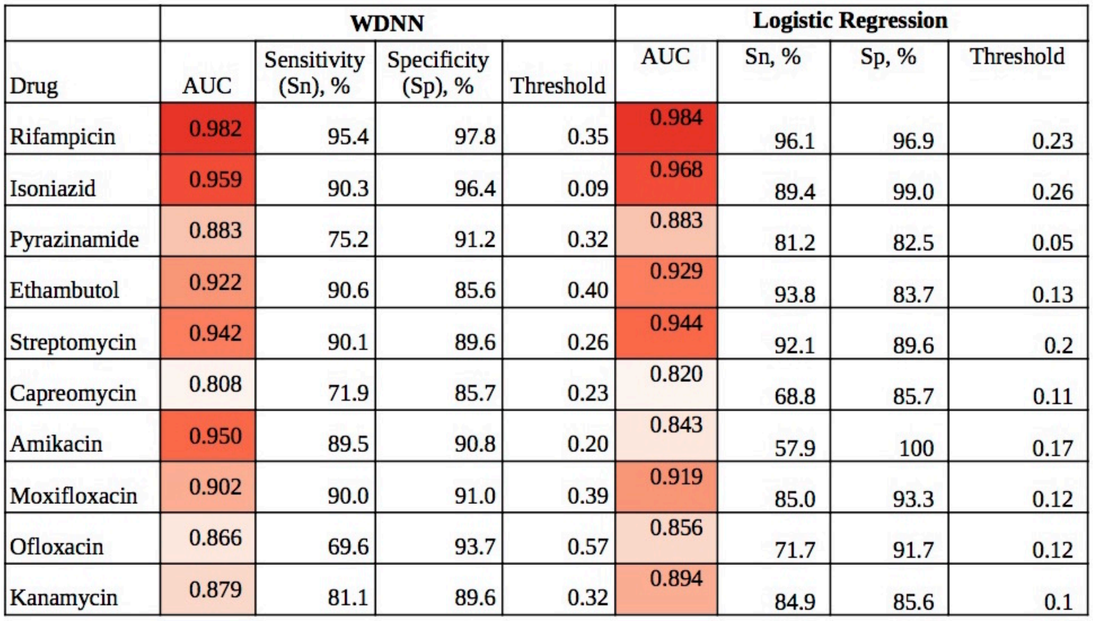


Figure 4.1: Tuberculosis drug resistance ROC performance curve of the MD-WDNN and L2-regularized logistic regression. A ROC plot of MD-WDNN (top) and logistic regression (bottom) predictive performance on the independent validation set for first-line (left) and second-line (right) anti-tuberculosis drugs.



	WDNN				Logistic Regression			
Drug	AUC	Sensitivity (Sn), %	Specificity (Sp), %	Threshold	AUC	Sn, %	Sp, %	Threshold
Rifampicin	0.982	95.4	97.8	0.35	0.984	96.1	96.9	0.23
Isoniazid	0.959	90.3	96.4	0.09	0.968	89.4	99.0	0.26
Pyrazinamide	0.883	75.2	91.2	0.32	0.883	81.2	82.5	0.05
Ethambutol	0.922	90.6	85.6	0.40	0.929	93.8	83.7	0.13
Streptomycin	0.942	90.1	89.6	0.26	0.944	92.1	89.6	0.2
Capreomycin	0.808	71.9	85.7	0.23	0.820	68.8	85.7	0.11
Amikacin	0.950	89.5	90.8	0.20	0.843	57.9	100	0.17
Moxifloxacin	0.902	90.0	91.0	0.39	0.919	85.0	93.3	0.12
Ofloxacin	0.866	69.6	93.7	0.57	0.856	71.7	91.7	0.12
Kanamycin	0.879	81.1	89.6	0.32	0.894	84.9	85.6	0.1

Table 4.1: Tuberculosis drug resistance predictive performance of the MD-WDNN and logistic regression on the independent validation set. Area under the ROC curve classification performance on the independent validation set. I also report sensitivity and specificity performance with the probability threshold chosen to maximize the sum of sensitivity and specificity for all anti-tuberculosis drugs. The cells are colored from lightest to darkest for lowest to highest AUC across the 10 drugs for each model.

sitivity (Sn) and specificity (Sp) performance on the independent validation set. While the ROC curve and AUC score are useful for summarizing performance at different probability thresholds, Sn and Sp are useful in understanding the performance of the models given a probability threshold used to distinguish resistance versus susceptibility. For Sn and Sp, the first pair I reported used a probability threshold to maximize the sum of Sn and Sp for each drug. For the second pair, I determined the probability threshold to maximize Sn given that the Sp is at least 90%. The 90% specificity threshold stems from the value assessment that over-diagnosis of antibiotic resistance is more harmful than under-diagnosis due the treatment toxicity and side effects, e.g. renal failure and hearing loss, for the drugs used in antibiotic resistant cases.

The ROC curves for the two final models on the independent validation data across the 10 anti-tuberculosis drugs are shown in Figure 4.1, illustrating the different Sn and Sp performance values at probability thresholds between 0 and 1. Table 4.1 shows the AUC corresponding to the ROC curves for each drug. The average AUCs for the MD-WDNN were 0.937 for first-line drugs and 0.891 for second-line drugs on an independent validation set, which were slightly lower than the AUCs during repeated cross-validation (AUC = 0.979 for first-line drugs, AUC = 0.936 for second-line drugs). The AUCs for L2 regularized logistic regression were 0.941 for first-line drugs and 0.879 for second-line drugs.

Due to class imbalance for some of the drugs, I also measured and reported performance using the precision-recall curve (Appendix Figure B.1), as this metric may be more informative for rare events[21]. The comparison between the MD-WDNN and logistic regression performance according to the precision-recall curve largely aligns with the AUC metric during cross-validation (Table 3.1 and Appendix Table A.6). I do note, however, there is a sizeable gap in average precision (AP) between the MD-WDNN and logistic regression models for three drugs on the independent validation set: capreomycin, amikacin, and moxifloxacin. The MD-WDNN achieved APs of 0.5, 0.74, 0.63 while logistic regression had APs of 0.45, 0.64, and 0.55 for those three drugs, respectively.

Table 4.1 shows the Sn, Sp, and corresponding probability threshold that maximizes the sum of Sn and Sp for each model-drug combination. For the MD-WDNN, the average Sn and Sp, respectively, on the independent validation set were 87.9% and 92.7% for first-line drugs and 82.0% and 90.1% for second-line drugs. For L2 regularized logistic regression, the average Sn and Sp, respectively, on the independent validation set were 90.1% and 90.5% for first-line drugs and 76.7% and 91.0% for second-line drugs. Notably, the two models perform similarly, with L2 regularized logistic regression slightly higher on average, for drugs except amikacin. For amikacin, the MD-WDNN significantly outperforms L2 regularized logistic regression, with an increased AUC of 0.107 and increased sum of Sn and Sp of 22.4%. Sn and Sp values for the second probability threshold, which maximizes Sn given that Sp is at least 90%, are available in Appendix Table A.7.

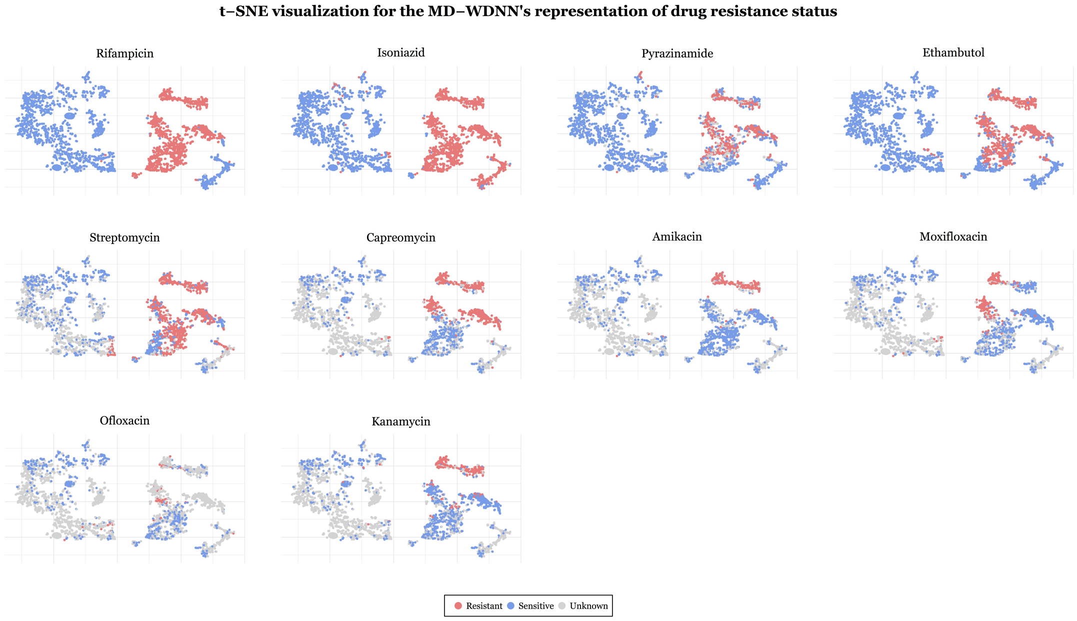


Figure 4.2: t-SNE visualization for the final output layer of the MD-WDNN. The final layer predictions, originally in 11 dimensions, were projected onto two dimensions. Each point is an MTB isolate, colored according to its resistance status with respect to the corresponding drug.

I also tested the prediction run time for each model on the independent validation set. The MD-WDNN prediction time was 0.0352 seconds, and the L2 regularized logistic regression prediction time was 0.00291 seconds.

4.2 t-SNE visualization

One method to visualize the components of a deep learning model with high dimensionality is through the t-distribution stochastic neighborhood embedding (t-SNE) method, which is a non-linear dimensionality reduction technique[22]. I applied t-SNE separately to (1) the input genetic predictors, which included the 222 "common" mutations and "derived" features and (2) the MD-WDNN final output layer predictions, which was originally in 11 dimensions. Each point represented one MTB isolate and was colored based on its phenotypic status for each drug. t-SNE on the input genetic markers showed well-defined clusters, and each cluster contained both susceptible and multidrug resistant isolates with little discernable pattern of resistance classification as shown in Appendix Figure B.2. Conversely, Figure 4.2 demonstrates clear separation by the MD-WDNN's output representation between resistant and susceptible isolates, consistent with my reported measurements of high model Sn and Sp.

The t-SNE plots demonstrate the multitask WDNN's ability to classify resistance across multiple drugs, separating them into nested groups of pan-susceptible isolates, followed by mono-isoniazid resistant isolates, multidrug resistant isolates, pre-XDR isolates, and XDR isolates. This is consistent with the order of administration of the drugs clinically as well as the usual order of *Mycobacterium tuberculosis* drug resistance acquisition[23]. The second-line injectable drugs, amikacin, capreomycin, and kanamycin, also show similarly-classified clusters, highlighting the moderate level of cross resistance between them. I also observe moderate levels of cross resistance among the fluoroquinolones despite the fact that fewer isolates were tested for resistance to these agents[24].

I overlaid the lineage clustering on the two t-SNE plots to determine the effect of lineage on both the input genetic marker representation and MD-WDNN final layer representation of the isolates. The input genetic data t-SNE coordinates largely recapitulated the genetic clustering due to lineage (Figure 4.3), which aligns with the understanding that the largest genetic differences between isolates were related to lineage. On the other hand, overlying t-SNE coordinates for the MD-WDNN's probabilistic representation (Appendix Figure B.3) with lineage coloring showed little pattern between t-SNE's representation of the MD-WDNN output layer and the determined lineage. The lack of association confirmed that the MD-WDNN's prediction of phenotype was not simply predicting on the basis of lineage-related variation.

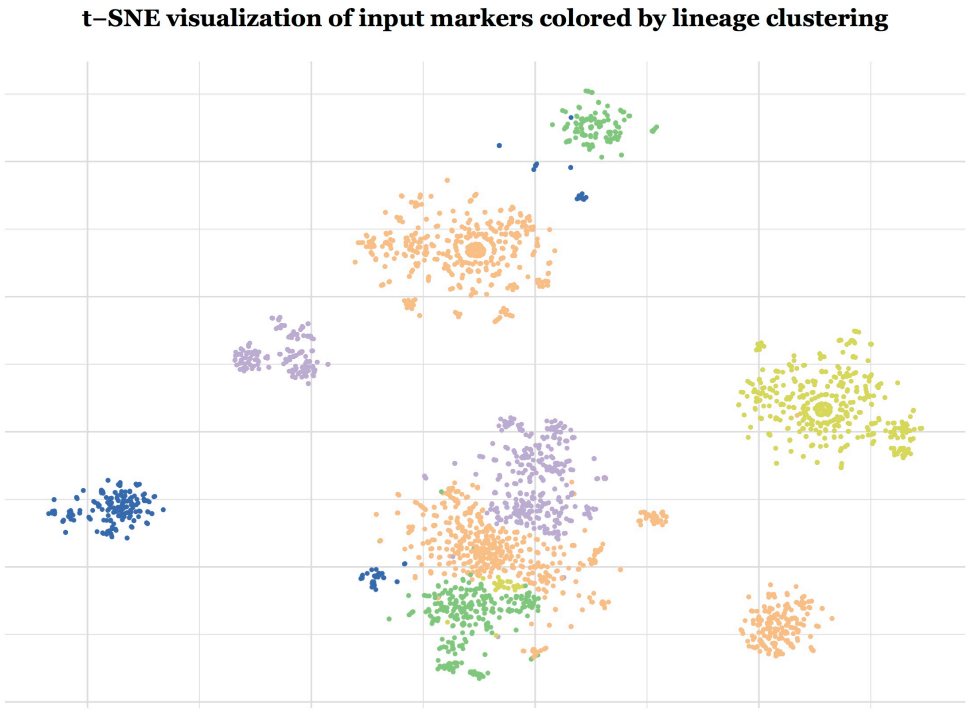


Figure 4.3: t-SNE visualization for input markers colored by lineage clustering. t-SNE plot of the 222 input genetic markers, including "common" mutations and "derived" categories. The coordinates are the same as in Appendix Figure B.2. Each isolate is colored based on the five lineage clusters determined in Figure 3.1, illustrating that the largest genetic differences between isolates were related to lineage.

4.3 Genetic variant importance to resistance

I examined predictor importance to resistance by analyzing the prediction outputs of the MD-WDNN and the presence or absence of mutations through permutation testing. I investigated the following null and alternative hypotheses:

H_0: MD-WDNN's probability of resistance unrelated to presence of variant

H_A: MD-WDNN's probability of resistance related to presence of variant

To build the reference distribution, I permuted the resistance labels randomly and calculated the distribution of the following difference as the desired test statistic, S, where R represents the event where the isolate is resistant, V represents the event that a given variant is present in an isolate, and V^C is the complement of V:

$$S = P(R|V) - P(R|V^C)$$

Note that each term, $P(R|V)$ or $P(R|V^C)$, is the MD-WDNN's predicted probability of resistance after permuting the resistance labels.

I then compared the actual difference above with the permutation distribution of test statistics. I created the sampling distribution with 100,000 randomized permutations per mutation and the actual differences were evaluated at a significance level of $\alpha = 0.05$ using a Bonferroni correction for the 222 multiple comparisons. I conducted the permutation test for each predictor ("common" mutations and "derived" categories) that was present in at least 30 tuberculosis isolates.

Of the 156 mutations and 56 derived categories, the majority were found to be significant "resistance predictors" for one or more drug: rifampicin (103 mutations, 40 derived), isoniazid (102 mutations, 42 derived), pyrazinamide (94 mutations, 38 derived), ethambutol (96 mutations, 44 derived), as well as the second-line drug streptomycin (98 mutations, 42 derived). Of the remaining

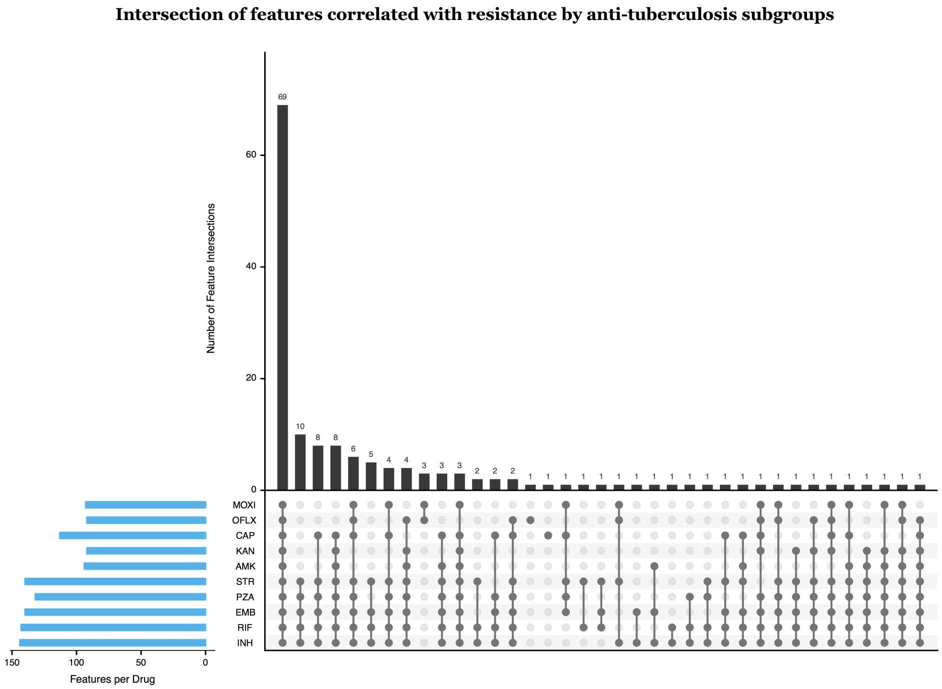


Figure 4.4: Intersection of predictors correlated with resistance by anti-tuberculosis drug subgroups. I permuted the resistance labels and calculated the distribution of the difference of resistance probability given the presence or absence of each predictor. I show the number of mutations per subgroup of drugs ordered from most to least mutations per subgroup. Number of significant predictors per drug is also shown.

predictors, the highest number of "susceptibility predictors" were found in isoniazid (39 mutations, 0 derived), rifampicin (37 mutations, 1 derived), streptomycin (37 mutations, 0 derived), ethambutol (36 mutations, 0 derived), and pyrazinamide (32 mutations, 2 derived). Figure 4.4 illustrates the number of significant resistance predictors per drug in the MD-WDNN and their intersections among different drug subsets. Subsets of drugs that included a second-line injectable drug and shared at least two predictors consistently included both INH and RIF. This is consistent with previous findings that tuberculosis isolates acquire resistance to first-line drugs before second-line drugs[23] and indicates that the multidrug model was able to capture these relationships. The subset of fluoroquinolones shared 3 resistance-correlated predictors not found in other first-line or second-line drugs, and reflect that fluoroquinolones have a mechanism of action that differs from those of first-line and second-line drugs[25].

I also examined predictor importance to the L2 regularized logistic regression model using bootstrapping of the models' fitted coefficients. The evalulation of predictor performance was as follows: first, I created a bootstrap sample of isolates equal to the size of the original dataset of 3,601 isolates. Second, I exponentiated the β coefficients of the logistic regression model to obtain the odds-ratio for each mutation. These odds-ratios are directly interpretable to determine the importance of the variant, where an odds-ratio greater than 1 indicates association with susceptibility and an odds-ratio less than 1 indicates association with resistance. Third, I repeated this process for 10,000 bootstrapped samples. Fourth, I built a confidence interval for each predictor-drug pair's exponentiated β coefficient using a significance level of $\alpha = 0.05$ and a Bonferroni correction for the 222 multiple comparisons. I made the final determination of a variant's importance to resistance or susceptibility based on whether the confidence interval contained 1, where a confidence interval containing only values less than 1 meant that the variant was associated with antibiotic resistance.

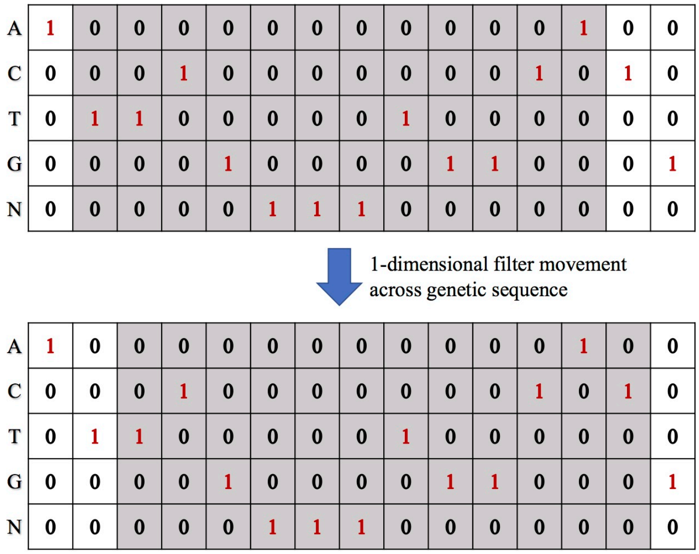
There was a large degree of overlap between important predictors for the MD-WDNN and L2 regularized logistic regression. The number of significant resistance predictors in overlap between the two models were 141 predictors for isoniazid, 128 for pyrazinamide, and 139 for streptomycin, including multiple derived categories. Both models successfully excluded variables known to be neutral or lineage markers, such as excluding gyrA S95T from association with fluoroquinolone resistance. The MD-WDNN and permutation measure of importance classified a larger proportion of the variants as associated with susceptibility than did L2 regularized logistic regression. For example, 39 mutations in the MD-WDNN measure were associated with isoniazid susceptibility, whereas 2 mutations were associated with isoniazid susceptibility by L2 regularized logistic regression. Overall, 52 genetic variants were associated with susceptibility to one or more drugs, including 19 known lineage markers. Both lists included non-canonical and rare variants among the top most important variables for resistance prediction.

5

Extension to CNN

5.1 CNN architecture

Convolutional neural networks (CNNs) are more complex alternatives to deep neural networks and have seen great success in many areas of medicine, including in identifying skin cancer from clinical images[26] and in identifying diabetic retinopathy from retinal images[27]. One benefit of a CNN is its ability to incorporate spatial information from the training data. In the case of images, CNNs



A	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
C	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0
T	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
G	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1
N	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0

A	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
C	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0
T	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
G	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1
N	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0

Figure 5.1: Schematic of 1D convolution and genetic sequence encoding. The key difference in the CNN compared to the wide and deep neural network is the existence of convolutional layers, which contain 1D filters. These filters slide across the genetic sequence with a stride length of one nucleotide. The schematic depicts the first convolutional layer of my CNN, with a filter of length 12 and the one-hot encoding of a genetic sequence.

capture relative spacing of pixels; in the case of genomic sequencing, CNNs capture relative spacing of nucleotides and genetic motifs. Using sequence-formatted data rather than variant-formatted data, I hypothesize that one major benefit of a CNN approach on sequence-style whole genome sequencing data is its ability to capture more complex genetic spatial interactions. By design, the CNN will also aim to capture genetic motifs that contribute to antibiotic resistance.

The key architectural feature of CNNs in genomic data analyses is convolutional layers, which contain one-dimensional filters that are scanned across the sequence of nucleotides. Figure 5.1 shows a schematic representation of how a filter moves, where the filter placement is designated by the grey region. Since the filter overlaps the genetic sequence, the actual values of the filter are not shown. The schematic shows a filter of size 12 nucleotides that has scanned one nucleotide across the genetic sequence represented in a one-hot encoding. The fifth base, N, represents the null character. The presence of the null character means there is a relative deletion in the particular tuberculosis isolate compared to the full-length reference strain. After each one-nucleotide translation of the filter, the Hadamard product,

$$H = F \odot S$$

is calculated between the filter (F) and the nucleotide encoding (S) within the current filter window. All elements are summed in the H matrix:

$$o = \sum_i \sum_j h_{ij}$$

The output, o, is then used in downstream calculations within the network; in the current case, this includes adding a bias term, downsampling through a max pooling layer, and either continuing through another convolutional layer or proceeding onto the fully connected layers of the network. All filters within the convolutional layers are fit through backpropagation. While the dimensions of the second convolutional layer change, the same intuitive concept of fitting feature-detecting filters upon the previous layer's output applies.

Table 5.1 shows the architectural features of the CNN. As mentioned in Chapter 2, the four antibiotics of interest for resistance prediction in this section are rifampicin, pyrazinamide, capreomycin, and kanamycin. A new CNN was trained for each antibiotic, since the input genetic sequence differed for the each antibiotic based on previously known genetic regions that are important for resistance for each particular drug (details are available in Chapter 2).

5.2 Empirical analysis

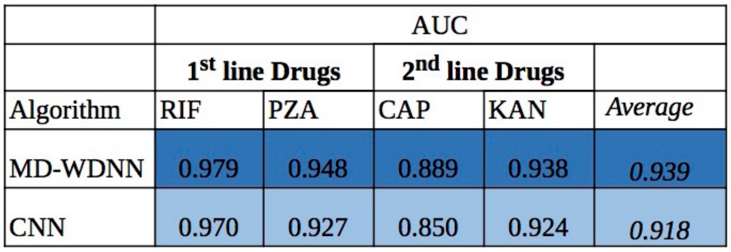
In the expanded set of 10,198 tuberculosis whole genome sequences, I investigated the performance of the convolutional neural network approach (Table 5.2). The MD-WDNN shows high perfor-

Layer	Hyperparameters	Activation
Input layer	Input size = length of nucleotide sequence	
1D-Convolutional layer	Number of filters = 64 Filter size = 12 Stride length = 1	ReLU
1D-Max pooling layer	Pool size = 3	
1D-Convolutional layer	Number of filters = 32 Filter size = 3 Stride length = 1	ReLU
1D-Max pooling layer	Pool size = 3	
Flatten layer		
Dense layer	Number of nodes = 256	ReLU
Dense layer	Number of nodes = 256	ReLU
Dense layer	Number of nodes = 1	Sigmoid

Table 5.1: Convolutional neural network architecture. All hyperparameters and nonlinear transformations, where applicable, are shown above. No padding is added in convolutional layers. The final output layer contains one node with a sigmoid activation, representing the final predicted probability of antibiotic resistance for one drug. A different CNN model with the same architecture is trained for each antibiotic, because the input genomic sequences differ for each drug of interest.

mance across all four drugs on the expanded dataset, with an average AUC of 0.939. The CNN shows promising but slightly lower predictive performance for all four antibiotics tested, with an average AUC of 0.918.

Compared to the independent validation results for the MD-WDNN in Chapter 4, the MD-WDNN and CNN both show a higher AUC for all antibiotics except for rifampicin. For rifampicin, the MD-WDNN had an AUC of 0.982 during independent validation, whereas on the expanded dataset, the AUCs weere 0.979 and 0.970 for the MD-WDNN and CNN, respectively. The average increase in AUC across the remaining 3 drugs for the CNN was 0.044, whereas the average increase for the MD-WDNN was 0.068. This provides evidence that with an expanded and enriched set of isolates, there can be large gains in performance, especially in the context of second-line drugs and multidrug resistance.



	AUC				
	1st line Drugs		**2nd line Drugs**		
Algorithm	RIF	PZA	CAP	KAN	*Average*
MD-WDNN	0.979	0.948	0.889	0.938	*0.939*
CNN	0.970	0.927	0.850	0.924	*0.918*

Table 5.2: AUROC performance of MD-WDNN and CNN models on expanded dataset. A table of predictive performance across the MD-WDNN and CNN during five-fold cross-validation. The MD-WDNN was trained on the variant-style format of whole genome sequencing data, whereas the CNN was trained on the sequence-style format. Performance is shown in average AUC across all cross-validation folds. The cells are colored by rank of the model for each drug, colored from lightest to darkest corresponding with lowest to highest AUROC value.

One important note is the limited genetic scope of the CNN. I trained the CNN on only 1-2 genes per antibiotic based on previously known genetic regions that inform resistance. As shown in Chapter 3, the incorporation of "common" mutations and "derived" predictors results in a significant increase in predictive performance. For future work, I believe with the incorporation of more genomic regions and potential multidrug modifications to the model, the CNN is a promising model architecture for identifying multidrug resistance.

5.3 Saliency map

Understanding which regions of the genome are associated with which resistance phenotypes can help identify resistance-determining regions, which can provide insight into the biological mechanisms behind the acquisition of resistance[28]. Because the CNN's sequence data has a linear spatial representation, interpreting the CNN model is possible through a saliency map. A saliency map provides a visualization of the importance of each nucleotide position of the input genetic sequence to the possible output classes (in my case, resistance or susceptibility). I implemented a procedure for building a saliency map, with adaptation of prior research by Zou et al. (2018)[29], for each of the four drugs and their corresponding genetic regions as follows:

1. For the antibiotic of interest, extract the gradients of the loss function (L) with respect to each nucleotide, j, in the input sequence (S_i) for each tuberculosis isolate:

$$\text{grad} = \frac{\partial L}{\partial S_{ij}}$$

2. Determine the contribution of the gradient, C_{ij}, for the particular input sequence by conducting element-wise multiplication of the one-hot encoded genetic sequence with the gradient sequence. I do so by taking the Hadamard product of the gradient and genetic sequence:

$$C_{ij} = \text{grad} \odot S_{ij}$$

3. Due to the one-hot encoding structure, there exists only one non-zero contribution value per nucleotide position. Take the only non-zero value per position and place into a one-dimensional vector the length of the sequence:

$$C_i = \sum_j C_{ij}$$

4. (a) If determining nucleotide positions that are correlated with resistance, take only contributions from resistant isolates. This is justified because it is reasonable to look at contributors of resistance only if the isolate is actually phenotypically resistant. For those isolates, take only negative contributions of the gradient, as 0 encodes resistance:

$$C_i = \begin{cases} min(C_i, 0) & i \text{ is resistant} \\ 0 & i \text{ is susceptible} \end{cases}$$

   (b) If determining nucleotide positions that are correlated with susceptibility, take only contributions from susceptible isolates. For those isolates, take only positive contributions of the gradient, as 1 encodes susceptibility:

$$C_i = \begin{cases} max(0, C_i) & i \text{ is susceptible} \\ 0 & i \text{ is resistant} \end{cases}$$

5. Visualize the results in a one-dimensional saliency map averaged over all relevant isolates, with *x*-coordinates shown as the coordinates in relation to the *Mycobacterium tuberculosis* H37Rv reference strain. The y-axis shows the magnitude of the gradients' contribution, which is a measure of importance to antibiotic resistance or susceptibility.

The saliency map showing the important predictor regions of resistance is reported in Figure 5.2. All values of the saliency map are negative by design (see above procedure for justification), and a larger magnitude of saliency means that a particular mutation in that position is more highly associated with antibiotic resistance. One notable feature of the figure is that rifampicin, capreomycin, and kanamycin have generally well-defined regions within each gene that are important to resistance. On the other hand, pyrazinamide shows less-defined regions of importance to resistance within the two genes analyzed. This is in line with the fact that resistance to pyrazinamide is thought to be caused by the aggregation of a number of mutations[6]. On the other hand, there are known to be tight genetic regions within the *rpoB* gene that are causative of resistance for rifampicin[28]. The findings from the saliency map are thus corroborated by the molecular understanding of genetic acquisiton of resistance, and the saliency map serves as a potential gateway for identifying new determinants of resistance.

The genetic predictors that are important to susceptibility within susceptible tuberculosis strains are also informative for understanding antibiotic resistance. By design, all saliency values are positive. The results for the susceptible strains show agreement with Figure 5.2. The full saliency map

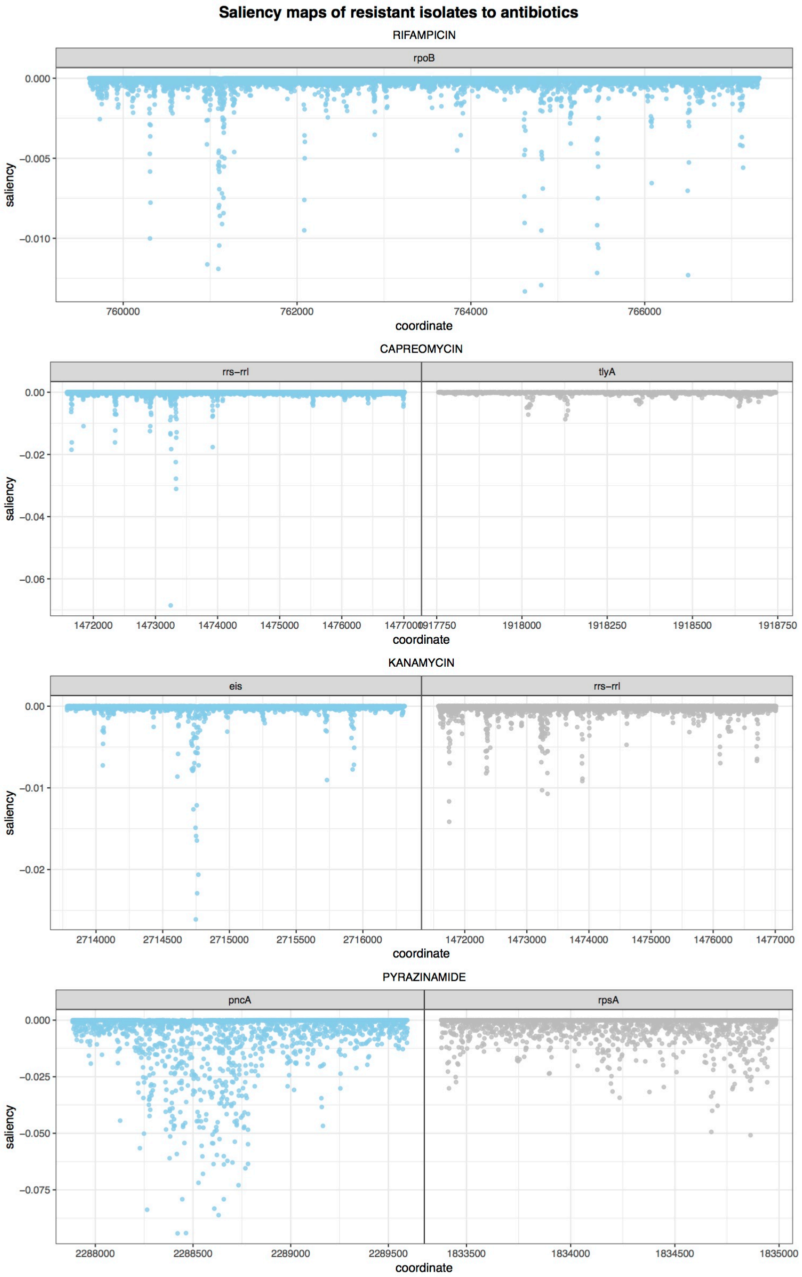


Figure 5.2: Relative importance of genomic coordinates to resistance phenotype within each antibiotic for the CNN. A visualization showing the relative importance, as measured in a gradient-based calculation of saliency, of each coordinate within the input genetic sequence. The x-axis shows the genomic coordinates relative to the H37Rv strain, which is standard practice within tuberculosis genomics. Any insertions relative to the H37Rv strains are given a fractional coordinate value between the two flanking H37Rv coordinates for ease of visualization and numerical consistency with the reference strain.

regarding importance to susceptibility is available in Appendix Figure B.4.

6

Discussion

The primary aim of this thesis was to construct a highly accurate model of drug resistance through the implementation and analysis of different statistical and deep learning methods trained on both genomic variant-style and sequence-style data. I demonstrated that L2 regression and MD-WDNN trained on a large diverse dataset using a method of aggregating rare variants outperforms my department's previously reported random forest model[6]. The CNN approach showed promising performance as well, especially in context of the limited genomic data available to the model.

A few prior studies have utilized algorithmic or machine learning methods using genomic data to account for the complex relationship between genotype and drug resistance in MTB[6,9,30,14,31]. Compared to one study that used a direct association (DA) algorithm, the machine learning approaches presented here offer improvement in Sn and Sp for the majority of drugs when prediction is attempted on all isolates, including those with rarer and not previously observed variants[9]. For example, DA had Sn and Sp for predicting pyrazinamide resistance of 24% and 99%, respectively, if prediction was attempted on all isolates including those with uncharacterized variants. The MD-WDNN performance on an independent dataset achieved Sn of 75.2% and Sp of 91.2%. The best sum of Sn and Sp for the L2 regularized logistic regression model showed Sn of 81.2% and Sp of 82.5%, and fixing Sp to at least 90% for comparability with MD-WDNN results in LR Sn of 70.7%. Similarly, the MD-WDNN and logistic regression Sn and Sp were 69.6%/93.7% and 71.7%/91.7%, respectively, for ofloxacin, whereas with DA, the Sn and Sp were 45% and 100%, respectively[9]. Another study used single-task machine learning, demonstrating the validity of this approach for identifying MDR and XDR-TB, but the study did not verify their findings using independent validation data, raising concerns about generalizability[31]. Additionally, the best models in the study used dimensionality reduction (sparse PCA) for two drugs (capreomycin and amikacin) to address the problem of rare and sparse inputs, limiting the interpretability for models of these drugs. In contrast, the MD-WDNN and CNN approaches used an interpretable set of inputs (no-dimensionality reduction), while also achieving substantially higher MD-WDNN performance in cross-validation with AUCS of 0.96 and 0.95 for CAP and AMK, compared to AUCs of 0.85 and 0.91 reported in their study[31]. Across all drugs tested, the MD-WDNN approach showed higher performance in 8 of the 10 drugs during cross-validation compared to their highest performing model for each drug (Table A.9). The increase in average AUC of the MD-WDNN was 0.014 for first-line drugs and 0.025 for second-line drugs. Third, their analysis did not demonstrate the lack of confounding by lineage and report some lineage variants as predictive of resistance.

The current MD-WDNN approach has several novel features. First, I included all variants in the set of 32 genetic loci as potential predictors of resistance to any drug and did not subset the variants according to *a priori* knowledge of causative relationships between genetic loci and drugs. The predictive performance gains offered by this more "permissive" approach were considerable especially for the second line drugs, and the first-line drug pyrazinamide. Second, I utilized rare variants through the method of forming derived groups of mutations, resulting in large performance gains for certain drugs. Third, this is the first neural network model for resistance prediction from MTB genotypic data. I attempted to incorporate prior information about the genetic etiology of MDR and XDR directly into the structure of the deep neural network, as it is known that both individual markers and gene-gene interactions confer resistance[7]. The wide portion of the network allows the effect of individual mutations (e.g. marginal effects) to be easily learned, while the deep portion of the network allows for arbitrarily complex epistatic effects to influence the predictions. Fourth, I am the first to examine a multidrug approach that allows drugs with less phenotypic data to borrow pathway information from others with a higher number of phenotyped isolates. To some extent, this proved to be true as demonstrated by Figure 3.3.

I acknowledge that with the use of a more complex model, there is an increased risk of overfitting to the data during repeated cross-validation. I used techniques such as dropout and L2 regularization at each layer of the MD-WDNN to mitigate the effect of overfitting. Furthermore, I sought to evaluate potential overfitting through the analysis on an independent validation set, which showed performance with high clinical relevance. Finally, I re-trained the MD-WDNN on an expanded dataset of 10,198 isolates and showed high performance through cross-validation as well. In light of these considerations, the MD-WDNN model presented here is the first multitask tool that provides the full antibiogram for 10 anti-tuberculosis drugs in one run. I successfully built high performing deep learning models to predict anti-tuberculosis drug resistance, although the performance gains from these more complex methods are not yet fully justified over simpler models, except in the case of amikacin, where the improvement was considerable. I expect the benefits of these deep learning models to increase when incorporating more genetic loci into the predictor set.

Although the gains that I attribute to the multitask architecture *per se* were not significant, the gains were quantitatively larger for second line drugs like kanamycin and ofloxacin. As second-line injectables and fluoroquinolones are cornerstone agents for the treatment of MDR-TB treatment, and accurate prediction of susceptibility to these agents is key in determining a patient's candidacy for the recently recommended shortened MDR-TB regimen, this approach holds promise as more genomic data is incorporated[32]. Prediction of resistance to second-line injectables has thus far been challenged by a limited genetic knowledge base and consequently limited Sn when using simple direct association approaches[9]. Thus, in aggregate, the use of a more complex approach, such as the multidrug WDNN, shows promise for performance gains in pyrazinamide and second line drugs. Furthermore, even for drugs like isoniazid and rifampicin that had high performance across the model architectures and the feature categories I tested, the multidrug WDNN validation performance exceeds prior models. This is likely a result of using a larger and richer TB dataset than has been previously used and using a multivariate approach to prediction.

In addition to the approach of the MD-WDNN, I am the first to my knowledge that investigated a convolutional neural network approach. There are a couple notable features to the CNN. First, this is the first model to my knowledge that directly analyzes genomic sequences to identify *Mycobacterium tuberculosis* antibiotic resistance. Second, I incorporated spatial features of the genetic regions to inform the prediction. Third, I implemented a visualization method that makes the CNN highly interpretable. I believe that by adding more genetic regions into the training data and through incorporating prior information into the structure of the model, such as using dilated convolutions, the performance of the CNN approach will improve.

The translation of the modeling approach is also a function of advancements in whole genome sequencing and accessibility to more MTB isolate data. Improvements in whole-genome sequencing technologies have significantly reduced costs[33], allowing for more routine whole genome sequencing in MTB isolates[34,35]. The prediction time for MTB drug resistance depends primarily on the sequencing turnaround time, which is significantly shorter than phenotypic susceptibility testing[36]. In addition, as more routine sequencing increases the amount of MTB isolate data, all reported models can be rapidly updated as the datasets become accessible. I expect that as more data are incorporated, the Sn and Sp gap in second-line injectable drugs and fluoroquinolones will become smaller.

I acknowledge some limitations of my thesis work. First, one source of bias could be errors during phenotyping, as susceptibility testing for some drugs has been shown to have low reproducibility and high variance[37,38]. However, I used strains with phenotypic data measured at national or supranational TB reference laboratories following strict quality control or carefully curated from research and reference laboratories[6,10]. Beyond technical or laboratory limitations in testing, certain resistance mutations, especially for ethambutol and second-line drugs, may result in minimum inhibitory concentrations (MICs) very close to the clinical testing concentration, which may result in lower Sn and Sp[39] when predicting a binary resistance phenotype. The use of MIC data for building future learning models may help circumvent this. Second, I only included mutations that occurred in over 0.8% (30 of 3601 isolates) individually or when aggregated with other rare variants in the same gene or intergenic region. Although I may have missed some important predictors, this threshold amounted to only ignoring variants that are very rare in a diverse sample of MTB genomes with good representation from the major genetic lineages. Third, I did not include third-line anti-tuberculosis drugs such as cycloserine or para-aminosalicylic acid due to the lack of phenotypic data.

7

Conclusion

In summary, I present an implementation and exploration of deep learning and traditional statistical models to identify the resistance of MTB isolates to 10 anti-tuberculosis drugs from whole genome sequencing data. The models were trained on rare and common genetic variants, as well as on sequence-formatted genetic sequences. The models achieved state-of-the-art performance on large, aggregated TB datasets, with prediction times of less than a tenth of a second, demonstrating the efficacy of the models as diagnostic tools for MTB drug resistance. The MD-WDNN represented the first multidrug model to my knowledge that incorporated a high number of genotypic predictors known to be important to determining resistance for one or more included drugs. The extension to the CNN approach laid a promising foundation for future work to incorporate more advanced techniques and larger regions of the tuberculosis genome. Further work identifying the impact of a wide range of genetic determinants will not only allow for improved predictive performance but may also provide a greater understanding of the biological mechanisms underlying drug resistance in MTB isolates.

A

Supplementary Tables

Drug	Susceptible Isolates	Resistant Isolates
RIF	2257	1285
INH	2011	1553
PZA	2445	702
EMB	2551	975
STR	1155	1025
CAP	799	589
AMK	1174	235
MOXI	1118	268
OFLX	651	88
KAN	1060	272

Table A.1: Phenotype of 3,601 tuberculosis isolates during training for 10 anti-tuberculosis drugs using variant-style genome sequencing data.

Drug	Susceptible Isolates	Resistant Isolates
RIF	453	282
INH	384	330
PZA	434	133
EMB	576	160
STR	433	152
CAP	420	32
AMK	273	19
MOXI	178	20
OFLX	363	92
KAN	396	53

Table A.2: Phenotype of 792 tuberculosis isolates in independent validation set for 10 anti-tuberculosis drugs using variant-style genome sequencing data.

Drug	Susceptible Isolates	Resistant Isolates
RIF	6427	3471
PZA	5393	1504
CAP	2836	737
KAN	2500	796

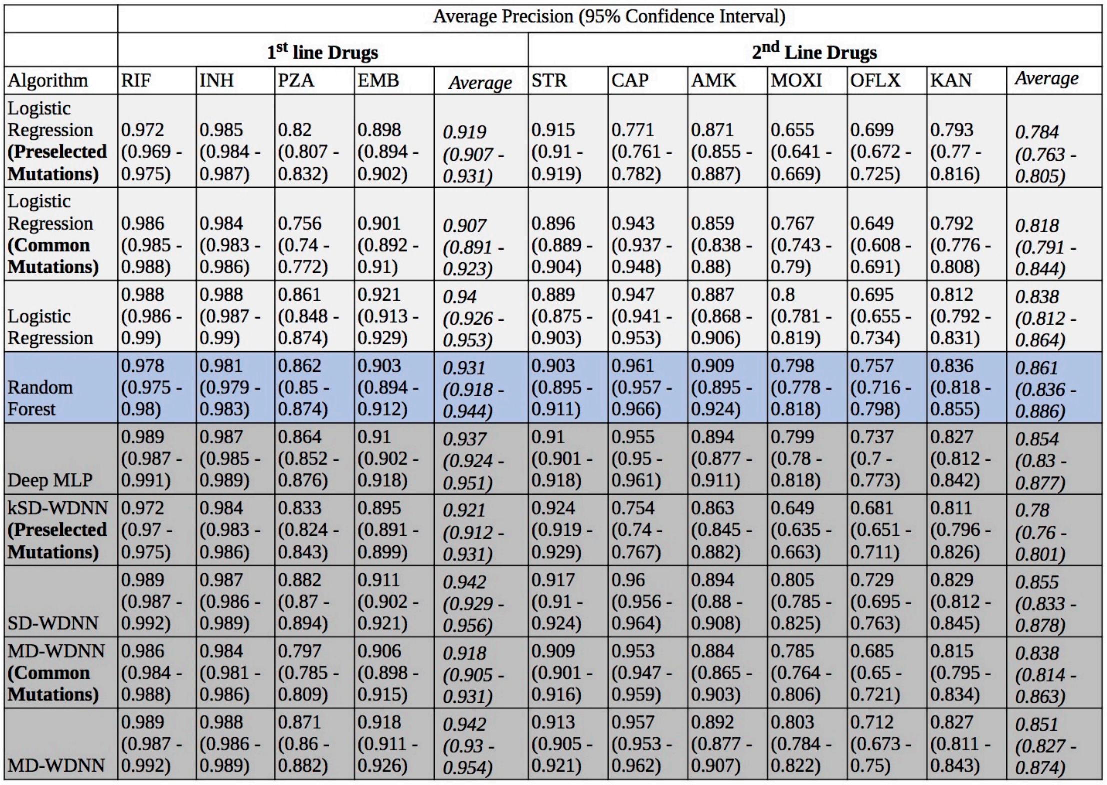
Table A.3: Phenotype of 10,198 tuberculosis isolates in cross-validation for 4 anti-tuberculosis drugs using sequence-style genome sequencing data.

Lineage-defining mutations
inhA_V78A
ndh_R284W
ndh_V18A
katG_R463L
pncA_H57D
iniA_H481Q
embC_V104M
embC_T270I
embC_N394D
embC_R567H
embC_R738Q
embC_V981L
embA_V206M
embA_T608N
embA_P913S
embB_Q139H
embB_E378A
gid_A119T
gid_S100F
gid_E92D
gid_L16R
gyrB_M330I
gyrB_A442S
gyrB_C48T
gyrA_E21Q
gyrA_T80A
gyrA_S95T
gyrA_G247S
gyrA_A384V
gyrA_G668D
rrs_C492T
ahpC_G-88A
rpoB_C-61T

Table A.4: Table of 33 variants that were used in the hierarchical clustering analysis. These variants were used to determine the geographic diversity of the isolates within the variant-style dataset.

MD-WDNN, MD-WDNN (Common Mutations), SD-WDNN, and kSD-WDNN	
Hyperparameter	Value
L2 regularization	10^-8
Hidden units per layer	256
Number of hidden layers	3
Dropout	0.5
Learning rate	e^(-9)
Optimizer	Adam
Epochs	100
Weight Initialization	Xavier uniform initializer
Random Forest	
Hyperparameter	Value
Number of trees	1000
Percentage of predictors to consider for best split	20%
Percentage of samples to split a node	0.2%
Regularized Logistic Regression	
Hyperparameter	Value
L2 regularization	Best penalty factor between 10^-5 and 10^5

Table A.5: A table of hyperparameters for each model. The L2 regularization factor for logistic regression was determined using cross-validation to maximize the AUC within the 80% training data for each fold.

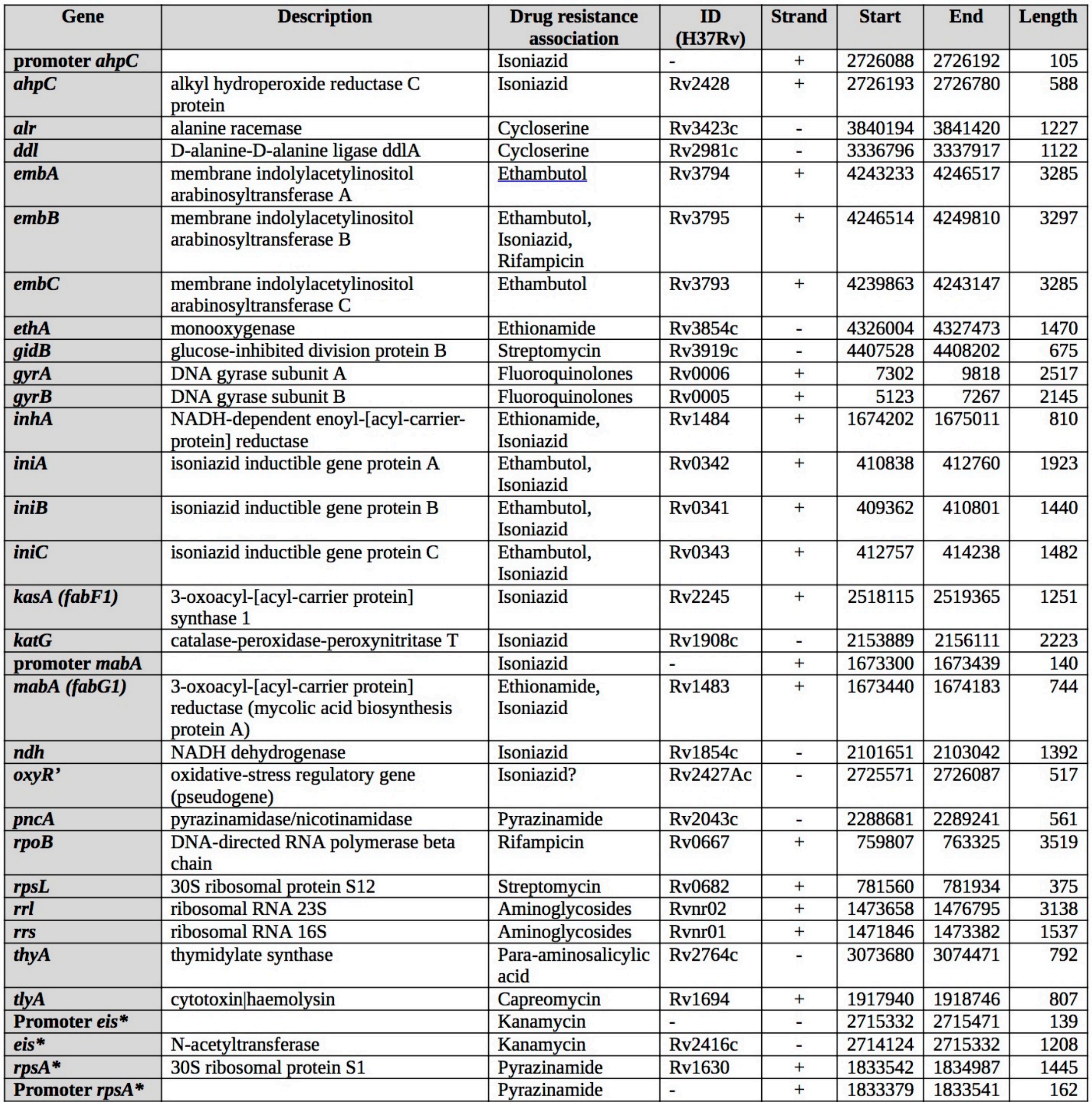


	Average Precision (95% Confidence Interval)											
	1st line Drugs					**2nd Line Drugs**						
Algorithm	RIF	INH	PZA	EMB	*Average*	STR	CAP	AMK	MOXI	OFLX	KAN	*Average*
Logistic Regression **(Preselected Mutations)**	0.972 (0.969 - 0.975)	0.985 (0.984 - 0.987)	0.82 (0.807 - 0.832)	0.898 (0.894 - 0.902)	*0.919 (0.907 - 0.931)*	0.915 (0.91 - 0.919)	0.771 (0.761 - 0.782)	0.871 (0.855 - 0.887)	0.655 (0.641 - 0.669)	0.699 (0.672 - 0.725)	0.793 (0.77 - 0.816)	*0.784 (0.763 - 0.805)*
Logistic Regression **(Common Mutations)**	0.986 (0.985 - 0.988)	0.984 (0.983 - 0.986)	0.756 (0.74 - 0.772)	0.901 (0.892 - 0.91)	*0.907 (0.891 - 0.923)*	0.896 (0.889 - 0.904)	0.943 (0.937 - 0.948)	0.859 (0.838 - 0.88)	0.767 (0.743 - 0.79)	0.649 (0.608 - 0.691)	0.792 (0.776 - 0.808)	*0.818 (0.791 - 0.844)*
Logistic Regression	0.988 (0.986 - 0.99)	0.988 (0.987 - 0.99)	0.861 (0.848 - 0.874)	0.921 (0.913 - 0.929)	*0.94 (0.926 - 0.953)*	0.889 (0.875 - 0.903)	0.947 (0.941 - 0.953)	0.887 (0.868 - 0.906)	0.8 (0.781 - 0.819)	0.695 (0.655 - 0.734)	0.812 (0.792 - 0.831)	*0.838 (0.812 - 0.864)*
Random Forest	0.978 (0.975 - 0.98)	0.981 (0.979 - 0.983)	0.862 (0.85 - 0.874)	0.903 (0.894 - 0.912)	*0.931 (0.918 - 0.944)*	0.903 (0.895 - 0.911)	0.961 (0.957 - 0.966)	0.909 (0.895 - 0.924)	0.798 (0.778 - 0.818)	0.757 (0.716 - 0.798)	0.836 (0.818 - 0.855)	*0.861 (0.836 - 0.886)*
Deep MLP	0.989 (0.987 - 0.991)	0.987 (0.985 - 0.989)	0.864 (0.852 - 0.876)	0.91 (0.902 - 0.918)	*0.937 (0.924 - 0.951)*	0.91 (0.901 - 0.918)	0.955 (0.95 - 0.961)	0.894 (0.877 - 0.911)	0.799 (0.78 - 0.818)	0.737 (0.7 - 0.773)	0.827 (0.812 - 0.842)	*0.854 (0.83 - 0.877)*
kSD-WDNN **(Preselected Mutations)**	0.972 (0.97 - 0.975)	0.984 (0.983 - 0.986)	0.833 (0.824 - 0.843)	0.895 (0.891 - 0.899)	*0.921 (0.912 - 0.931)*	0.924 (0.919 - 0.929)	0.754 (0.74 - 0.767)	0.863 (0.845 - 0.882)	0.649 (0.635 - 0.663)	0.681 (0.651 - 0.711)	0.811 (0.796 - 0.826)	*0.78 (0.76 - 0.801)*
SD-WDNN	0.989 (0.987 - 0.992)	0.987 (0.986 - 0.989)	0.882 (0.87 - 0.894)	0.911 (0.902 - 0.921)	*0.942 (0.929 - 0.956)*	0.917 (0.91 - 0.924)	0.96 (0.956 - 0.964)	0.894 (0.88 - 0.908)	0.805 (0.785 - 0.825)	0.729 (0.695 - 0.763)	0.829 (0.812 - 0.845)	*0.855 (0.833 - 0.878)*
MD-WDNN **(Common Mutations)**	0.986 (0.984 - 0.988)	0.984 (0.981 - 0.986)	0.797 (0.785 - 0.809)	0.906 (0.898 - 0.915)	*0.918 (0.905 - 0.931)*	0.909 (0.901 - 0.916)	0.953 (0.947 - 0.959)	0.884 (0.865 - 0.903)	0.785 (0.764 - 0.806)	0.685 (0.65 - 0.721)	0.815 (0.795 - 0.834)	*0.838 (0.814 - 0.863)*
MD-WDNN	0.989 (0.987 - 0.992)	0.988 (0.986 - 0.989)	0.871 (0.86 - 0.882)	0.918 (0.911 - 0.926)	*0.942 (0.93 - 0.954)*	0.913 (0.905 - 0.921)	0.957 (0.953 - 0.962)	0.892 (0.877 - 0.907)	0.803 (0.784 - 0.822)	0.712 (0.673 - 0.75)	0.827 (0.811 - 0.843)	*0.851 (0.827 - 0.874)*

Table A.6: Tuberculosis drug resistance prediction precision-recall performance of the models examined using repeated cross-validation. A table of average precision, which summarizes the precision-recall curve, across all nine models during repeated cross-validation. The MD-WDNN, SD-WDNN, deep MLP, random forest, and logistic regression models were trained on the full set of predictors. The MD-WDNN (Common Mutations) and logistic regression (Common Mutations) models were trained on mutations not including the derived categories. The kSD-WDNN (Preselected mutations) and logistic regression (Preselected mutations) models were trained on preselected mutations known to be determinants of resistance for each drug. Performance is shown in average precision and 95% confidence interval across all cross-validation folds.

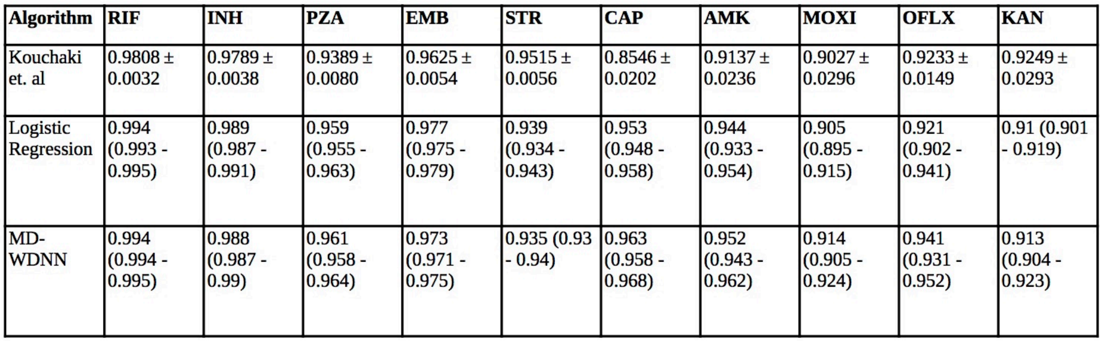
	WDNN			Logistic Regression		
Drug	Sensitivity	Specificity	Threshold	Sensitivity	Specificity	Threshold
Rifampicin	0.968	0.921	0.03	0.968	0.914	0.1
Isoniazid	0.924	0.906	0.03	0.9454	0.901	0.1
Pyrazinamide	0.752	0.901	0.31	0.707	0.901	0.11
Ethambutol	0.819	0.903	0.67	0.813	0.905	0.34
Streptomycin	0.895	0.901	0.27	0.908	0.903	0.21
Capreomycin	0.594	0.902	0.36	0.625	0.9045	0.14
Amikacin	0.895	0.908	0.2	0.579	0.908	0.1
Moxifloxacin	0.900	0.904	0.36	0.850	0.923	0.1
Ofloxacin	0.717	0.904	0.51	0.717	0.917	0.12
Kanamycin	0.792	0.909	0.33	0.773	0.907	0.12

Table A.7: Tuberculosis drug resistance maximum sensitivity with a specificity greater than 90% of the MD-WDNN and L2 regularized logistic regression on the independent validation set.



Gene	Description	Drug resistance association	ID (H37Rv)	Strand	Start	End	Length
promoter *ahpC*		Isoniazid	-	+	2726088	2726192	105
ahpC	alkyl hydroperoxide reductase C protein	Isoniazid	Rv2428	+	2726193	2726780	588
alr	alanine racemase	Cycloserine	Rv3423c	-	3840194	3841420	1227
ddl	D-alanine-D-alanine ligase ddlA	Cycloserine	Rv2981c	-	3336796	3337917	1122
embA	membrane indolylacetylinositol arabinosyltransferase A	Ethambutol	Rv3794	+	4243233	4246517	3285
embB	membrane indolylacetylinositol arabinosyltransferase B	Ethambutol, Isoniazid, Rifampicin	Rv3795	+	4246514	4249810	3297
embC	membrane indolylacetylinositol arabinosyltransferase C	Ethambutol	Rv3793	+	4239863	4243147	3285
ethA	monooxygenase	Ethionamide	Rv3854c	-	4326004	4327473	1470
gidB	glucose-inhibited division protein B	Streptomycin	Rv3919c	-	4407528	4408202	675
gyrA	DNA gyrase subunit A	Fluoroquinolones	Rv0006	+	7302	9818	2517
gyrB	DNA gyrase subunit B	Fluoroquinolones	Rv0005	+	5123	7267	2145
inhA	NADH-dependent enoyl-[acyl-carrier-protein] reductase	Ethionamide, Isoniazid	Rv1484	+	1674202	1675011	810
iniA	isoniazid inductible gene protein A	Ethambutol, Isoniazid	Rv0342	+	410838	412760	1923
iniB	isoniazid inductible gene protein B	Ethambutol, Isoniazid	Rv0341	+	409362	410801	1440
iniC	isoniazid inductible gene protein C	Ethambutol, Isoniazid	Rv0343	+	412757	414238	1482
kasA (fabF1)	3-oxoacyl-[acyl-carrier protein] synthase 1	Isoniazid	Rv2245	+	2518115	2519365	1251
katG	catalase-peroxidase-peroxynitritase T	Isoniazid	Rv1908c	-	2153889	2156111	2223
promoter *mabA*		Isoniazid	-	+	1673300	1673439	140
mabA (fabG1)	3-oxoacyl-[acyl-carrier protein] reductase (mycolic acid biosynthesis protein A)	Ethionamide, Isoniazid	Rv1483	+	1673440	1674183	744
ndh	NADH dehydrogenase	Isoniazid	Rv1854c	-	2101651	2103042	1392
oxyR'	oxidative-stress regulatory gene (pseudogene)	Isoniazid?	Rv2427Ac	-	2725571	2726087	517
pncA	pyrazinamidase/nicotinamidase	Pyrazinamide	Rv2043c	-	2288681	2289241	561
rpoB	DNA-directed RNA polymerase beta chain	Rifampicin	Rv0667	+	759807	763325	3519
rpsL	30S ribosomal protein S12	Streptomycin	Rv0682	+	781560	781934	375
rrl	ribosomal RNA 23S	Aminoglycosides	Rvnr02	+	1473658	1476795	3138
rrs	ribosomal RNA 16S	Aminoglycosides	Rvnr01	+	1471846	1473382	1537
thyA	thymidylate synthase	Para-aminosalicylic acid	Rv2764c	-	3073680	3074471	792
tlyA	cytotoxin|haemolysin	Capreomycin	Rv1694	+	1917940	1918746	807
Promoter *eis**		Kanamycin	-	-	2715332	2715471	139
eis*	N-acetyltransferase	Kanamycin	Rv2416c	-	2714124	2715332	1208
rpsA*	30S ribosomal protein S1	Pyrazinamide	Rv1630	+	1833542	1834987	1445
Promoter *rpsA**		Pyrazinamide	-	+	1833379	1833541	162

Table A.8: List of genomic regions used for resistance prediction. Regions marked with (*) were not sequenced in 1,379 isolates, but are known to be associated with resistance to kanamycin and pyrazinamide. Thus, these strains were assigned a status of 0.5 for variants within these four regions. This allowed the model to learn the contribution of these regions in the remaining 2,222 isolates to antibiotic resistance.



Algorithm	RIF	INH	PZA	EMB	STR	CAP	AMK	MOXI	OFLX	KAN
Kouchaki et. al	0.9808 ± 0.0032	0.9789 ± 0.0038	0.9389 ± 0.0080	0.9625 ± 0.0054	0.9515 ± 0.0056	0.8546 ± 0.0202	0.9137 ± 0.0236	0.9027 ± 0.0296	0.9233 ± 0.0149	0.9249 ± 0.0293
Logistic Regression	0.994 (0.993 - 0.995)	0.989 (0.987 - 0.991)	0.959 (0.955 - 0.963)	0.977 (0.975 - 0.979)	0.939 (0.934 - 0.943)	0.953 (0.948 - 0.958)	0.944 (0.933 - 0.954)	0.905 (0.895 - 0.915)	0.921 (0.902 - 0.941)	0.91 (0.901 - 0.919)
MD-WDNN	0.994 (0.994 - 0.995)	0.988 (0.987 - 0.99)	0.961 (0.958 - 0.964)	0.973 (0.971 - 0.975)	0.935 (0.93 - 0.94)	0.963 (0.958 - 0.968)	0.952 (0.943 - 0.962)	0.914 (0.905 - 0.924)	0.941 (0.931 - 0.952)	0.913 (0.904 - 0.923)

Table A.9: Comparison of performance to prior study. A table containing the AUCs for the best performing model in Kouchaki et al. for each drug and models' performances during cross-validation. The MD-WDNN showed higher performance for 8 of the 10 drugs. For the drugs in which Kouchaki et al. used dimensionality reduction (capreomycin and amikacin), the MD-WDNN showed significantly higher performance.

B

Supplementary Figures

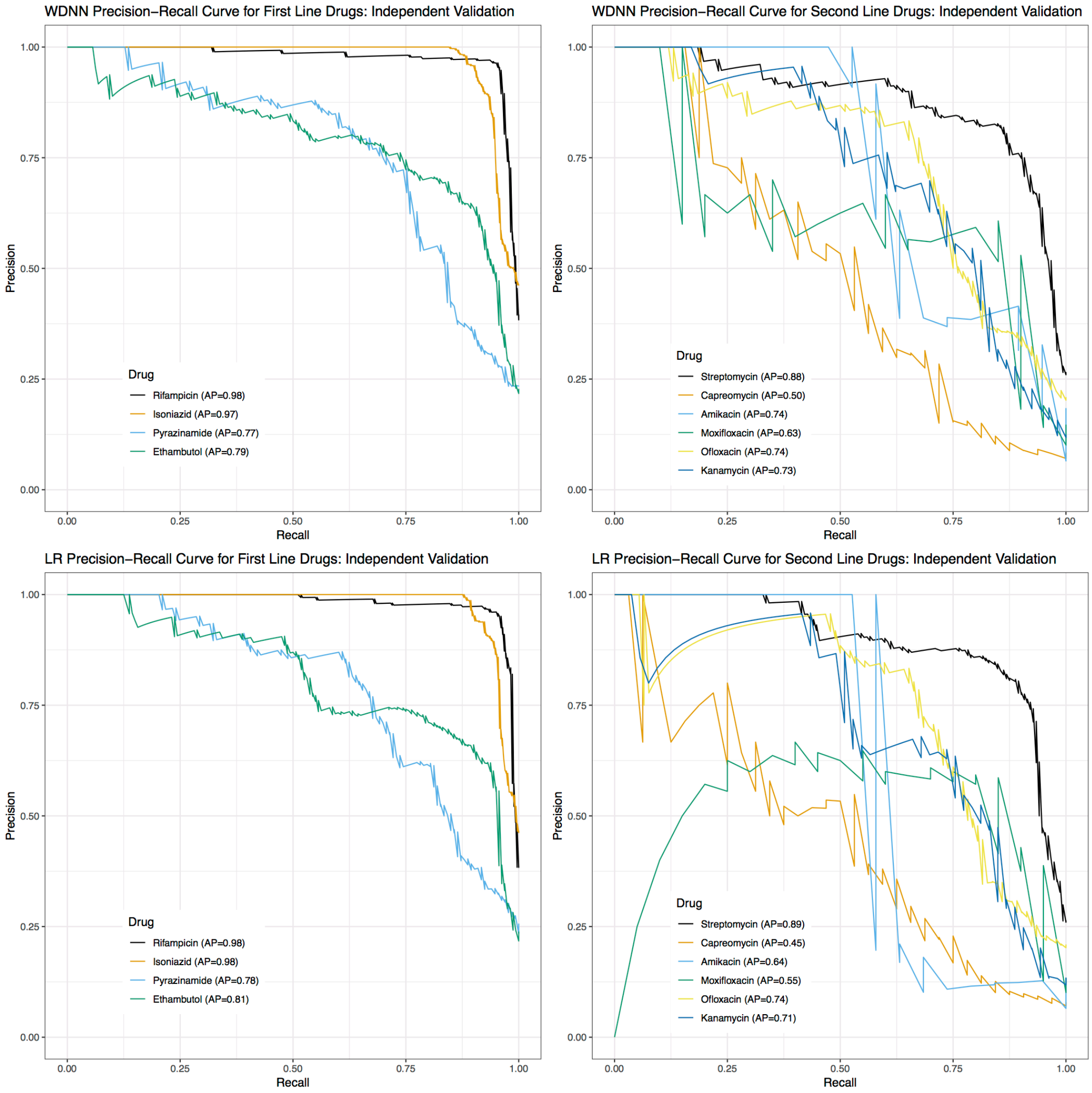


Figure B.1: Precision-recall performance curve of the MD-WDNN and logistic regression on the independent validation set. A precision-recall plot of MD-WDNN (top) and logistic regression (bottom) predictive performance on the independent validation set for first-line (left) and second-line (right) anti-tuberculosis drugs.

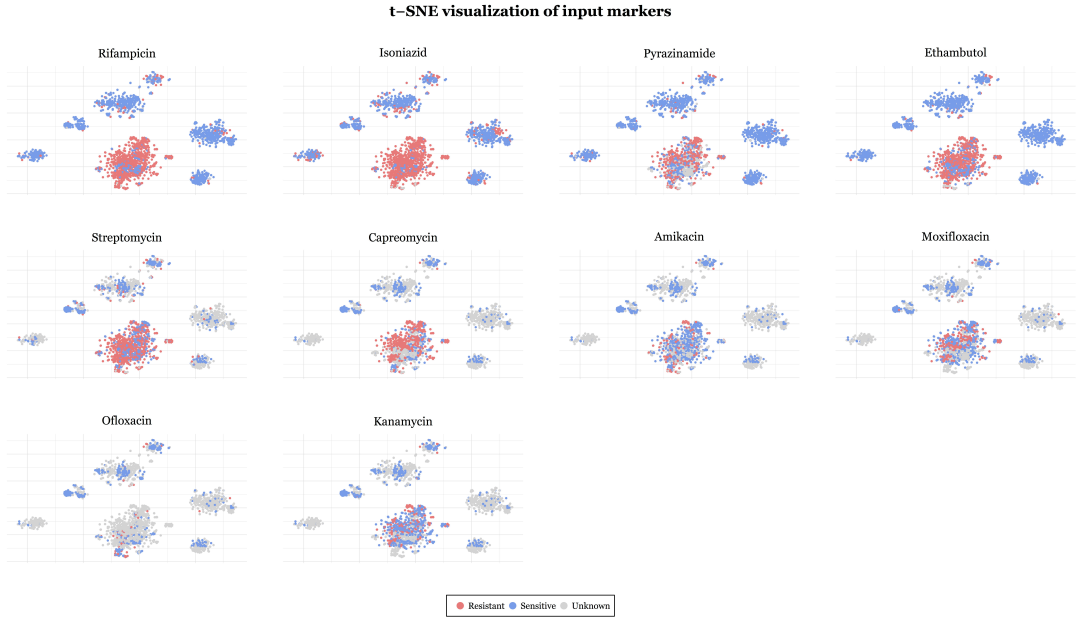


Figure B.2: t-SNE visualization for inputted genetic markers colored by resistance status. The input genetic markers, originally in 222 dimensions, were projected onto two dimensions. The t-SNE plots have the same coordinates as in Figure 4.3. Each point is an MTB isolate, colored according to its resistance status with respect to the corresponding drug. t-SNE on the input genetic markers showed well-defined clusters with little discernable pattern of resistance classification between clusters.

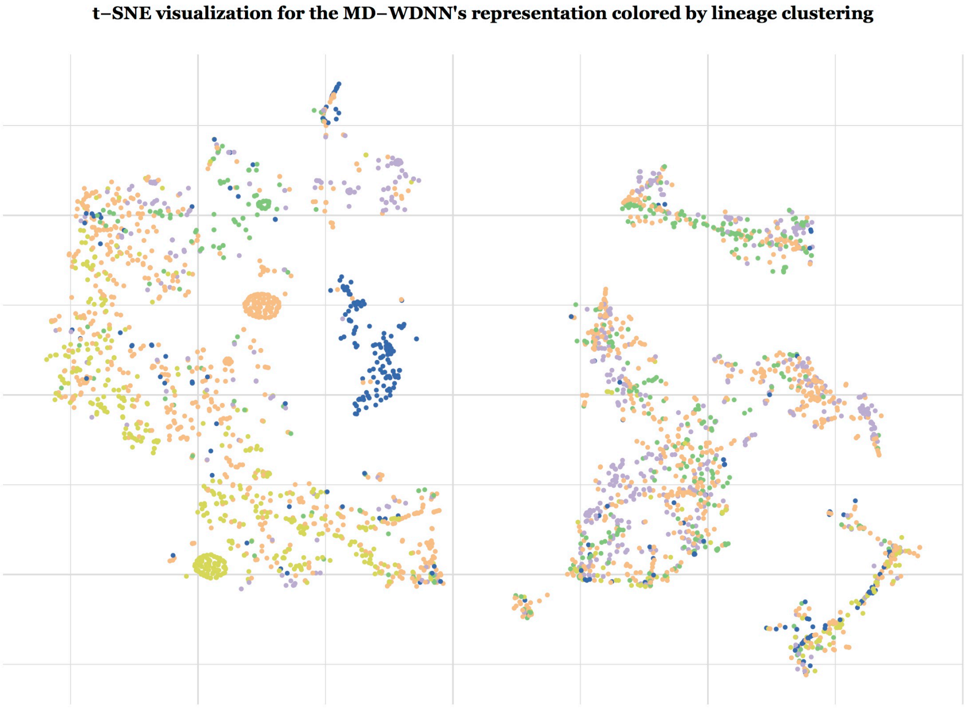


Figure B.3: t-SNE visualization for the final output layer of the MD-WDNN colored by lineage clustering. t-SNE plot with the same coordinates as in Figure 4.2. Each isolate is colored based on the five lineage clusters determined in Figure 3.1, illustrating the diversity of MTB isolates within the MD-WDNN's resistance-susceptibility clustering.

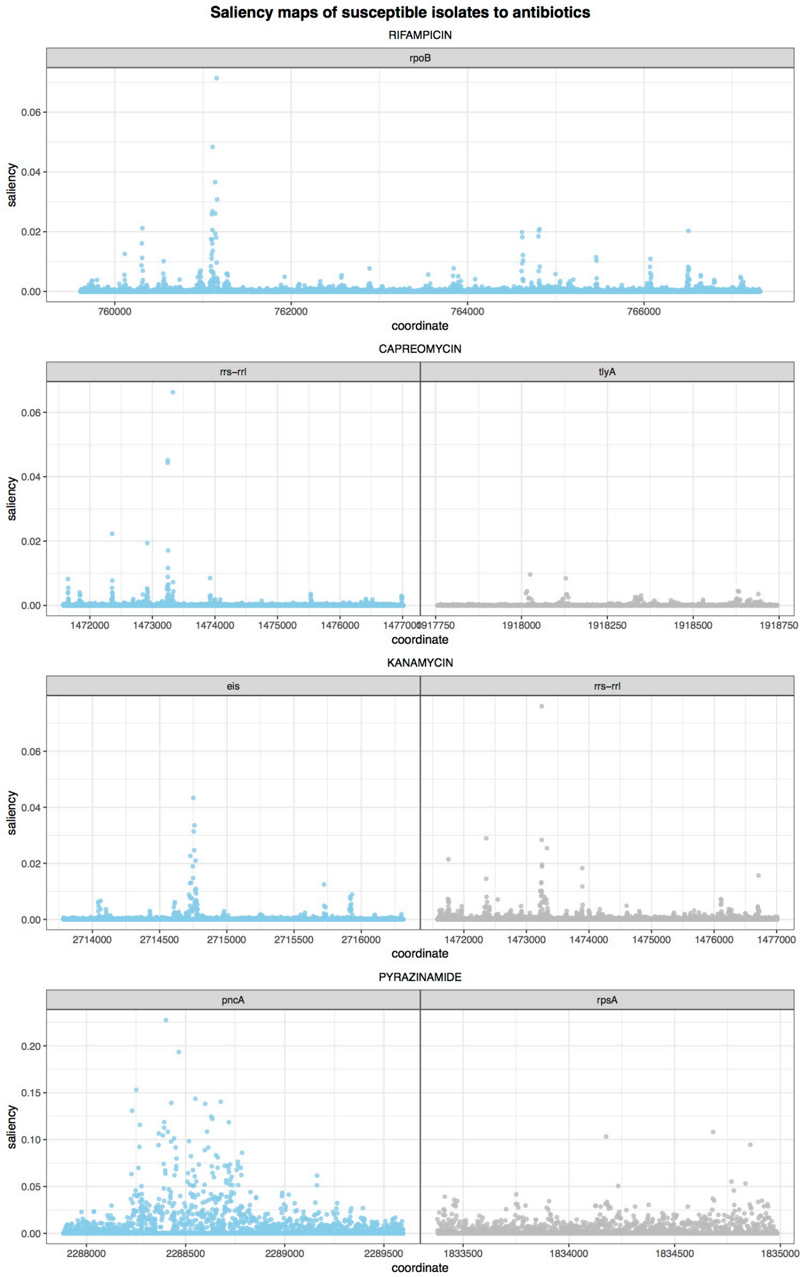


Figure B.4: Relative importance of genomic coordinates to the susceptible phenotype within each antibiotic for the CNN. A visualization showing the relative importance, as measured in a gradient-based calculation of saliency, of each coordinate within the input genetic sequence. The x-axis shows the genomic coordinates relative to the H37Rv strain, which is standard practice within tuberculosis genomics. Any insertions relative to the H37Rv strains are given a fractional coordinate value between the two flanking H37Rv coordinates for ease of visualization and numerical consistency with the reference strain.

References

[1] World Health Organization. Global Tuberculosis Report; 2019.

[2] Klein EY, Van Boeckel TP, Martinez EM, Pant S, Gandra S, Levin SA, et al. Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. Proc Natl Acad Sci. 2018;115(15):E3463–E3470.

[3] World Health Organization. Multidrug and extensively drug-resistant TB (M/XDR-TB) 2010 Global Report On Surveillance and Response; 2010.

[4] Beste DJV, Espasa M, Bonde B, Kierzek AM, Stewart GR, McFadden J. The Genetic Requirements for Fast and Slow Growth in Mycobacteria. PLoS ONE. 2009;4(4):e5349.

[5] Tagliani E, Cabibbe AM, Miotto P, Borroni E, Toro JC, Mansjö M, et al. Diagnostic Performance of the New Version (v2.0) of GenoType MTBDRsl Assay for Detection of Resistance to Fluoroquinolones and Second-Line Injectable Drugs: a Multicenter Study. Journal of Clinical Microbiology. 2015;53(9):2961–2969.

[6] Farhat MR, Sultana R, , Iartchouk O, Bozeman S, Galagan J, et al. Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. Am J Respir Crit Care Med. 2016;194(5):620–630.

[7] Farhat MR, Jacobson KR, Franke MF, Kaur D, Sloutsky A, Mitnick CD, et al. Gyrase Mutations Are Associated with Variable Levels of Fluoroquinolone Resistance in Mycobacterium tuberculosis. J Clin Microbiol. 2016;54(3):727–733.

[8] Mongan AE, Tuda JSB, Runtuwene LR. Portable sequencer in the fight against infectious disease. J Hum Genet. 2020;65(1):35–40.

[9] Walker TM, Kohl TA, Omar SV, Hedge J, Elias CDO, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015;15(10):1193–1202.

[10] Starks AM, Avilés E, Cirillo DM, Denkinger CM, Dolinger DL, Emerson C, et al. Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform. Clin Infect Dis. 2015;61(Suppl 3):S141–146.

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

[12] Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, et al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. N Engl J Med. 2011;364(8):730–739.

[13] Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, et al. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis. Nat Med. 2016;22(12):1470–1474.

[14] Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, et al. Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet. 2013;45(10):1255–1260.

[15] Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 2017;45(D1):D535–D542.

[16] Cheng HT, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, et al. Wide & Deep Learning for Recommender Systems. arXiv. 2016;1606.07792.

[17] Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. NeurIPS. 2012;p. 2951–2959.

[18] Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. AISTATS. 2011;15:315–323.

[19] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. JMLR. 2014;15:1929–1958.

[20] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. PMLR. 2015;37:448–456.

[21] Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE. 2015;10(3):e0118432.

[22] van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–2605.

[23] Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, III CEB, et al. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. Nat Genet. 2017;49(3):395–402.

[24] Farhat MR, Mitnick CD, Franke MF, Kaur D, Sloutsky A, Murray M, et al. Concordance of Mycobacterium tuberculosis fluoroquinolone resistance testing: implications for treatment. Int J Tuberc Lung Dis. 2015;19(3):339–341.

[25] Aldred KJ, Blower TR, Kerns RJ, Berger JM, Osheroff N. Fluoroquinolone interactions with Mycobacterium tuberculosis gyrase: enhancing drug activity against wild-type and resistant gyrase. Proc Natl Acad Sci. 2016;113(7):E839–E846.

[26] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–118.

[27] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402–2410.

[28] Telenti A, Imboden P, andL Matter andK Schopfer FM, Bodmer T, Lowrie D, Colston MJ, et al. Detection of rifampicin-resistance mutations in Mycobacterium tuberculosis. Lancet. 1993;341:647–651.

[29] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–18.

[30] Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics. 2018;34(10):1666–1671.

[31] Kouchaki S, Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics. 2019;35(13):2276–2282.

[32] Moodley R, Godec TR, Team ST. Short-course treatment for multidrug-resistant tuberculosis: the STREAM trials. Eur Respir Rev. 2016;25(139):29–35.

[33] Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13(9):601–612.

[34] Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, et al. Whole-Genome Sequencing for Rapid Susceptibility Testing of M. tuberculosis. N Engl J Med. 2013;369(3):290–292.

[35] Shea J, Halse TA, Lapierre P, Shudt M, Kohlerschmidt D, Roey PV, et al. Comprehensive Whole-Genome Sequencing and Reporting of Drug Resistance Profiles on Clinical Cases of Mycobacterium tuberculosis in New York State. J Clin Microbiol. 2017;55(6):1871–1882.

[36] Votintseva AA, Bradley P, Pankhurst L, del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol. 2017;55(5):1285–1298.

[37] World Health Organization. A roadmap for ensuring quality tuberculosis diagnostics services within national laboratory strategic plans; 2010.

[38] Angra PK, Taylor TH, Iademarco MF, Metchock B, Astles JR, Ridderhof JC. Performance of Tuberculosis Drug Susceptibility Testing in U.S. Laboratories from 1994 to 2008. J Clin Microbiol. 2012;50(4):1233–1239.

[39] Ängeby K, Juréen P, Kahlmeter G, Hoffner SE, Schönd T. Challenging a dogma: antimicrobial susceptibility testing breakpoints for Mycobacterium tuberculosis. Bull World Health Organ. 2012;90(9):693–698.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.