**Preprocessing**

Based on the following two papers:  
  
[Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN](https://www.nature.com/articles/s41598-022-06449-4?fromPaywallRec=false) (1)

&

[A convolutional neural network highlights mutations relevant to antimicrobial resistance in *Mycobacterium tuberculosis*](https://www.nature.com/articles/s41467-022-31236-0?fromPaywallRec=false) (2)

**Preprocessing for paper 1**

### Data collection

To prepare the training data and labels, we downloaded the whole-genome sequencing (WGS) data for 10,575 MTB isolates from the sequence read archive (SRA) database and obtained corresponding lineage and phenotypic drug susceptibility test (DST) data from CRyPTIC Consortium and the 100,000 Genomes project in an excel file, which is also available in the supplementary of their publication. The phenotypic DST results for the drugs were used as labels when training and evaluating our ML models. All the data were collected and shared by the CRyPTIC Consortium and the 100,000 Genomes Project. Like the datasets used by previous studies, this dataset is imbalanced in that most isolates are susceptible, and the minority of them are resistant for all the four first-line drugs and four second-line drugs. The numbers of isolate samples with phenotypic DST results available are 7138, 7137, 6347 and 7081 for EMB, INH, PZA and RIF, respectively. There are 6291 shared isolates among the four sample sets. In addition, 6820 out of the 10,575 isolates have phenotypic DST result available for each of the four second-line drugs.

### Genetic feature extraction

To detect the potential genetic features that could contribute to MTB drug resistance classification, we used a command-line tool called ARIBA. ARIBA is a very rapid, flexible and accurate AMR genotyping tool that generates detailed and customizable outputs from which we extracted genetic features. First, we downloaded all reference data from CARD, which included not only references from different MTB strains but also from other bacteria (e.g., *Staphylococcus aureus*). Secondly, we clustered reference sequences based on their similarity. Then we used this collection of reference clusters as our pan-genome reference and aligned read pairs of an isolate to them. For each cluster that had reads mapped, we ran local assemblies, found the closest reference, and identified variants. After running these steps, ARIBA generated files including a summary file for alignment quality, a report file containing information of detected variants and AMR-associated genes, and a read depth file. For each cluster, the read depth file provides counts of the four DNA bases on each locus of the closest reference where reads were mapped.

Next, we filtered out low-quality mappings that did not pass the ‘match’ criteria defined in ARIBA’s GitHub wiki. From these high-quality mappings, we collected novel variants in coding regions, well-studied resistance-causing variants and AMR-associated gene presences that were detected from at least one out of the 10,575 isolates as 263 genetic features. In addition, we included indicator variables for each of the 19 lineages into our feature vector resulting in a total of 282 features.

Here are the preprocessing steps based on the data collection and genetic feature extraction process described:

1. **Download and Aggregate Data**:
   * Download WGS data for 10,575 MTB isolates from the Sequence Read Archive (SRA).
   * Collect corresponding lineage and phenotypic drug susceptibility test (DST) data from the CRyPTIC Consortium and the 100,000 Genomes Project.
   * Organize data into a structured format (e.g., Excel) for use in ML models.
   * Ensure alignment of WGS data and DST labels for consistency.
2. **Data Cleaning**:
   * Identify and handle missing or incomplete DST data.
   * Filter out isolates with incomplete phenotypic DST data (e.g., isolates without results for first- and second-line drugs).
   * Remove duplicate or redundant samples if any.
3. **Handling Data Imbalance**:
   * Address the imbalance in the dataset (most isolates are susceptible, few are resistant) using techniques like:
     + Oversampling the minority class.
     + Undersampling the majority class.
     + Applying Synthetic Minority Over-sampling Technique (SMOTE) or other balancing methods during ML model training.
4. **Reference Data Preparation**:
   * Download reference data from the Comprehensive Antibiotic Resistance Database (CARD), which includes MTB strains and other bacteria (e.g., Staphylococcus aureus).
   * Cluster reference sequences based on their similarity to create a pan-genome reference.
5. **Mapping and Variant Calling**:
   * Align paired-end reads from each isolate to the pan-genome reference.
   * Perform local assemblies and identify the closest reference for each isolate.
   * Detect variants and antibiotic resistance-associated genes using ARIBA’s genotyping tool.
6. **Filtering Low-Quality Mappings**:
   * Apply quality control to the mapping outputs from ARIBA by filtering out alignments that do not meet the criteria specified in ARIBA’s ‘match’ criteria.
   * Retain only high-quality mappings for further analysis.
7. **Feature Extraction**:
   * Extract genetic features from ARIBA’s output, including:
     + Novel variants in coding regions.
     + Known resistance-causing variants.
     + AMR-associated gene presences.
   * Incorporate additional features such as lineage indicators for the 19 lineages.
8. **Feature Vector Construction**:
   * Combine the 263 genetic features and 19 lineage indicators into a single feature vector for each isolate, resulting in 282 features.
9. **Scaling and Normalization**:
   * Scale and normalize the feature values (e.g., genetic feature presence, variant counts) for ML model input to ensure uniformity across features.
10. **Splitting Data**:

* Split the dataset into training, validation, and testing subsets, ensuring that the split respects the class imbalance and maintains representation for both susceptible and resistant isolates.

These steps ensure that the data is preprocessed and ready for model training and evaluation in the classification of MTB drug resistance.

**Preprocessing for paper 2**

## Methods

### Sequence data

The training, cross-validation, and test datasets consist of a combined 23,049 *M. tuberculosis* isolates for which whole genome sequence data and antibiotic resistance phenotype data are available. The sequencing data are obtained through the National Center for Biotechnology Information (NCBI) database, PATRIC, and published literature: 10,201 strains are in the “train” dataset (for training and cross-validation), 7537 are in the hold-out “test\_1” dataset (for hold-out testing), and the remaining 5312 are in the hold-out “test\_GenTB” dataset (for hold-out testing). Isolates were added to the “test\_1” and “test\_GenTB” datasets on a rolling basis—i.e., as the sequencing data became available gradually over time.

We process sequences in the train and test\_1 datasets using a previously validated pipeline as described by Ezewudo et al. (2018), with modifications as elaborated by Freschi et al. (2021). Reads are trimmed and filtered using PRINSEQ, contaminated isolates are removed using Kraken, and aligned to the reference genome H37Rv using BWA-MEM. Duplicate reads are removed using Picard, and we drop isolates with <95% coverage of the reference genome at 10× coverage.

For the “test\_GenTB” dataset, we use the sequencing data prepared by Groschel et al., which employs a different variant of the Ezewudo et al. pipeline. The differences between these two pipelines (most notably the use of minimap-2 instead of BWA-MEM) make a negligible difference on final variant calls[59](https://www.nature.com/articles/s41467-022-31236-0#ref-CR59).

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

### Antimicrobial resistance phenotype data

Culture-based antimicrobial drug susceptibilities to at least one of 13 anti-TB drugs are available for all 23,049 isolates in the combined training, cross-validation, and test dataset. Phenotypes (drug susceptibility test results) for isolates in the training and cross-validation dataset are from the ReSeqTB data portal, the PATRIC database, and manual curation of phenotypic data available in the literature. Phenotypes for the test dataset isolates are from data available in the literature. Each isolate’s phenotype is classified as resistant, susceptible, or unavailable, with respect to a combination of 13 possible first-line (rifampicin, isoniazid, pyrazinamide, ethambutol) and second-line drugs (streptomycin, ciprofloxacin, levofloxacin, moxifloxacin, ofloxacin, capreomycin, amikacin, kanamycin, ethionamide). In the hold-out test dataset, ethionamide and ciprofloxacin are excluded due to having fewer than 50 phenotyped resistant isolates (0/2 resistant to ciprofloxacin; 12/25 resistant to ethionamide).

### Selecting input loci

The loci of the isolate sequences are selected from genes known or suspected to cause resistance based on previous models and experiments. In order to incorporate any possible regulatory sequences from the immediate genetic neighborhood, the entire upstream and downstream region of each gene or operon is included (upstream region: from the beginning of the relevant gene to the end of the previous gene on the genome; downstream region: from the end of the relevant gene to the beginning of the next gene on the genome). Loci are aligned to the H37Rv reference genome for comparison of coordinates and genome annotations are based on H37Rv coordinates from Mycobrowser.

### Evaluation on CRyPTIC isolates

Binary phenotype data is downloaded from the CRyPTIC study. CRyPTIC has phenotype data for the following drugs predicted by the CNN: isoniazid, rifampicin, ethambutol, amikacin, kanamycin, moxifloxacin, levofloxacin, ethionamide. Phenotypes whose quality is not “high” are masked. We filter isolates that do not have a phenotype for any of the drugs of interest, for a total of 9498 isolates.

Isolates from the dataset of 9498 are analyzed using a variant of the Ezewudo et al. pipeline, with modifications as elaborated by Freschi et al. (2021)[48](https://www.nature.com/articles/s41467-022-31236-0#ref-CR48),[54](https://www.nature.com/articles/s41467-022-31236-0#ref-CR54), additionally using minimap2-2.24 for read mapping, SPAdes v 3.15.4 for assembly, and trimmomatic v. 0.40 for read trimming. Nucleotide sequences for the designated genomic loci are extracted and aligned against the previous input sequence alignments using MAFFT v7.490 with the --add and --keeplength options.

Confidence intervals for sensitivity and specificity on the full CRyPTIC dataset are generated by sampling 80% of the dataset 100 times.

### Analysis of mis-predicted isolates

For each SD-CNN model, we compute the genetic distance (number of different sites) between all isolates in the training and test sets. Following Vargas et al., 2021[70](https://www.nature.com/articles/s41467-022-31236-0#ref-CR70), to compute the genetic distance for the entire genome, we process the VCF files output by our pipeline described above, and take all sites that meet the following criteria: mean Base Quality > 20, mean Mapping Quality > 30, no reads supporting insertions or deletions, number of high quality reads ≥ 20, and at least 75% support for a non-reference allele. Sites falling between 25% and 75% support for non-reference alleles are labeled as uncertain and do not contribute to distance calculation. We further remove sites with an empirical base pair recall score < 90% and sites where at least 10% of the isolates have uncertain calls. We then compute the number of differences between pairs of isolates.

For computing the number of differences from the perspective of the SD-CNN model, only sites found in the loci used in each SD-CNN model are included, and each site with a confident insertion or deletion contributes one to the difference score.

### Designation of known resistance variants from WHO catalog

A list of known resistance-conferring variants is extracted from the WHO catalog. Only variants with a Final Confidence Grading of “Category 1: Associated with resistance” or “Category 2: Associated with resistance—interim” are taken to be known resistance-conferring variants.

For prediction of resistance with the WHO catalog, isolates are assumed to be sensitive unless they have one or more of the known resistance-conferring variants for a particular drug.

### Saliency calculation

Saliencies are calculated using DeepLIFT v. 0.6.12.0, using the recommended defaults for genomics: “rescale” rule applied to convolutional layers, and “reveal-cancel” rule applied to fully connected layers. We use the H37Rv reference genome, which is sensitive to all antibiotics, as the baseline[25](https://www.nature.com/articles/s41467-022-31236-0#ref-CR25).

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

### Lineage variant analysis

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

This method section outlines a comprehensive process for training and testing machine learning models using whole-genome sequencing (WGS) data from *M. tuberculosis* (MTB) isolates and their corresponding drug resistance phenotypes. Here's a breakdown of the key steps mentioned:

### **1. Data Collection and Preparation:**

* The dataset includes 23,049 MTB isolates with WGS and antimicrobial resistance phenotype data.
* The data are split into three sets:
  + Training set: 10,201 isolates (for training and cross-validation).
  + Hold-out test set 1: 7,537 isolates.
  + Hold-out test set "GenTB": 5,312 isolates.
* Sequence data are processed using different pipelines, including the Ezewudo pipeline and its variants. Sequence alignment is done using BWA-MEM or minimap-2, and sequences with <95% coverage are excluded.

### **2. Phenotypic Data:**

* Culture-based drug susceptibility test (DST) results are collected for 13 first- and second-line anti-TB drugs.
* Phenotypes are categorized as resistant, susceptible, or unavailable.
* Two drugs, ethionamide and ciprofloxacin, are excluded from the hold-out test set due to insufficient resistant isolates.

### **3. Genetic Feature Selection:**

* Loci are selected from genes known or suspected to cause resistance based on prior models and experiments.
* The surrounding upstream and downstream regions of these loci are included to capture regulatory sequences.
* Loci are aligned to the H37Rv reference genome.

### **4. Pipeline Modifications:**

* The Freschi et al. pipeline is used for aligning and assembling sequences in the CRyPTIC dataset, with tools like minimap2, SPAdes, and trimmomatic for trimming and alignment.
* MAFFT is used to align nucleotide sequences.

### **5. Machine Learning Input:**

* Curated genetic variants serve as features for the Wide and Deep Neural Network (WDNN) model.
* FASTA files for loci of interest are used for the Single-Drug and Multi-Drug Convolutional Neural Networks (SD-CNN and MD-CNN).

### **6. Evaluation on CRyPTIC Isolates:**

* CRyPTIC data for binary phenotypes are used to evaluate the models.
* Isolates without phenotypes are filtered out, leaving 9,498 isolates for analysis.
* Confidence intervals for sensitivity and specificity are generated through 100 bootstrapped samples.

### **7. Analysis of Mis-predicted Isolates:**

* Genetic distance between isolates in the training and test sets is computed using VCF files from the pipeline.
* Differences in genetic sites are analyzed to assess potential model mispredictions.

### **8. WHO Catalog Resistance Variants:**

* Known resistance-conferring variants are extracted from the WHO catalog and incorporated into the models.
* If an isolate contains one or more resistance variants, it is predicted as resistant for that drug.

### **9. Saliency Calculation:**

* Saliency maps are computed using DeepLIFT, comparing isolate sequences to the H37Rv reference genome.
* Saliency scores help identify which genome positions influence antibiotic resistance prediction.

### **10. Lineage Variant Analysis:**

* Lineage-defining variants are analyzed to assess their association with resistance.
* Lineage variants from the Freschi et al. barcode are cross-referenced with variants in the dataset.

This method provides a thorough framework for leveraging WGS data, curated genetic features, and machine learning models to predict drug resistance in MTB isolates. The process involves extensive data preprocessing, feature selection, and validation across multiple datasets using both classical and advanced genomic pipelines.