Deep Learning Predicts Tuberculosis Drug Resistance Status from Whole-Genome Sequencing Data

Michael L. Chen[1], Akshith Doddi[2], Jimmy Royer[3], PhD, Luca Freschi[1], PhD, Marco Schito[4], PhD, Matthew Ezewudo[4], PhD, Isaac S. Kohane[1], MD, PhD, Andrew Beam[1†], PhD, Maha Farhat[1,5†*], MD, MSc

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[2]University of Virginia School of Medicine, Charlottesville, VA
[3]Analysis Group Inc.
[4]Critical Path Institute, 1730 E River Rd., Tucson, AZ
[5]Division of Pulmonary & Critical Care, Massachusetts General Hospital, Boston, MA
[†]Denotes equal contribution.

*Corresponding author. E-mail: Maha_Farhat@hms.harvard.edu

One sentence summary: A unified multitask deep learning model can be used to identify multidrug resistant *Mycobacterium tuberculosis* using sequencing data.

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 drug susceptibility but gaps remain for predicting phenotype accurately from genotypic data. Using targeted or whole genome sequencing and conventional drug resistance phenotyping data from 3,601 *Mycobacterium tuberculosis* strains, 1,228 of which were multidrug resistant, we implemented the first multitask deep learning framework to predict phenotypic drug resistance to 10 anti-tubercular drugs. The proposed wide and deep neural network (WDNN) achieved improved predictive performance compared to regularized logistic regression and random forest: the average sensitivities and specificities, respectively, were 92.7% and 92.7% for first-line drugs and 82.0% and 92.8% for second-line drugs during cross-validation. On an independent validation set, the multitask WDNN showed significant performance gains over baseline models, with average sensitivities and specificities, respectively, of 84.5% and 93.6% for first-line drugs and 64.0% and 95.7% for second-line drugs. In addition to being able to learn from samples that have only been partially phenotyped, our proposed multitask architecture shares information across different anti-tubercular drugs and genes to provide a more accurate phenotypic prediction. We use *t*-distributed Stochastic Neighbor Embedding (*t*-SNE) visualization and feature importance analyses to examine inter-drug similarities. Deep learning has a clear role in improving drug resistance predictive performance over traditional methods and holds promise in bringing sequencing technologies closer to the bedside.

Introduction

Tuberculosis (TB) is among the top 10 causes of mortality worldwide with an estimated 10.4 million new incidents of TB in 2015 *(1)*. The growing use of antibiotics in healthcare has led to increased prevalence of drug resistant bacterial strains *(2)*, and the World Health Organization (WHO) estimates that 4.1% of new *Mycobacterium tuberculosis* (MTB) clinical isolates are multidrug-resistant (MDR) (*i.e.* resistant to rifampicin [RIF] and isoniazid [INH]). Furthermore, approximately 9.5% of MDR cases are extensively drug-resistant (XDR) (*i.e.* resistant to one second-line injectable drug, such as amikacin [AMK], kanamycin [KAN], or capreomycin [CAP], and one fluoroquinolone, such as moxifloxacin [MOXI], or ofloxacin [OFLX]) *(1)*. The WHO estimates that 48% of MDR-TB and 72% of XDR-TB patients have unfavorable treatment outcomes, citing the lack of MDR-TB detection and treatment as a global health crisis *(1)*.

Diagnosing drug resistance remains a barrier to providing appropriate TB treatment. Due to insufficient resources for building diagnostic laboratories, fewer than half of the countries with a high MDR-TB burden have modern diagnostic capabilities *(3)*. Even in the best equipped laboratories, conventional culture and culture based drug susceptibility testing (DST) constitutes a considerable biohazard and requires weeks to months before results are reported due to *Mycobacterium tuberculosis*'s slow growth *in vitro (1)*. Molecular diagnostics are now an increasingly common alternative to conventional cultures. The WHO has endorsed three such molecular tests: the GeneXpert MTB/RIF a rapid RT-PCR based diagnostic test assay that detects RIF resistance, the Hain line probe assay (LPA) that tests for both RIF and INH resistance, and the Hain MDRTB*sl* an LPA that tests for resistance to second-line injectable drugs and fluoroquinolones *(1)*. The LPAs recently approved by the WHO have seen moderate sensitivities, such as a range from 63.7% to 94.4% for second-line injectable drugs and fluoroquinolones *(4–6)*. However, current diagnostic approaches face challenges. First, these methods have limited sensitivity because they rely on a few genetic loci, ranging between 1-6 loci per test *(6, 7)*. Second, they do not detect most rare gene variants of the targeted loci, especially insertion and deletions and variants in promoter regions *(8)*. Third, current molecular tests only detect resistance to five anti-tubercular drugs rather than the full panel. Fourth, they do not account for variables such as genetic background and gene-gene interactions despite good evidence for this for several drugs including rifampicin, ethambutol and fluoroquinolone from allelic exchange experiments *(9–11)*. The limited scope of these tests suggests the need for a comprehensive drug susceptibility test.

An alternative to targeted mutation detection methods is whole genome sequencing, which captures both common and rare mutations involved in drug resistance. Past studies utilizing whole genome sequencing have shown a wide range of performance, with sensitivities for first-line drugs ranging from 54% to 98% *(8, 12, 13)*. Second-line injectable drugs and fluoroquinolones had lower sensitivities, most of which were between 30% and 96% *(8, 12, 13)*. We hypothesize that the limited predictive performance of anti-tubercular drugs outside of first-line drugs could be improved using a large dataset enriched for resistance to second-line drugs and a more complex model.

Deep learning models have become a powerful tool for many classification tasks. Modern deep neural networks have achieved state-of-the-art performance in image recognition *(14)*, speech recognition *(15)*, and natural language processing *(16)*. Researchers in medicine have begun to translate these approaches for use in personalized clinical care. Deep 'convolutional' neural networks have been used to in identifying diabetic retinopathy *(17)* and classifying skin cancers *(18)*. Deep learning applications in computational biology and bioinformatics have also been successful, such as in predicting RNA-binding protein sites *(19)*, inferring target gene expression from landmark genes *(20)*, and identifying biomarkers for predicting human chronological age *(21)*. The flexibility of deep learning architectures has allowed for a range of successful applications in clinical tasks, biomedicine, molecular genomics, and other fields.

We demonstrate here an improved predictive tool to evaluate drug resistance for 10 anti-tubercular drugs using a novel multitask 'wide and deep' neural network (WDNN) framework *(22)*. In contrast to previously reported single task models, our multitask framework that predicts the full resistance profile simultaneously allows the anti-tubercular drugs to share resistance pathway information from the phenotypes of other drugs and incorporates prior knowledge that drug resistance can be caused by both direct genotype-phenotype relationships as well as epistatic effects *(9–11)*. We use the deep learning architectural features to evaluate the relative influence of genomic markers, provide insights into the biological basis for our model, and gain a deeper understanding of the relationships amongst the 10 anti-tubercular drugs.

Results

Data Processing

The pooled data from the WHO network of supranational reference laboratories and the ReSeqTB knowledgebase *(8, 23)* used in training the initial model included 3,601 MTB isolates. All of the anti-tubercular drugs had a higher proportion of susceptible isolates compared to resistant isolates, ranging from 53.0% to 88.1% susceptible for the different drugs. Ofloxacin was tested in the smallest number of isolates at a total of 739. All other drugs were tested in at least 1,204 isolates, with rifampicin tested in 3,542 isolates and isoniazid in 3,564 isolates (Supplementary Table S1).

The independent validation set contained 792 MTB isolates, with 198 to 736 of these isolates tested for each of the 10 drugs (Supplementary Table S2). Because ciprofloxacin had limited phenotypic availability in the independent validation set and predictive performance could not be validated, we did not include performance for ciprofloxacin resistance.

We found 6,342 different insertions, deletions, and single nucleotide polymorphisms (SNPs) in 30 promoter, intergenic, and coding regions of the MTB isolates' genomes. Of these variants, 156 were present in at least 30 of the 3,601 isolates and were used as predictors. Of the 3,445 variants found in fewer than 30 isolates, we aggregated the variants into 141 derived categories (see Methods) and used 56 derived categories, those present in at least 30 isolates, as predictors. The final model used 222 total predictors in training and subsequent analyses.

Evaluation of MTB isolate diversity

Sequence data from 33 genetic lineage markers (Supplementary Table S3) were available in all 3,601 isolates and were used to assess isolate diversity *(12)*. Overall, the isolates showed considerable diversity with a low pairwise genetic distance ranging from 0 to 3.87. The isolates fell into five well-defined genetic clusters. The isolate clusters, shown in Figure 1 and colored as indicated, contained 632 (Euro-American LAM sub-lineages; purple), 1,501 (other Euro-American sub-lineages; orange), 331 (Indo-Oceanic, *Mycobacterium africanum*, and other animal lineages; blue), 643 (Central Asian; yellow), and 494 (East Asian; green) isolates, respectively. Overlying the lineage clusters and t-SNE coordinates (Supplementary Figure S1) confirmed that the multitask WDNN phenotyping was not biased by lineage related variation.

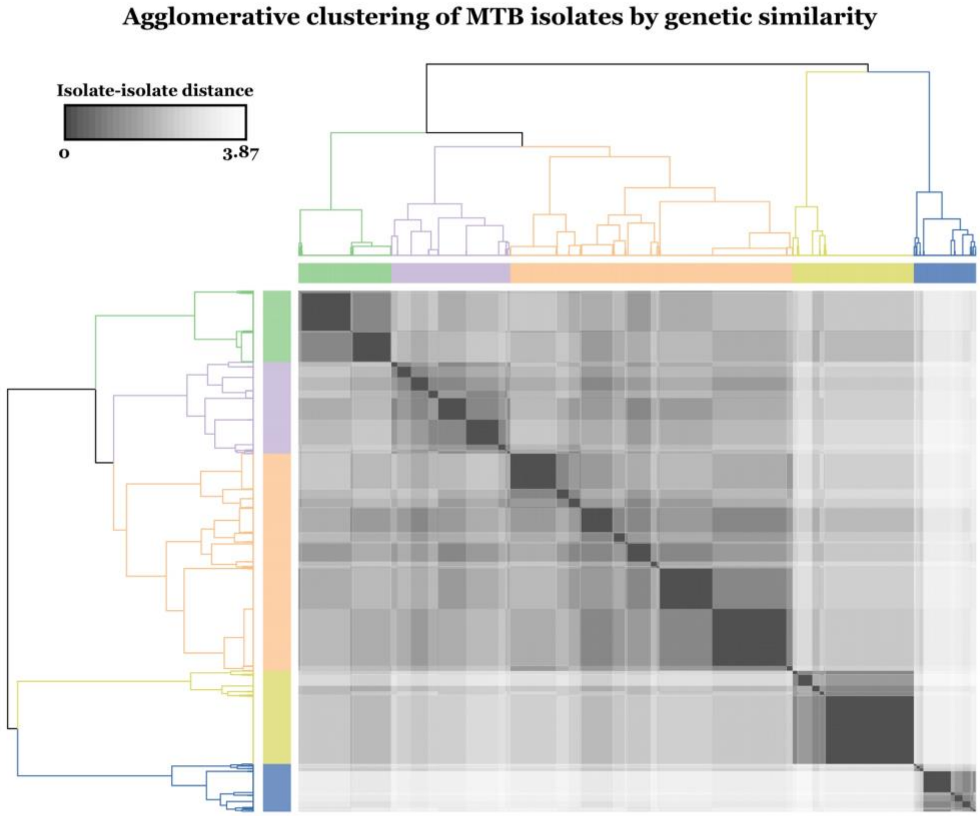

__Figure 1: Agglomerative clustering of MTB isolates by genetic similarity.__ We used known lineage-defining mutations to calculate isolate-isolate Euclidean distances, which is shown in the heat map. Using these distances of the lineage-defining mutation vectors between isolates, we applied Ward's method of hierarchical clustering to construct the dendrogram and determine the five lineage clusters.

Comparison of model predictive performance

A comparison of model sum of sensitivity and specificity performances across the 10 anti-tubercular drugs is shown in Figure 2. The multitask WDNN, a single task WDNN (trained for each drug individually), random forest, and regularized logistic regression were trained on the full set of predictors, whereas the multilayer perceptron (MLP) was trained only using predictors in genes known to be determinants of resistance for each drug. Using five-fold cross validation, the average sensitivities and specificities, respectively, for rifampicin and isoniazid were 97.1% and 95.9% (multitask WDNN), 95.6% and 95.4% (random forest), 96.7% and 95.7% (regularized logistic regression), 96.3% and 94.3% (preselected mutations MLP), and 97.2% and 95.2% (single task WDNN). The model performance trends were similar for the other eight anti-tubercular drugs. The average sensitivities and specificities, respectively, of the multitask WDNN for the different drugs were 89.8% and 90.6% (other first-line drugs: PZA, EMB, STR), 84.5% and 93.9 (second-line injectable drugs: CAP, AMK, KAN), and 78.2% and 91.1% (fluoroquinolones: OFLX and MOXI).

Using an independent validation set, the models showed similar trends in performance as in cross-validation. The average sensitivities and specificities, respectively, for rifampicin and

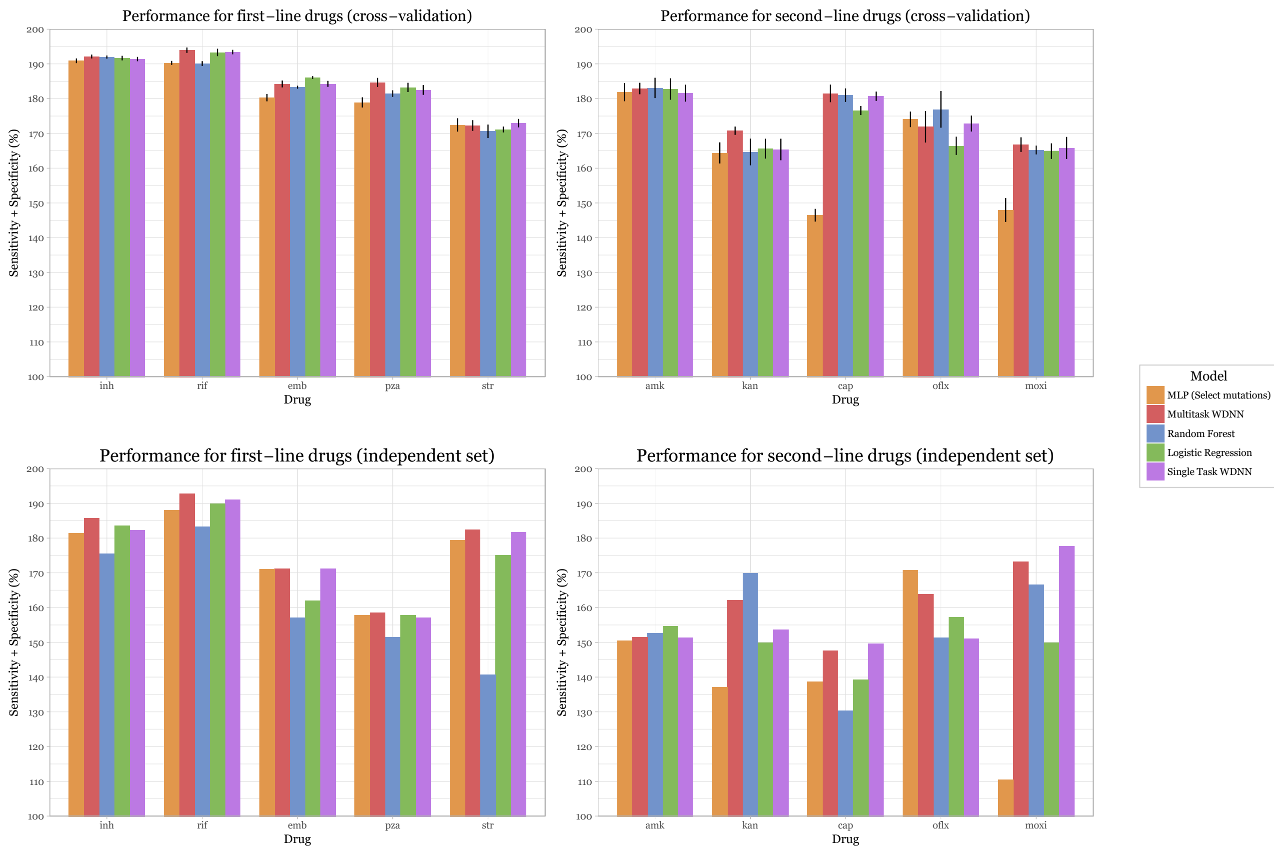

***Figure 2: Tuberculosis drug resistance predictive performance of the multitask WDNN and baseline models**. A bar plot of sensitivity + specificity performance across all four models during cross-validation (top) and on the independent validation set (bottom). The multitask WDNN, single task WDNN, random forest, and logistic regression models were trained on the full set of predictors, while the single task MLP was trained on preselected mutations. Thresholds were chosen for each model on the training data to maximize sensitivity + specificity with the condition that specificity is at least 90%. Individual sensitivity and specificity performance for all five models is available in the supplementary materials.*

isoniazid were 93.7% and 95.6% (multitask WDNN), 80.5% and 98.9% (random forest), 87.7% and 99.0% (regularized logistic regression), 90.9% and 93.8% (preselected mutations MLP), and 91.7% and 95.0% (single task WDNN). For the different subgroups of drugs, the multitask WDNN had average sensitivity and specificity performance of 78.4% and 92.3% (other first-line drugs), 57.9% and 95.9% (second-line injectable drugs), and 73.2% and 95.4% (fluoroquinolones).

Compared to the other models, the multitask WDNN achieved a higher sum of specificity and sensitivity for 9 of the 10 drugs (random forest), 9 of the 10 drugs (regularized logistic regression), 8 of the 10 drugs (preselected mutations MLP), and 7 of the 10 drugs (single task WDNN) during cross-validation. On the independent validation set, the multitask WDNN achieved a higher sum of specificity and sensitivity for 8 of the 10 drugs (random forest), 9 of the 10 drugs (regularized logistic regression), 9 of the 10 drugs (preselected mutations MLP), and 7 of the 10 drugs (single task WDNN). Details about individual sensitivity and specificity performance for the models are provided in Supplementary Tables S4 and S5.

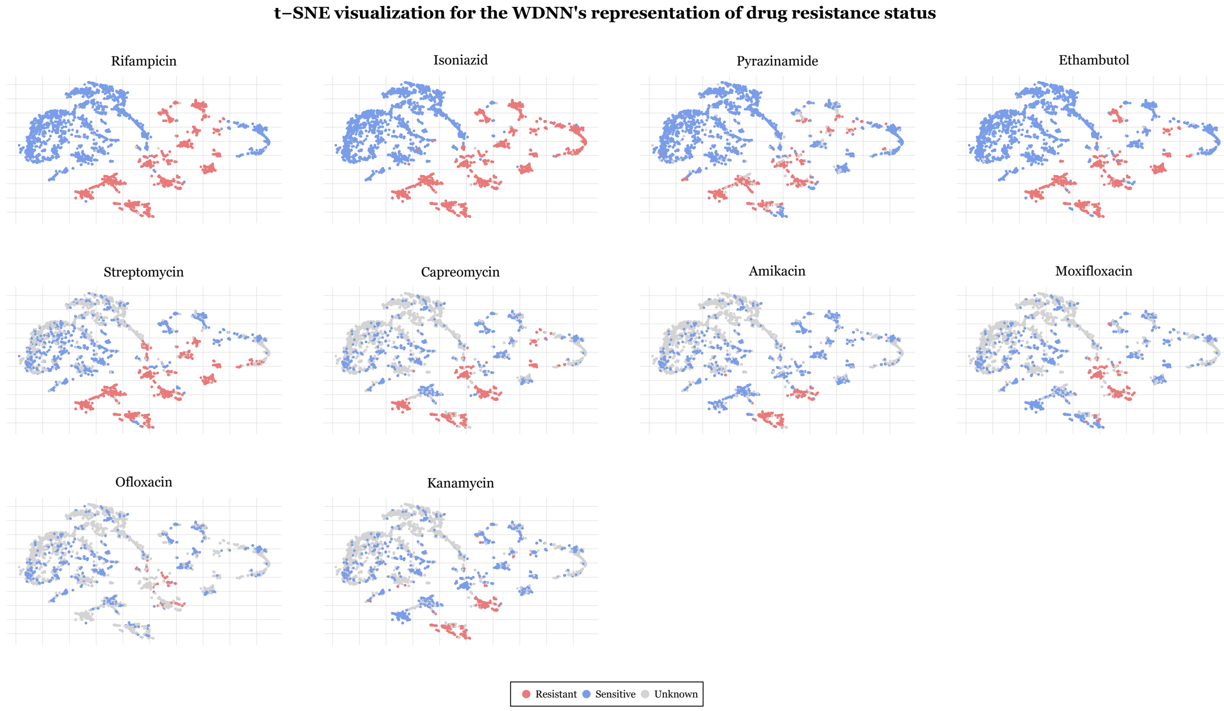

Figure 3: t-SNE visualization for the final output layer of the multitask WDNN. *The final layer predictions, originally in 11 dimensions, were projected onto two dimensions. Each point is an MTB isolate, colored according to its resistance status with respect to the corresponding drug.*

MTB isolate visualization using t-SNE

A popular way to visualize the various high-dimensional components of a deep learning model is the *t*-distribution stochastic neighborhood embedding (*t*-SNE) method, which is a nonlinear dimensionality reduction technique *(24)*. To visualize the multitask WDNN's integration of genetic features into a prediction, we applied *t*-SNE to the multitask WDNN predictions. Figure 3 shows the two-dimensional *t*-SNE projection colored by the MTB isolate resistance phenotype by drug. This demonstrated clear separation by the model between resistant and sensitive isolates, consistent with our measurements of high model sensitivity and specificity. The *t*-SNE plots also demonstrates the multitask WDNN's ability to classify resistance across multiple drugs, separating them into nested groups of pan-susceptible isolates, followed by mono-INH resistant isolates, multidrug resistant isolates, pre-XDR isolates, and XDR isolates, which is consistent with the order of administration of the drugs clinically as well as the usual order of MTB drug resistance acquisition *(25)*. The second-line injectable drugs, AMI, CAP, and KAN, also show similarly-classified clusters, highlighting the well-known moderate level of cross resistance between them. We also observe this among the fluoroquinolones despite the fact that fewer isolates were tested for resistance to these agents *(26)*.

Importance of MTB genetic variants to drug resistance

All 222 predictors were tested for importance to resistance to each of the 10 drugs through a permutation test as described in the methods section. The first-line anti-tubercular

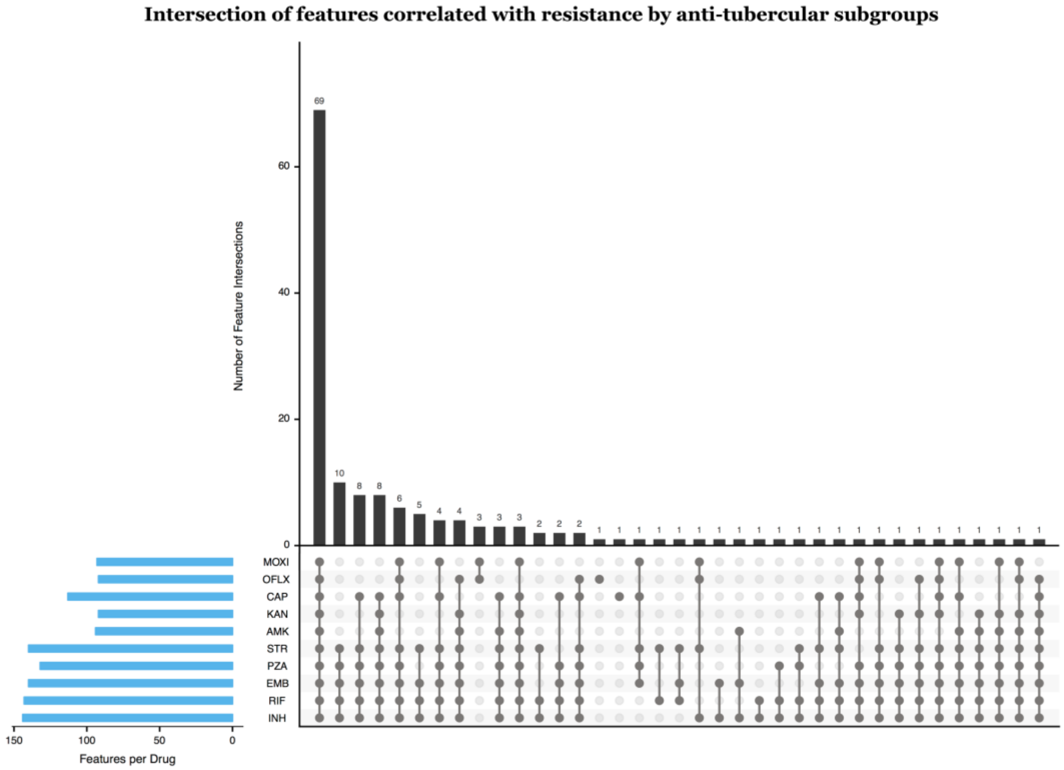

Figure 4: Intersection of predictors correlated with resistance by anti-tubercular drug subgroups. *We permuted the resistance labels and calculated the distribution of the difference, P(isolate is resistant | mutation is present) – P(isolate is resistant | mutation is absent). We show the number of mutations per subgroup of drugs ordered from most to least mutations per subgroup. Number of significant predictors per drug is also shown.*

drugs had the largest numbers of significant 'resistance predictors': rifampicin (143 predictors), isoniazid (144 predictors), pyrazinamide (132 predictors), ethambutol (140 predictors), and streptomycin (140 predictors).

Figure 4 illustrates the number of significant predictors per drug and the predictor intersections among different drug subsets. There were 37 drug subsets that shared at least one resistance predictor. The largest subset was of 10 anti-tubercular drugs that shared 69 resistance predictors. 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 MTB isolates acquire resistance to first-line drugs before second-line drugs *(25)* and indicates that the multitask 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, which is expected given that fluoroquinolones have a mechanism of action that differs from those of first-line and second-line drugs *(27)*.

Discussion

A few prior studies have utilized algorithmic or machine learning methods using MTB genomic data to account for the complex relationship between genotype and drug resistance *(8, 12, 13, 28)*. We demonstrate here that the multitask WDNN approach outperforms our previously reported random forest model *(8)*. Compared to one study that used a direct association (DA) algorithm, the multitask model presented here offers improvement in sensitivity and specificity for the majority of drugs when prediction is attempted on all isolates, including those with rarer and not previously observed variants *(12)*. One study used single-task machine learning, demonstrating the validity of this approach for identifying MDR and XDR-TB, but were limited by the use of a dataset with a low number of MDR isolates (81) and even lower numbers of isolates resistant to drugs other than RIF and INH (ranging from 19 to 59), raising concerns about generalizability *(13)*.

Our model has several novel features which are important to its success. The multitask structure allows drugs which have less phenotypic data to borrow information about resistance pathways from drugs that have higher numbers of phenotyped isolates. Additionally, the wide and deep structure allows us to include prior information about the genetic etiology of MDR and XDR, as it is known that both individual markers and gene-gene interactions confer resistance *(9–11)*. 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. Our deep learning model is the first multitask tool to our knowledge that predicts resistance for 10 anti-tubercular drugs simultaneously with state-of-the-art performance.

Multitask architectures in deep learning have not been used widely in pharmaceutical and drug-related industries due to many barriers, including the difficulty of implementing a high-quality deep multitask network *(29)*. However, past multitask deep learning algorithms have seen success over traditional single task baseline models, such as in applications to drug discovery and studying gene regulatory networks *(29–31)*. In addition, multitask neural networks have been shown to have larger performance gains over single task models when using smaller datasets *(32, 33)*. We directly compared performance of the multitask and single task wide and deep neural networks, showing improvements in sensitivity and specificity using the multitask architecture.

The increased predictive performance of the multitask WDNN over the single task preselected mutations MLP may arise from a number of possible explanations. First, phenotypic resistance data that was highly available in our dataset for certain drugs (*i.e.* RIF, INH, PZA, and EMB) served as a direct indicator for resistance to second-line injectables and fluoroquinolones. This explanation is unlikely, as our *t*-SNE analysis shows clustering patterns specific to second-line injectable drugs and fluoroquinolones, and the validated model specificity for these drugs was robust. Second, mutations that do not necessarily confer resistance to particular drugs may be indicative of other genomic predictors, thereby serving as a reliable predictor for resistance. Because of the large intersection of mutations (Figure 4) for all anti-tubercular drugs, it is likely that this explanation plays a role in the performance differences. The correlative effect of mutations can be treated as a positive feature in the multitask architecture due to the difficulty of acquiring comprehensive genomic data. On the other hand, the potential lack of causation also requires care when using the predictive model, which could account for the increased performance of the preselected mutations MLP over the multitask WDNN in detecting ofloxacin resistance. Third, there may exist mutations that are not yet known to confer resistance to particular anti-tubercular drugs but were captured by the multitask WDNN thereby improving performance.

Understanding the improved performance of our wide and deep neural network is a difficult task due to the architectural complexity and lack of visualization tools in deep learning *(34, 35)*. Our *t*-SNE visualization demonstrated the multitask model's ability to capture the biologically and clinically expected order of resistance acquisition and cross resistance providing further evidence to support the use of this prediction architecture *(25, 26, 36)*. The multitask WDNN's drug resistance classifications for all isolate–drug pairs allowed us to visualize isolate clustering through *t*-SNE even where phenotypic data for isolate–drug pairs were not available.

Our evaluation of predictor importance found significant groupings in drug subsets that we would expect based on prior knowledge of the drug mechanisms. We had a significant intersection subset including only first-line and second-line injectable drugs, one subset with only first-line drugs, and one subset including only fluoroquinolones. The high number of distinct subgroups of drugs reflects the complex decision process of the multitask WDNN but gives evidence for a predictive approach consistent with previously reported understanding of drug resistance acquisition. Overall, developments in deep learning visualization tools and techniques are needed for understanding drug resistance acquisition and ultimately allow for improved deep learning models with improved predictive performance.

The translation of our deep learning approach is also function of advancements in whole genome sequencing and accessibility to more MTB isolate data. Improvements in whole-genome sequencing technologies have significantly reduced costs *(37)*, allowing for more routine whole genome sequencing in MTB isolates *(38)*. The prediction time for MTB drug resistance depends primarily on the sequencing turnaround time, which is significantly shorter than phenotypic susceptibility testing *(39)*. In addition, as more routine sequencing increases the amount of MTB isolate data, our deep learning model can be rapidly updated as the datasets become accessible. We expect that as more data are incorporated, the sensitivity and specificity gap in second-line injectable drugs and fluoroquinolones will become smaller.

We acknowledge some limitations of our study. 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 *(40)*. However, we 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 *(8, 23)*. Beyond technical or laboratory limitations in testing, certain resistance mutations, especially for ethambutol and second-line drugs, may result in minimum inhibitory concentrations (MIC) very close to the clinical testing concentration, which may result in lower sensitivity and specificity *(41)* when predicting a binary resistance phenotype. The use of MIC data for building future learning models may help circumvent this. Second, we only included mutations that occurred in >0.8% (30 of 3,601 isolates) individually or when aggregated with other rare variants in the same gene or intergenic region. Although we 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 4 major genetic lineages. Third, we did not include third-line anti-tubercular drugs such as cycloserine or para-aminosalicylic acid due to the lack of phenotypic data.

In summary, we presented a new deep learning architecture to identify the resistance of MTB isolates to 10 anti-tubercular drugs. The wide and deep neural network achieved state-of-the-art performance on a large, aggregated TB dataset, demonstrating the efficacy of deep learning as a diagnostic tool for MTB drug resistance. The WDNN represented the first multitask model to our knowledge that incorporated a high number of genotypic predictors known to be important to determining resistance for one or more included drugs. Further work identifying the key processes of deep learning will not only allow for improved predictive performance but may also give us a greater understanding of the biological mechanisms underlying drug resistance in MTB isolates.

Materials and Methods

Overview of the Study Design

MTB targeted sequence and antibiotic resistance data from a sample enriched in first and second-line antibiotic resistance *(8)* was pooled with public whole genome sequence and resistance data for training of the prediction model. Model validation was performed on an independent set of public whole genome sequences for which phenotypic resistance data was available. The validation dataset was a convenience dataset not preselected based on antibiotic resistance or strain lineage and diversity distribution. We evaluated MTB isolate diversity through hierarchical clustering and using lineage-defining mutations in the drug resistance loci, as assessed by Walker *et al. (12)*. In order to predict drug resistance for each isolate, we built a unified wide and deep neural network to predict phenotypic status for all drugs simultaneously. We compared our model to baseline machine learning models (random forest and regularized logistic regression). We built a single-task MLP trained on mutations known to be resistance-determining for each drug to evaluate the impact of training on the full genome sequence. We visualized the multitask WDNN's final phenotypic representation in 2-dimensional *t*-SNE plots, and evaluated the importance of genetic variants to resistance through permutation testing.

Data Description

Sequence data: The training dataset consisted of 1,379 MTB isolates that underwent sequencing using molecular inversion probes that targeted 28 preselected antibiotic resistance genes and promoter regions, with 100 bases flanking both ends of each region *(8)*. This sequence data was pooled with 2,222 additional MTB whole genome sequences curated by the ReSeqTB knowledgebase, which maintains a public data sharing platform (www.reseqtb.org) curating genotypic and phenotypic data of WHO-endorsed *in vitro* diagnostic assays for MTB *(23)*. The validation dataset of 792 MTB isolates was obtained by pooling additional data from ReSeqTB, without overlap with the training set, and other MTB whole genome sequences and phenotype data curated manually from the following references *(28, 42–44)*.
Antibiotic resistance phenotype data: All isolates included underwent culture based antibiotic susceptibility testing to two or more drugs at WHO approved critical concentrations and met other quality control criteria as detailed in *(8)*. The pooled phenotype data included resistance status for eleven drugs: first-line drugs (rifampicin, isoniazid, pyrazinamide, ethambutol, and streptomycin); second-line injectable drugs (capreomycin, amikacin, and kanamycin); and fluoroquinolones (ciprofloxacin, moxifloxacin, and ofloxacin). Phenotypic data was classified as resistant, susceptible, or not available.

Variant calling

We used a custom bioinformatics pipeline to clean and filter the raw sequencing reads. We aligned filtered reads to the reference MTB isolate H37Rv and included in the analysis variants called by Stampy 1.0.23 *(45)* and Platypus 0.5.2 *(46)* using default parameters. Genome coverage was assessed using SAMtools 0.1.18 *(47)* and read mapping taxonomy was assessed using Kraken *(48)*. Strains with a coverage of less than 95% at 10x or more in the regions of interest (Supplementary Table S6), or that had a mapping percentage of less than 90% to *Mycobacterium tuberculosis* complex were excluded. Further, regions of the remaining genome not covered by 10 regions or more in at least 95% of the isolates were filtered out from the analysis. In the remaining regions, variants were further filtered if they had a quality of <15, purity of <0.4 or did not meet the PASS filter designation by Platypus.

Building the predictor set of features

Because 1,379 of the 3,601 of the MTB isolates in the training set underwent targeted sequencing only, we restricted the resistance predictors to variants in the regions targeted in these isolates (Supplementary Table S6). Since the *eis* and *rpsA* genes and promoters were recently determined to be associated with kanamycin and pyrazinamide resistance respectively *(49, 50)*, we added mutations in the *eis* and *rpsA* regions into our set of predictors. For those isolates with missing genotype data, we used a status of 0.5 for the missing mutations.

The predictors included in the neural network consisted of two groups. In the first group, each mutation was considered a predictor and its status was binary (either present or absent). For the second group, we created 'aggregate' categories by grouping the rarer mutations (present in <30 isolates) by gene locus (coding, intergenic and putative promoter regions). For each coding region, we split the variants by type into three groups: single nucleotide substitution (SNP), frameshift insertion/deletion or non-frameshift insertion/deletion. For each non-coding region, we split the variants by type into two groups: insertions/deletion or single nucleotide substitution). We used individual and 'aggregate' predictors found in at least 30 MTB isolates to make our final set of predictors.

Evaluation of MTB isolate diversity

We identified lineage-defining variants as assessed in a 2015 study by Walker *et al. (12)*. The genetic-lineage similarity between each pair of isolates was computed as the Euclidean distance between the two corresponding lineage-defining mutation vectors. We applied Ward's method of hierarchical clustering on the resultant distance matrix *(51)* to group the isolates and displayed the isolate-isolate Euclidean distance matrix based on the lineage-defining variants in a heat map. We used *hclust* in the R stats 3.4.2 package to perform hierarchical clustering. Each group was mapped back to the recognized MTB lineage classification by matching the expected pattern of SNPs in Walker *et al. (12)*.

Multitask and Single Task Wide and Deep Neural Network Model

Wide and deep neural networks (WDNN) marry two successful models, logistic regression and deep multilayer perceptrons (MLP), to leverage the strengths of each approach. In WDNNs, a 'wide' logistic regression model is trained in tandem with a 'deep' MLP and the two models are merged in a final classification layer, allowing the network to learn useful rules directly from the raw data and higher level nonlinear features. For genomic data, the logistic regression portion of network can be thought of as modeling the additive portion genotype-phenotype relationship, while the MLP models the nonlinear or epistatic portion. We implemented a wide and deep neural network *(22)* with two hidden layers with ReLU activations *(52)*, dropout *(53)*, and L1 regularization (Figure 5). The network was trained via stochastic gradient descent using the Adam optimizer.

Traditionally, dropout occurs only during training while no dropout occurs during test time *(53)*. However, recent advancements have shed light on dropout from a Bayesian

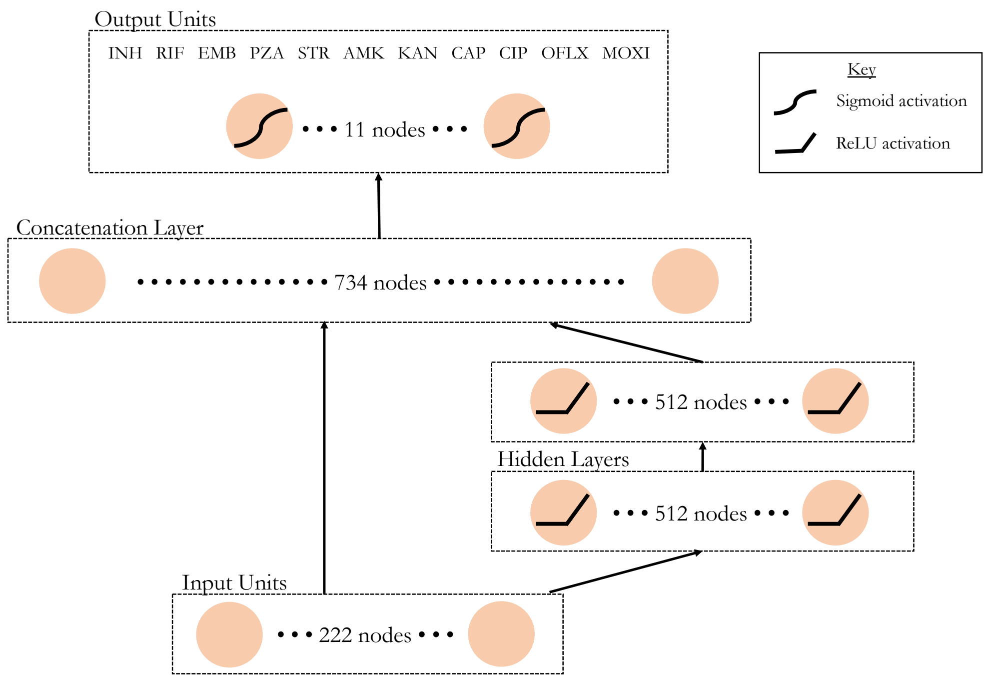

Figure 5: A schematic of the wide and deep neural network architecture. *Data flows from bottom to top through the wide (left) and deep (right) paths of the neural network. Nonlinear transformations, where applied, are depicted on the corresponding nodes. Each of the 11 nodes in the output layer represents resistance status predictions in all MTB isolates for one of the 11 anti-tubercular drugs.*

perspective, and have shown that averaging predictions from multiple dropout masks can reduce variance and improve predictive performance *(54)*. This is often referred to as "Monte Carlo (MC) dropout". Our wide and deep neural network (WDNN) included dropout during both training and test time, and our final predictions were an average of 100 MC dropout samples. L1 regularization was applied on the wide model (which is equivalent to the well-known 'LASSO' model) *(55)*, the hidden layer of the deep model, and the output sigmoid layer.

The multitask WDNN was trained simultaneously on resistance status for all 11 drugs, including ciprofloxacin. Each of the 11 nodes in the final layer represented one drug and outputted the probability that the MTB isolate was resistant to the corresponding drug. We constructed a single task WDNN with the same architecture as the multitask model except for the structure of the output layer, which predicts for one drug.

The multitask WDNN utilized a loss function that is a variant of traditional binary cross entropy. Our dataset had missing resistance status for some drugs in the MTB isolates, so we implemented a loss function that did not penalize the model for its prediction on drug-isolate pairs for which we did not have phenotypic data. Due to imbalance between the susceptible and resistant classes within each drug, we adjusted our loss function to upweight the sparser class according to the susceptible-resistant ratio within each drug. Thus, the final loss function was a class-weight binary cross entropy that masked outputs where the resistance status was missing.

Baseline Models

In addition to the multitask and single task wide and deep neural networks, we implemented three other classification models – a single task random forest, a single task regularized logistic regression, and a single task multilayer perceptron (MLP with MC dropout) with preselected predictors based on prior biological knowledge of drug resistance mechanisms *(8)*. The single task MLP was used as a baseline to identify drugs for which model performance benefited from predictors not already known to affect the drug resistance.

Training and Model Evaluation

The multitask WDNN, single task WDNN, random forest, and regularized logistic regression classifiers were trained on predictors in the dataset present in at least 30 MTB isolates. The single task MLP was trained on mutations based on preselected genes, as described above. A single task MLPs was trained accordingly for each drug with different subsets of predictors.

We used five-fold cross validation to train the models and evaluate performance. The single task WDNN, single task MLP, random forest, and regularized logistic regression models were stratified by class label to address imbalances between resistance and susceptible classes, as they were all single task classifiers. Model performance was validated through an independent validation set.

We reported specificity and sensitivity for the all the models. The probability threshold was chosen to maximize the sum of specificity and sensitivity with the condition that specificity is at least 90% on the training data and applied to the validation data. 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. During five-fold cross-validation, the mean and standard error of specificity and sensitivity were reported based on validation set results across the five folds.

MTB isolate visualization using t-SNE

We examined the final output layer of the multitask WDNN using *t*-distributed Stochastic Neighbor Embedding (*t*-SNE), a method for visualizing data with high dimensionality *(24)*. The final layer weights, originally in 11 dimensions, were extracted from the multitask WDNN and projected onto two dimensions. Each point represented one MTB isolate and was colored based on its phenotypic status for each drug.

Importance of MTB genetic variants to drug resistance

We examined predictor importance to resistance by analyzing the prediction outputs of the multitask WDNN and the presence or absence of mutations through a permutation test. We permuted the resistance labels and calculated the distribution of following difference:

$$P(isolate\ is\ resistant \mid mutation\ is\ present) - P(isolate\ is\ resistant \mid mutation\ is\ absent)$$

where P(isolate is resistant | mutation is present) is the WDNN's outputted probability of resistance for a given mutation. We then compared the actual differences with the permuted differences. The sampling distribution included 100,000 randomized permutations per mutation and the actual differences were evaluated at a significance level of $\alpha = 0.05$ corrected for multiple comparisons. We conducted the permutation test for each predictor (mutations or derived categories) that was present in at least 30 MTB isolates. We focused on the mutations and derived mutation categories that were correlated with resistance to anti-tubercular drugs.

Implementation Details

Our multitask and single task wide and deep neural network implementations used the Keras 1.2.0 library in Python 2.7 with a TensorFlow 0.10.0 backend. The random forest and regularized logistic regression classifiers were implemented with Python Scikit-Learn 0.18.1. The isolate diversity analysis was implemented using the R stats 3.4.2 package, the *t*-SNE analysis used the Rtsne 0.13 package in R, and the permutation tests were implemented in Python 2.7. All models were trained on a NVIDIA GeForce GTX Titan X graphics processing unit (GPU). Hyperparameters are available in Supplementary Table S7.

Statistical Analyses

Predictive performance during cross-validation was reported in mean and standard error of the validation dataset over the five folds of training (Figure 2). Determination of resistance-correlated mutations during permutation tests used a significance level of $\alpha = 0.05$ corrected for multiple comparisons.

References

1. WHO, Global Tuberculosis Report 2016, *CDC 2016*, 214 (2016).
2. P. Bradley, N. C. Gordon, T. M. Walker, L. Dunn, S. Heys, B. Huang, S. Earle, L. J. Pankhurst, L. Anson, M. De Cesare, P. Piazza, A. A. Votintseva, T. Golubchik, D. J. Wilson, D. H. Wyllie, R. Diel, S. Niemann, S. Feuerriegel, T. A. Kohl, N. Ismail, S. V. Omar, E. G. Smith, D. Buck, G. McVean, A. S. Walker, T. E. A. Peto, D. W. Crook, Z. Iqbal, Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis, *Nat. Commun.* **6** (2015), doi:10.1038/ncomms10063.
3. WHO, Multidrug and extensively drug-resistant TB (M/XDR-TB) 2010 Global Report on Surveillance and Response, (2010) (available at http://apps.who.int/iris/bitstream/10665/44286/1/9789241599191_eng.pdf?ua=1&ua=1).
4. Q. Liu, G. L. Li, C. Chen, J. M. Wang, L. Martinez, W. Lu, L. M. Zhu, Diagnostic performance of the genotype MTBDRplus and MTBDRs/assays to identify tuberculosis drug resistance in eastern China, *Chin. Med. J. (Engl).* **130**, 1521–1528 (2017).
5. G. Theron, J. Peter, M. Richardson, M. Barnard, S. Donegan, R. Warren, K. R. Steingart, K. Dheda, The diagnostic accuracy of the GenoType((R)) MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs, *Cochrane Database Syst Rev* **10**, Cd010705 (2014).
6. E. Tagliani, A. M. Cabibbe, P. Miotto, E. Borroni, J. C. Toro, M. Mansjö, S. Hoffner, D. Hillemann, A. Zalutskaya, A. Skrahina, D. M. Cirillo, 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, *J. Clin. Microbiol.* **53**, 2961–2969 (2015).
7. D. I. Ling, A. A. Zwerling, M. Pai, GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: A meta-analysis, *Eur. Respir. J.* **32**, 1165–1174 (2008).
8. M. R. Farhat, R. Sultana, O. Iartchouk, S. Bozeman, J. Galagan, P. Sisk, C. Stolte, H. Nebenzahl-Guimaraes, K. Jacobson, A. Sloutsky, D. Kaur, J. Posey, B. N. Kreiswirth, N. Kurepina, L. Rigouts, E. M. Streicher, T. C. Victor, R. M. Warren, D. Van Soolingen, M. Murray, Genetic determinants of drug resistance in mycobacterium tuberculosis and their diagnostic value, *Am. J. Respir. Crit. Care Med.* **194**, 621–630 (2016).
9. M. R. Farhat, K. R. Jacobson, M. F. Franke, D. Kaur, A. Sloutsky, C. D. Mitnick, M. Murray, Gyrase Mutations Are Associated with Variable Levels of Fluoroquinolone Resistance in Mycobacterium tuberculosis, *J. Clin. Microbiol.* **54**, 727–733 (2016).
10. H. Safi, S. Lingaraju, A. Amin, S. Kim, M. Jones, M. Holmes, M. McNeil, S. N. Peterson, D. Chatterjee, R. Fleischmann, D. Alland, Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-Arabinose biosynthetic and utilization pathway genes, *Nat. Genet.* **45**, 1190–1197 (2013).
11. H. Nebenzahl-Guimaraes, K. R. Jacobson, M. R. Farhat, M. B. Murray, Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in Mycobacterium tuberculosis*J. Antimicrob. Chemother.* **69**, 331–342 (2014).
12. T. M. Walker, T. A. Kohl, S. V. Omar, J. Hedge, C. Del Ojo Elias, P. Bradley, Z. Iqbal, S. Feuerriegel, K. E. Niehaus, D. J. Wilson, D. A. Clifton, G. Kapatai, C. L. C. Ip, R. Bowden, F. A. Drobniewski, C. Allix-Béguec, C. Gaudin, J. Parkhill, R. Diel, P. Supply, D. W. Crook, E. G. Smith, A. S. Walker, N. Ismail, S. Niemann, T. E. A. Peto, J. Davies, C. Crichton, M. Acharya, L. Madrid-Marquez, D. Eyre, D. Wyllie, T. Golubchik, M. Munang, Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: A retrospective cohort study, *Lancet Infect. Dis.* **15**, 1193–1202 (2015).

13. Y. Yang, K. E. Niehaus, T. M. Walker, Z. Iqbal, A. S. Walker, D. J. Wilson, T. E. Peto, D. W. Crook, E. G. Smith, T. Zhu, D. A. Clifton, Machine Learning for Classifying Tuberculosis Drug-Resistance from DNA Sequencing Data, *Bioinformatics*, Advance online publication. (2017).
14. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, *Adv. Neural Inf. Process. Syst.*, 1–9 (2012).
15. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, Deep Neural Networks for Acoustic Modeling in Speech Recognition, *IEEE Signal Process. Mag.*, 82–97 (2012).
16. R. Socher, C. Lin, Parsing natural scenes and natural language with recursive neural networks, *ICML*, 129–136 (2011).
17. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, D. R. Webster, Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs., *JAMA* **304**, 649–656 (2016).
18. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, *Nature* **542**, 115–118 (2017).
19. S. Zhang, J. Zhou, H. Hu, H. Gong, L. Chen, C. Cheng, J. Zeng, A deep learning framework for modeling structural features of RNA-binding protein targets, *Nucleic Acids Res.* **44**, 1–14 (2015).
20. Y. Chen, Y. Li, R. Narayan, A. Subramanian, X. Xie, Gene expression inference with deep learning, *Bioinformatics* **32**, 1832–1839 (2016).
21. E. Putin, P. Mamoshina, A. Aliper, M. Korzinkin, A. Moskalev, A. Kolosov, A. Ostrovskiy, C. Cantor, J. Vijg, A. Zhavoronkov, Deep biomarkers of human aging: Application of deep neural networks to biomarker development, *Aging (Albany. NY).* **8**, 1021–1033 (2016).
22. H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, R. Anil, Z. Haque, L. Hong, V. Jain, X. Liu, H. Shah, Wide & Deep Learning for Recommender Systems, *arXiv Prepr.*, 1–4 (2016).
23. A. M. Starks, E. Aviles, D. M. Cirillo, C. M. Denkinger, D. L. Dolinger, C. Emerson, J. Gallarda, D. Hanna, P. S. Kim, R. Liwski, P. Miotto, M. Schito, M. Zignol, Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform, *Clin. Infect. Dis.* **61**, S141–S146 (2015).
24. L. J. P. Van Der Maaten, G. E. Hinton, Visualizing high-dimensional data using t-sne, *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
25. A. L. Manson, K. A. Cohen, T. Abeel, C. A. Desjardins, D. T. Armstrong, C. E. Barry, J. Brand, TBResist Global Genome Consortium, S. B. Chapman, S.-N. Cho, A. Gabrielian, J. Gomez, A. M. Jodals, M. Joloba, P. Jureen, J. S. Lee, L. Malinga, M. Maiga, D. Nordenberg, E. Noroc, E. Romancenco, A. Salazar, W. Ssengooba, A. A. Velayati, K. Winglee, A. Zalutskaya, L. E. Via, G. H. Cassell, S. E. Dorman, J. Ellner, P. Farnia, J. E. Galagan, A. Rosenthal, V. Crudu, D. Homorodean, P.-R. Hsueh, S. Narayanan, A. S. Pym, A. Skrahina, S. Swaminathan, M. Van der Walt, D. Alland, W. R. Bishai, T. Cohen, S. Hoffner, B. W. Birren, A. M. Earl, Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance., *Nat. Genet.* **49**, 395–402 (2017).
26. M. R. Farhat, C. D. Mitnick, M. F. Franke, D. Kaur, A. Sloutsky, M. Murray, K. R. Jacobson, Concordance of Mycobacterium tuberculosis fluoroquinolone resistance testing:

implications for treatment, *Int J Tuberc Lung Dis* **19**, 339–341 (2015).
27. K. J. Aldred, T. R. Blower, R. J. Kerns, J. M. Berger, N. Osheroff, Fluoroquinolone interactions with *Mycobacterium tuberculosis* gyrase: Enhancing drug activity against wild-type and resistant gyrase, *Proc. Natl. Acad. Sci.* **113**, E839–E846 (2016).
28. H. Zhang, D. Li, L. Zhao, J. Fleming, N. Lin, T. Wang, Z. Liu, C. Li, N. Galwey, J. Deng, Y. Zhou, Y. Zhu, Y. Gao, T. Wang, S. Wang, Y. Huang, M. Wang, Q. Zhong, L. Zhou, T. Chen, J. Zhou, R. Yang, G. Zhu, H. Hang, J. Zhang, F. Li, K. Wan, J. Wang, X. E. Zhang, L. Bi, Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance, *Nat. Genet.* **45**, 1255–1260 (2013).
29. B. Ramsundar, B. Liu, Z. Wu, A. Verras, M. Tudor, R. P. Sheridan, V. S. Pande, Is Multitask Deep Learning Practical for Pharma?, *J. Chem. Inf. Model.* **57**, 2068–2076 (2017).
30. S. Kearnes, B. Goldman, V. Pande, Modeling Industrial ADMET Data with Multitask Networks, *arXiv* (2016), doi:1606.08793v1.pdf.
31. Q. Qin, J. Feng, Imputation for transcription factor binding predictions based on deep learning, *PLoS Comput. Biol.* **13** (2017), doi:10.1371/journal.pcbi.1005403.
32. J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, V. Svetnik, Deep neural nets as a method for quantitative structure-activity relationships, *J. Chem. Inf. Model.* **55**, 263–274 (2015).
33. G. Dahl, N. Jaitly, R. Salakhutdinov, Multi-task Neural Networks for QSAR Predictions, *arXiv Prepr. arXiv1406.1231*, 1–21 (2014).
34. M. D. Zeiler, R. Fergus, Visualizing and Understanding Convolutional Networks arXiv:1311.2901v3 [cs.CV] 28 Nov 2013, *Comput. Vision–ECCV 2014* **8689**, 818–833 (2014).
35. J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding Neural Networks Through Deep Visualization, *ICML - Deep Learn. Work. 2015*, 12 (2015).
36. A. Kolyva, P. Karakousis, Old and new TB drugs: Mechanisms of action and resistance, *InTechOpen*, 210–232 (2012).
37. X. Didelot, R. Bowden, D. J. Wilson, T. E. A. Peto, D. W. Crook, Transforming clinical microbiology with bacterial genome sequencing, *Nat. Rev. Genet.* **13**, 601–612 (2012).
38. C. U. Köser, J. M. Bryant, J. Becq, M. E. Török, M. J. Ellington, M. A. Marti-Renom, A. J. Carmichael, J. Parkhill, G. P. Smith, S. J. Peacock, Whole-genome sequencing for rapid susceptibility testing of M. tuberculosis., *N. Engl. J. Med.* **369**, 290–2 (2013).
39. A. A. Votintseva, P. Bradley, L. Pankhurst, C. Del Ojo Elias, M. Loose, K. Nilgiriwala, A. Chatterjee, E. G. Smith, N. Sanderson, T. M. Walker, M. R. Morgan, D. H. Wyllie, A. S. Walker, T. E. A. Peto, D. W. Crook, Z. Iqbal, Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples, *J. Clin. Microbiol.* **55**, 1285–1298 (2017).
40. World Health Organization (WHO), A roadmap for ensuring quality tuberculosis diagnostics services within national laboratory strategicplans. (2010).
41. K. Ängeby, P. Juréen, G. Kahlmeter, S. E. Hoffner, T. Schön, Challenging a dogma: antimicrobial susceptibility testing breakpoints for Mycobacterium tuberculosis., *Bull. World Health Organ.* **90**, 693–8 (2012).
42. T. D. Lieberman, D. Wilson, R. Misra, L. L. Xiong, P. Moodley, T. Cohen, R. Kishony, Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis, *Nat. Med.* **22**, 1470–1474 (2016).
43. A. Chatterjee, K. Nilgiriwala, D. Saranath, C. Rodrigues, N. Mistry, Whole genome sequencing of clinical strains of Mycobacterium tuberculosis from Mumbai, India: A potential tool for determining drug-resistance and strain lineage, *Tuberculosis* **107**, 63–72 (2017).

44. J. L. Gardy, J. C. Johnston, S. J. H. Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. M. Jones, F. S. L. Brinkman, R. C. Brunham, P. Tang, Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak, *N. Engl. J. Med.* **364**, 730–739 (2011).
45. G. Lunter, M. Goodson, Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads, *Genome Res.* **21**, 936–939 (2011).
46. A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg, A. O. M. Wilkie, G. Mcvean, G. Lunter, Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications, *Nat. Genet.* **46**, 912–918 (2014).
47. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools, *Bioinformatics* **25**, 2078–2079 (2009).
48. D. E. Wood, S. L. Salzberg, Kraken: Ultrafast metagenomic sequence classification using exact alignments, *Genome Biol.* **15** (2014), doi:10.1186/gb-2014-15-3-r46.
49. M. B. Gikalo, E. Y. Nosova, L. Y. Krylova, A. M. Moroz, The role of eis mutations in the development of kanamycin resistance in Mycobacterium tuberculosis isolates from the moscow region, *J. Antimicrob. Chemother.* **67**, 2107–2109 (2012).
50. W. Shi, X. Zhang, X. Jiang, H. Yuan, J. S. Lee, C. E. Barry, H. Wang, W. Zhang, Y. Zhang, Pyrazinamide Inhibits Trans-Translation in Mycobacterium tuberculosis, *Science (80-.).* **333**, 1630–1632 (2011).
51. F. Murtagh, P. Legendre, Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?, *J. Classif.* **31**, 274–295 (2014).
52. X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, *AISTATS '11 Proc. 14th Int. Conf. Artif. Intell. Stat.* **15**, 315–323 (2011).
53. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
54. Y. Gal, Z. Ghahramani, Dropout as a Bayesian Approximation : Representing Model Uncertainty in Deep Learning, *ICML* **48**, 1–10 (2015).
55. R. Tibshirani, Regression Selection and Shrinkage via the Lasso, *J. R. Stat. Soc. B* **58**, 267–288 (1996).

Supplementary Materials

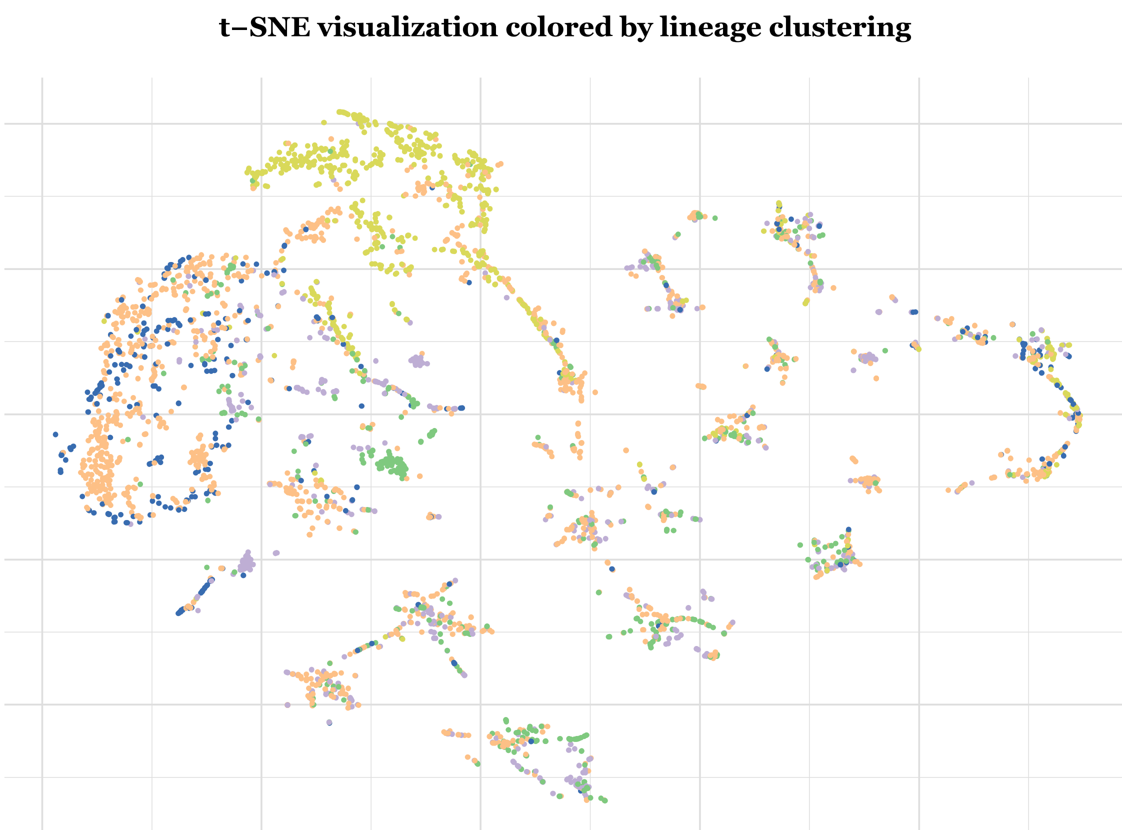

Figure S1: *t*-SNE visualization colored by lineage clustering. t-SNE plot with the same coordinates as in Figure 3. Each isolate is colored based on the six lineage clusters determined in Figure 1, illustrating the diversity of MTB isolates within the multitask WDNN's resistance-susceptibility clustering.

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 S1: **Phenotype of 3,601 *Mycobacterium* tuberculosis isolates in training and cross-validation.** Phenotype availability for the 10 anti-tubercular drugs.

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 S2: **Phenotype of 792 *Mycobacterium* tuberculosis isolates in held-out validation set.** Phenotype availability for the 10 anti-tubercular drugs in an independent validation set.

Lineage-defining mutations to determine isolate diversity
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 S3: **Lineage-defining mutations to determine isolate diversity**. A table of 33 mutations used to determine isolate diversity by genetic covariance and hierarchical clustering.

	MLP (Select Mutations)		Multitask WDNN		Random Forest		Logistic Regression		Single task WDNN	
Drugs	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
RIF	97.2 ± 0.5	93.1 ± 0.2	97.7 ± 0.6	96.2 ± 0.5	95.9 ± 0.6	94.2 ± 0.3	97.1 ± 1.0	96.1 ± 0.4	98.3 ± 0.5	95.1 ± 0.5
INH	95.4 ± 0.5	95.5 ± 0.5	96.5 ± 0.4	95.6 ± 0.4	95.3 ± 0.3	96.7 ± 0.3	96.3 ± 0.4	95.4 ± 0.5	96.1 ± 0.5	95.3 ± 0.4
PZA	87.7 ± 1.3	91.2 ± 0.7	91.3 ± 1.2	93.4 ± 0.6	91.0 ± 0.7	90.4 ± 0.7	93.4 ± 1.0	89.9 ± 0.9	90.3 ± 1.3	92.2 ± 0.4
EMB	89.4 ± 1.0	90.9 ± 0.3	90.9 ± 0.9	93.3 ± 0.5	94.9 ± 0.2	88.4 ± 0.4	94.4 ± 0.2	91.7 ± 0.3	92.8 ± 0.8	91.5 ± 0.3
STR	88.2 ± 0.9	84.2 ± 1.7	87.1 ± 1.3	85.2 ± 0.8	86.5 ± 1.2	84.1 ± 1.5	82.7 ± 0.5	88.4 ± 0.7	91.3 ± 0.8	81.7 ± 0.9
CAP	60.1 ± 1.4	86.4 ± 1.2	91.8 ± 2.1	89.7 ± 1.4	91.5 ± 1.4	89.5 ± 1.4	88.6 ± 1.1	88.0 ± 0.6	94.5 ± 1.1	86.2 ± 0.8
AMK	86.8 ± 2.6	95.1 ± 0.5	85.6 ± 1.5	97.3 ± 0.7	88.4 ± 2.7	94.7 ± 1.0	85.8 ± 3.0	96.9 ± 0.8	89.9 ± 2.0	91.6 ± 1.3
MOXI	58.6 ± 3.3	89.4 ± 0.8	77.3 ± 1.6	89.5 ± 1.4	74.9 ± 1.1	90.3 ± 0.5	74.8 ± 2.1	90.1 ± 0.6	76.0 ± 3.1	89.8 ± 0.9
OFLX	84.2 ± 1.7	89.9 ± 1.4	79.1 ± 4.5	92.8 ± 0.5	81.7 ± 5.3	95.2 ± 0.4	73.4 ± 2.5	93.0 ± 0.9	82.0 ± 2.0	90.8 ± 1.1
KAN	71.4 ± 2.4	93.0 ± 1.8	76.2 ± 0.9	94.6 ± 0.8	73.6 ± 3.6	91.1 ± 1.3	75.7 ± 2.6	90.0 ± 1.2	77.2 ± 2.8	88.2 ± 1.4

Table S4: **Tuberculosis drug resistance prediction performance of the multitask WDNN and baseline models from cross-validation**. A table of predictive performance across all four models during cross-validation. The multitask WDNN, single task WDNN, random forest, and logistic regression models were trained on the full set of predictors, while the single task MLP was trained on preselected mutations.

	MLP (Select Mutations)		Multitask WDNN		Random Forest		Logistic Regression		Single task WDNN	
Drugs	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
RIF	97.5	90.5	96.1	96.7	85.5	97.8	91.8	98.0	96.1	94.9
INH	84.2	97.1	91.2	94.5	75.5	100.0	83.6	100.0	87.3	95.1
PZA	61.7	96.1	63.9	94.7	54.9	96.5	61.7	96.1	65.4	91.7
EMB	90.6	80.4	83.1	88.0	62.5	94.6	70.0	92.0	84.4	86.8
STR	82.9	96.5	88.2	94.2	42.8	97.9	77.6	97.5	88.8	92.8
CAP	59.4	79.3	53.1	94.5	31.3	99.0	40.6	98.6	56.3	93.3
AMK	52.6	97.8	52.6	98.9	52.6	100.0	63.2	91.6	57.9	93.4
MOXI	15.0	95.5	80.0	93.3	70.0	96.6	55.0	94.9	85.0	92.7
OFLX	79.3	91.5	66.3	97.5	53.3	98.1	59.8	97.5	57.6	93.4
KAN	47.2	89.9	67.9	94.2	71.7	98.2	50.9	99.0	62.3	91.4

Table S5: **Tuberculosis drug resistance prediction performance of the multitask WDNN and baseline models on the independent validation set**. A table of predictive performance across all four models on the independent validation set. The multitask WDNN, single task WDNN, random forest, and logistic regression models were trained on the full set of predictors, while the single task MLP was trained on preselected mutations.

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 S6: **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.

Multitask WDNN and Single task WDNN	
Hyperparameter	**Value**
L1 regularization	10^-6
Hidden units per layer	512
Number of hidden layers	2
Dropout	0.6
Learning rate	e^{-7}
Optimizer	Adam
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**
L1 regularization	Best penalty factor between 10^-5 and 10^5
Multilayer Perceptron (MLP)	
Hyperparameter	**Value**
Hidden units per layer	512
Number of hidden layers	3
Dropout	0.5
Learning rate	0.001
Optimizer	Adam

Table S7: **Hyperparameters for the multitask and single task WDNN, baseline models, and the MLP**. A table of hyperparameters for each model. The L1 regularization factor for logistic regression was determined using cross-validation to maximize the area-under-the-ROC-curve (AUC) within the 80% training data for each fold.