# **Questions**

1. What are the CARD database and the WHO catalog?

**CARD (Comprehensive Antibiotic Resistance Database):**

* The **CARD** database is a collection of known **antibiotic-resistance genes** and mutations. It contains detailed information about how bacteria develop resistance to different antibiotics by acquiring specific genetic changes.
* **Purpose**: It is used to identify and understand antibiotic resistance mechanisms in bacterial genomes. Researchers use this database to find links between genetic mutations and antibiotic resistance.

### **WHO Catalogue (World Health Organization Catalogue of Mutations):**

* The **WHO Catalogue of Mutations** is a comprehensive list of genetic mutations that are associated with **drug resistance in tuberculosis (TB)**.
* **Purpose**: This catalogue identifies specific mutations in the **Mycobacterium tuberculosis (MTB)** genome that are confirmed to cause resistance to various TB drugs. It helps guide diagnostic tests and treatment decisions by indicating which mutations are linked to resistance.

In summary, the **CARD database** focuses on antibiotic resistance genes across various bacteria, while the **WHO Catalogue** is specifically about TB and tracks mutations that lead to drug resistance in **Mycobacterium tuberculosis**.

### **Scope:**

* **CARD Database**:
  + **Focus**: Broad, covering a wide range of **antibiotic resistance genes and mutations** across **various bacterial species**.
  + **Application**: It is used to study antibiotic resistance mechanisms in many types of bacteria, not just **Mycobacterium tuberculosis**.
  + **Data Type**: Includes genetic information on resistance mechanisms like efflux pumps, antibiotic degradation, and target modification for a variety of antibiotics and bacteria.
* **WHO Catalogue of Mutations**:
  + **Focus**: Narrow, specifically focused on **drug resistance mutations** in **Mycobacterium tuberculosis**.
  + **Application**: It is primarily used to guide **TB diagnosis and treatment** by identifying mutations known to cause resistance to TB drugs.
  + **Data Type**: Lists mutations that are directly linked to **drug resistance in tuberculosis**, often used to improve the detection of resistant TB strains in clinical settings.

2. What is a pan-genome reference?

A **pan-genome reference** is a collection of all the genetic material (genes and variations) found across multiple strains or individuals of a particular species, rather than using the genome of just a single strain. It represents the **complete set of genes** that may be present in a species, covering both **core genes** (found in all individuals of the species) and **accessory genes** (genes found in only some individuals or strains).

### **Key Components of a Pan-Genome:**

1. **Core Genome**: These are the genes shared by all strains of a species. They represent essential functions that are necessary for the survival of the organism.
2. **Accessory Genome**: These are genes that are present in some strains but not others. They often provide strain-specific functions, such as adaptations to particular environments or antibiotic resistance.
3. **Pan-Genome**: The sum of the core and accessory genomes, representing all the possible genetic information across different strains.

### **Why Use a Pan-Genome Reference?**

In traditional genetic studies, a **single reference genome** (from one strain of a species) is often used for comparison. This works well when studying closely related strains, but can cause problems when analyzing genetically diverse strains because it may miss important genes or variations that are unique to other strains.

* **Pan-genome reference** includes genetic diversity from multiple strains, so it can better represent the entire species.
* It reduces errors in **variant detection** (finding mutations or genetic changes) because the analysis is not restricted to just one genome.
* For species with high genetic variability, like **Mycobacterium tuberculosis**, a pan-genome can capture rare or strain-specific mutations, improving the accuracy of **antimicrobial resistance (AMR) prediction** and other analyses.

### **Benefits in AMR Prediction:**

In the context of **antimicrobial resistance (AMR) prediction**, using a pan-genome reference allows researchers to detect a broader range of **resistance-related genetic mutations** across different bacterial strains. This improves the quality and accuracy of predictions for new or diverse strains, which might not be well-represented by a single reference genome.

### **Summary:**

A **pan-genome reference** is a collection of all genetic material from multiple strains of a species, providing a more comprehensive representation of the species' genetic diversity. It helps improve the accuracy of genomic studies, especially when dealing with diverse bacterial strains, and is particularly useful in predicting **antibiotic resistance** by capturing all possible resistance-related mutations.

3.

# 

# 

# 

# 

# **Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN**

* [Xingyan Kuang](https://www.nature.com/articles/s41598-022-06449-4#auth-Xingyan-Kuang-Aff1),
* [Fan Wang](https://www.nature.com/articles/s41598-022-06449-4#auth-Fan-Wang-Aff1),
* [Kyle M. Hernandez](https://www.nature.com/articles/s41598-022-06449-4#auth-Kyle_M_-Hernandez-Aff1-Aff2),
* [Zhenyu Zhang](https://www.nature.com/articles/s41598-022-06449-4#auth-Zhenyu-Zhang-Aff1) &
* [Robert L. Grossman](https://www.nature.com/articles/s41598-022-06449-4#auth-Robert_L_-Grossman-Aff1-Aff2)

Effective and timely antibiotic treatment depends on accurate and rapid in silico antimicrobial-resistant (AMR) predictions. Existing statistical rule-based *Mycobacterium tuberculosis* (MTB) drug resistance prediction methods using bacterial genomic sequencing data often achieve varying results: high accuracy on some antibiotics but relatively low accuracy on others. Traditional machine learning (ML) approaches have been applied to classify drug resistance for MTB and have shown more stable performance. However, there is no study that uses deep learning architecture like Convolutional Neural Network (CNN) on a large and diverse cohort of MTB samples for AMR prediction. We developed 24 binary classifiers of MTB drug resistance status across eight anti-MTB drugs and three different ML algorithms: logistic regression, random forest and 1D CNN using a training dataset of 10,575 MTB isolates collected from 16 countries across six continents, where an extended pan-genome reference was used for detecting genetic features. Our 1D CNN architecture was designed to integrate both sequential and non-sequential features. In terms of F1-scores, 1D CNN models are our best classifiers that are also more accurate and stable than the state-of-the-art rule-based tool Mykrobe predictor (81.1 to 93.8%, 93.7 to 96.2%, 93.1 to 94.8%, 95.9 to 97.2% and 97.1 to 98.2% for ethambutol, rifampicin, pyrazinamide, isoniazid and ofloxacin respectively). We applied filter-based feature selection to find AMR relevant features. All selected variant features are AMR-related ones in CARD database. 78.8% of them are also in the catalogue of MTB mutations that were recently identified as drug resistance-associated ones by WHO. To facilitate ML model development for AMR prediction, we packaged every step into an automated pipeline and shared the source code at <https://github.com/KuangXY3/MTB-AMR-classification-CNN>.

This passage describes a new approach to predicting drug resistance in **Mycobacterium tuberculosis (MTB)** using **machine learning (ML)** and **deep learning** techniques, which are applied to a large global dataset. Here's a simplified explanation:

### **1. Problem: Predicting Drug Resistance**

* **Why it's important**: In order to effectively treat tuberculosis (TB), it’s crucial to quickly and accurately predict which antibiotics will work against specific strains of the TB-causing bacteria. Some existing methods struggle with predicting drug resistance accurately for all antibiotics.

### **2. Machine Learning vs. Traditional Methods**

* **Traditional Methods**: Current prediction methods often rely on rule-based systems, which work well for some antibiotics but not for others.
* **Machine Learning (ML)**: Traditional machine learning approaches like **logistic regression** and **random forest** have been used to improve drug resistance prediction and offer more stable performance compared to rule-based methods.

### **3. Deep Learning with CNNs**

* **Deep Learning Innovation**: The new method involves using a **1D Convolutional Neural Network (CNN)**, a type of deep learning model. This is the first time a CNN has been used on such a large and diverse dataset of TB samples for drug resistance prediction.
* **What makes CNNs special**: CNNs can analyze both **sequential (genetic sequence)** and **non-sequential features** (other types of data) to make better predictions.

### **4. The Dataset**

* The study used a large dataset of **10,575 MTB samples** collected from **16 countries across six continents**, making it highly diverse. This gives the model a broad range of data to learn from and improves its ability to predict drug resistance accurately.

### **5. Performance of the CNN Model**

* **Better Results**: The CNN model showed better accuracy in predicting resistance to several TB drugs (such as ethambutol, rifampicin, pyrazinamide, isoniazid, and ofloxacin) compared to existing rule-based methods. It measured success with a performance metric called **F1-scores**, which ranged from **81.1% to 98.2%**, depending on the drug.
* **Comparison**: The CNN outperformed the **Mykrobe predictor**, a well-known rule-based tool, in both accuracy and stability.

### **6. Feature Selection**

* **AMR Features**: The study used a technique called **filter-based feature selection** to identify the most important genetic features related to drug resistance. Most of these features matched known drug resistance-related mutations in the **CARD database** and **WHO catalogue**.

### **7. Sharing the Work**

* To make the model easy to use, the researchers packaged all the steps of building and training the model into an **automated pipeline**. They also shared the source code online, allowing others to use or improve the model for predicting drug resistance.

### **Summary:**

The researchers developed a new **1D CNN deep learning model** to predict drug resistance in tuberculosis. It outperformed existing methods by using a large global dataset and integrating various data features. The model achieved high accuracy for several anti-TB drugs and identified important genetic features related to resistance. The entire process was made available online for others to use and improve.

Antimicrobial resistance (AMR) is recognized as one of the greatest concerns for public health globally[1](https://www.nature.com/articles/s41598-022-06449-4#ref-CR1). Previous work estimated that the deaths attributable to antimicrobial resistance might rise from the current estimate of 700,000 lives per year to ten million annually by 2050[2](https://www.nature.com/articles/s41598-022-06449-4#ref-CR2). The prevalence of bacterial strains’ resistance to antibiotics has reduced the efficacy of antibiotics treatment dramatically[3](https://www.nature.com/articles/s41598-022-06449-4#ref-CR3), which leads to the urgent need for antimicrobial susceptibility testing to guide the treatment of antibiotics for serious bacterial infections. The conventional culture-based methods have limitations including extended turnaround time for slow-growing bacteria such as Mycobacterium tuberculosis (MTB) and bias due to potential contamination. MTB remains the world’s most deadly infectious disease, with an estimated 1.5 million deaths in 2019[4](https://www.nature.com/articles/s41598-022-06449-4#ref-CR4). The currently recommended treatment for drug-susceptible TB disease is a 6-month course of four first-line drugs: isoniazid (INH), rifampicin (RIF), ethambutol (EMB) and pyrazinamide (PZA)[5](https://www.nature.com/articles/s41598-022-06449-4#ref-CR5). As resistance to first-line drugs has become more prevalent, second-line drugs were developed to treat first-line drug-resistant TB disease, which requires a course of second-line drugs for at least nine months and up to 20 months[4](https://www.nature.com/articles/s41598-022-06449-4#ref-CR4). The emergence of drug-resistant TB continues to threaten global TB control efforts. The World Health Organization reported that nearly half a million people developed rifampicin-resistant TB (RR-TB), of which 78% had multidrug-resistant TB (MDR-TB) around the world in 2019[4](https://www.nature.com/articles/s41598-022-06449-4#ref-CR4). There is an urgent need to rapidly identify drug sensitivity profiles of TB, given the fact that culture-based diagnostic tests are usually time-consuming.

This passage highlights the critical challenge of **antimicrobial resistance (AMR)**, particularly focusing on **tuberculosis (TB)**, and emphasizes the need for rapid methods to identify drug resistance. Here's an explanation in simpler terms:

### **1. Antimicrobial Resistance (AMR) as a Global Health Threat:**

* **AMR** is a major global health concern, as bacteria are becoming resistant to antibiotics.
* Currently, **700,000 deaths** per year are attributed to drug-resistant infections, and this number is expected to increase to **10 million deaths annually by 2050** if the issue is not addressed.

### **2. Impact on Antibiotic Treatment:**

* Many bacterial strains have developed resistance to antibiotics, making it much harder to treat infections. This has dramatically reduced the effectiveness of traditional antibiotic treatments.
* To address this, there is an urgent need for **antimicrobial susceptibility testing**, which helps doctors determine which antibiotics will be effective against a particular bacterial infection.

### **3. Challenges with Traditional Testing Methods:**

* **Conventional culture-based methods** for testing bacterial resistance have limitations, especially with slow-growing bacteria like **Mycobacterium tuberculosis (MTB)**.
  + These methods take a long time to produce results and can be prone to contamination, leading to inaccurate results.

### **4. Tuberculosis (TB) and Drug Resistance:**

* **TB** remains one of the most deadly infectious diseases, causing **1.5 million deaths in 2019**.
* **Treatment for drug-susceptible TB** involves a 6-month course of four main drugs: isoniazid (INH), rifampicin (RIF), ethambutol (EMB), and pyrazinamide (PZA).
* However, as resistance to these first-line drugs has become more common, second-line drugs are needed, which require longer treatments (9-20 months).

### **5. Drug-Resistant TB (MDR-TB and RR-TB):**

* Drug-resistant TB is a growing problem, making it more difficult to control the disease globally.
* In 2019, nearly **500,000 people** were diagnosed with **rifampicin-resistant TB (RR-TB)**, and of these, **78%** also had **multidrug-resistant TB (MDR-TB)**, which means they were resistant to both rifampicin and isoniazid.

### **6. Need for Rapid Drug Sensitivity Testing:**

* Identifying which drugs can still effectively treat TB is critical, but traditional **culture-based tests** take a long time to produce results.
* There is an **urgent need** for faster diagnostic methods that can quickly determine the **drug sensitivity profile** of TB, allowing for more effective and timely treatment of drug-resistant cases.

### **Summary:**

Antimicrobial resistance is a growing global health crisis, and TB is one of the most affected diseases. Drug-resistant TB is a serious challenge, and current methods for diagnosing drug resistance are too slow. Faster testing methods are urgently needed to guide effective treatment and control the spread of resistant TB strains.

To overcome these restrictions and identify antibiotic resistance more efficiently, researchers use conventional association rule methods to predict antimicrobial resistance[6](https://www.nature.com/articles/s41598-022-06449-4#ref-CR6). These methods are based on the identification of variants associated with AMR from whole genome sequencing (WGS) data. The WGS data from clinical strains has been curated in dedicated databases including the Comprehensive Antibiotic Resistance Database (CARD)[7](https://www.nature.com/articles/s41598-022-06449-4#ref-CR7) and the Pathosystems Resource Integration Center (PATRIC) [8](https://www.nature.com/articles/s41598-022-06449-4#ref-CR8).

Traditional machine learning (ML) algorithms, e.g., support vector machine (SVM), logistic regression (LR) and random forests (RF), have been compared with variant-based association rules for AMR prediction using WGS data of pathogen isolates in recent years[9](https://www.nature.com/articles/s41598-022-06449-4#ref-CR9),[10](https://www.nature.com/articles/s41598-022-06449-4#ref-CR10). Yang et al. developed and compared different traditional ML methods using a cohort of 1839 UK MTB isolates for the prediction of resistance on eight anti-TB drugs. Kouchaki et al. trained their models by using a dataset of over 13,402 isolates for more stable prediction on seen and unseen samples[10](https://www.nature.com/articles/s41598-022-06449-4#ref-CR10). Three basic ML classifiers based on the feature space after dimension reduction and three ensemble learning methods were considered on this dataset. Another study conducted by Zhang et al. investigated deep learning strategy by using 2D Convolutional Neural Network (CNN) on whole-genome sequencing data of 149 MTB isolates for resistance classification on a less studied drug PZA[11](https://www.nature.com/articles/s41598-022-06449-4#ref-CR11). Variants were called by aligning reads on a single reference genome H37Rv. Although ML, including deep learning, has been applied to the prediction of AMR, most studies used a limited number of isolates collected from a specific area, and all of them used single strain reference when detecting variants instead of using pan-genome reference[12](https://www.nature.com/articles/s41598-022-06449-4#ref-CR12),[13](https://www.nature.com/articles/s41598-022-06449-4#ref-CR13), which could result in poor mapping and variant calling quality in new strains. The use of a pan-genome reference can decrease errors in the mapping and variant detection process, especially for more diverged strains.

This passage explains how researchers are trying to improve the prediction of **antimicrobial resistance (AMR)** by using **machine learning (ML)** and data from **whole genome sequencing (WGS)**. Here’s a breakdown in simpler terms:

### **1. Problem with Traditional Methods**

* **Antibiotic resistance** is becoming a major problem, and traditional methods for identifying which bacteria are resistant to antibiotics are slow and not always efficient.
* To improve this process, researchers are using **association rule methods**, which look at genetic variations (called **variants**) in bacteria’s DNA. These variants can help predict which bacteria are resistant to certain antibiotics by using data from **whole genome sequencing (WGS)**.
* **WGS** gives a complete picture of the bacteria's genetic makeup. Data from bacterial strains has been stored in databases like **CARD (Comprehensive Antibiotic Resistance Database)** and **PATRIC (Pathosystems Resource Integration Center)**, which researchers use to find patterns linked to resistance.

### **2. Using Machine Learning for Better Prediction**

* In recent years, researchers have applied **machine learning (ML)** to predict AMR more efficiently. **ML algorithms** can learn from data and improve their predictions over time.
* Common ML methods used include:
  + **Support Vector Machine (SVM)**
  + **Logistic Regression (LR)**
  + **Random Forests (RF)**
* These ML algorithms have been tested against traditional association rule methods using WGS data from bacteria to see which method predicts resistance better.

### **3. Examples of Studies Using Machine Learning:**

* **Yang et al.**: This study used a group of **1,839 tuberculosis (TB)** samples from the UK to train ML models and predict resistance to **eight anti-TB drugs**. The researchers compared different ML methods to see which worked best.
* **Kouchaki et al.**: They used a much larger dataset of **13,402 TB samples** to train their models, allowing them to make more reliable predictions, even for new data they hadn’t seen before. They tested different **ML classifiers** (which are like tools that help the computer make decisions) and **ensemble learning methods** (which combine multiple models for better accuracy).
* **Zhang et al.**: This study applied a more advanced type of ML called **deep learning** using a method called **2D Convolutional Neural Network (CNN)**. They focused on a lesser-known TB drug, **pyrazinamide (PZA)**, and used WGS data from **149 TB samples** to predict resistance.

### **4. Limitations of Previous Studies:**

* **Limited Datasets**: Most of these studies used a small number of bacterial samples collected from a specific region. This limits the generalizability of the models because the bacteria from one area might be different from those in other areas.
* **Single Reference Genome**: In these studies, the WGS data was compared to a single reference genome (usually the standard **H37Rv** strain). However, this can lead to problems when analyzing new strains that are different from the reference, as it can result in errors in mapping the variants and lower the quality of the predictions.

### **5. Benefits of Using Pan-Genome Reference**

* **Pan-Genome Reference**: To fix this issue, researchers suggest using a **pan-genome reference**, which includes genetic information from **multiple strains** of bacteria, not just one.
* This approach makes it easier to find genetic variants accurately, especially in bacteria that are more genetically diverse. It helps reduce errors in identifying the parts of the genome linked to drug resistance and improves the overall quality of the prediction.

### **Summary:**

Researchers are using **machine learning** and **deep learning** to predict antibiotic resistance in bacteria using **whole genome sequencing data**. Studies have shown that these methods can improve the accuracy of predictions, but most research so far has used limited datasets and relied on a single reference genome, which can lead to errors when working with diverse bacterial strains. Using a **pan-genome reference**, which incorporates genetic information from multiple bacterial strains, can reduce these errors and improve the prediction of antimicrobial resistance.

Here, we present our study of MTB drug resistance classification using traditional ML methods (LR and RF) and a deep neural network architecture of 1D CNN on a large and diverse dataset of MTB isolates. To compare the performance of our ML classifiers with a state-of-the-art statistical modeling method Mykrobe predictor, we evaluated the accuracy of Mykrobe predictor on the same dataset[14](https://www.nature.com/articles/s41598-022-06449-4#ref-CR14). Mykrobe predictor uses a De Bruijn graph representation of bacterial diversity to identify species and resistance profiles of clinical isolates for Staphylococcus aureus and Mycobacterium tuberculosis. We used a dataset of 10,575 MTB isolates[15](https://www.nature.com/articles/s41598-022-06449-4#ref-CR15), which is imbalanced with more susceptible isolates than resistant ones for all four first-line drugs mentioned above and four second-line drugs: amikacin (AMK), capreomycin (CM), kanamycin (KM) and ofloxacin (OFX). To reduce computation, we performed feature selection first to reduce the dimensions of input data and applied multi-input 1D CNN. Instead of using a single strain reference, we used all references from CARD database[16](https://www.nature.com/articles/s41598-022-06449-4#ref-CR16), even including references of other bacteria to build reference clusters as a pan-genome reference. Sequencing reads were then aligned to these reference clusters for variant detection. The results showed that our best ML classifiers outperformed the state-of-the-art rule-based method Mykrobe predictor, especially for EMB resistance, and showed more stable accuracy to all the four first-line drugs. Although our basic 1D CNN architecture didn’t significantly outperform our traditional ML methods LR and RF, there are potential ways to optimize it in the future, e.g., hyperparameter tuning.

This passage explains a study that compares different **machine learning (ML) methods** and a **deep learning architecture** for predicting drug resistance in **Mycobacterium tuberculosis (MTB)** using a large and diverse dataset of MTB samples. Here's a simpler and more detailed explanation of what's happening:

### **1. The Study’s Goal**

The researchers wanted to **classify drug resistance** in **MTB** using two traditional machine learning methods:

* **Logistic Regression (LR)**
* **Random Forest (RF)**

They also used a more advanced deep learning model called a **1D Convolutional Neural Network (1D CNN)**.

The goal was to compare the performance of these methods with a well-known statistical modeling tool called the **Mykrobe predictor**.

### **2. What is the Mykrobe Predictor?**

* **Mykrobe predictor** is a tool that uses a method called a **De Bruijn graph** to represent the genetic diversity of bacteria. It helps to:
  + Identify the species of the bacteria.
  + Determine its resistance to antibiotics.
* The Mykrobe predictor has been used to identify antibiotic resistance for bacteria like **Staphylococcus aureus** and **Mycobacterium tuberculosis**.

### **3. The Dataset**

* The researchers used a dataset of **10,575 MTB samples**, which contained genetic information from these bacterial samples.
* The dataset was **imbalanced**, meaning there were more samples of **drug-susceptible** (non-resistant) bacteria than **drug-resistant** bacteria. This imbalance was seen for both:
  + **First-line drugs**: These are the primary drugs used to treat TB (like isoniazid, rifampicin, ethambutol, and pyrazinamide).
  + **Second-line drugs**: These drugs (like amikacin, capreomycin, kanamycin, and ofloxacin) are used when the bacteria are resistant to first-line drugs.

### **4. The Methodology**

* **Feature Selection**: Since the dataset contains a lot of genetic information, the researchers first performed **feature selection**. This step helps to reduce the amount of data by selecting only the most relevant genetic features for predicting drug resistance. This makes the models run faster and reduces the complexity.
* **Pan-Genome Reference**: Instead of using a single reference genome (which is the standard practice), the researchers built a **pan-genome reference** using genetic information from many different bacteria. They used data from the **CARD database**, which contains information about antibiotic resistance from multiple bacterial species, not just MTB. This helped create **reference clusters** that better represent the genetic diversity of the bacteria.
* **Variant Detection**: Once they had the reference clusters, they compared the genetic data (sequencing reads) from the samples against these clusters to detect **variants** (mutations). These variants help identify whether the bacteria are resistant to drugs.

### **5. The Results**

* **ML Methods vs. Mykrobe Predictor**: The researchers found that their best **machine learning classifiers** (LR, RF, and CNN) outperformed the **Mykrobe predictor**, especially when it came to predicting resistance to the drug **ethambutol (EMB)**.
* **Stability Across Drugs**: The machine learning models were also more consistent in their accuracy for predicting resistance to the **first-line TB drugs** (isoniazid, rifampicin, ethambutol, and pyrazinamide).
* **1D CNN Performance**: Although the **1D CNN** deep learning model was expected to outperform the traditional methods (LR and RF), it did not show a significant improvement. However, the researchers believe that with further **optimization**, like **hyperparameter tuning** (adjusting the model’s settings to improve performance), the CNN model could potentially perform better in the future.

### **Summary:**

The study compared different machine learning and deep learning models to predict **drug resistance in tuberculosis**. The models used a **large dataset** and a **pan-genome reference** to improve the accuracy of predictions. The machine learning models, especially **logistic regression** and **random forests**, performed better than the **Mykrobe predictor**, especially for the drug **ethambutol**. The **1D CNN** deep learning model didn’t outperform the traditional methods in this study but has the potential to do so with further improvements.

## **Methods**

### 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[17](https://www.nature.com/articles/s41598-022-06449-4#ref-CR17) 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[15](https://www.nature.com/articles/s41598-022-06449-4#ref-CR15). 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[15](https://www.nature.com/articles/s41598-022-06449-4#ref-CR15). 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 (Fig. [1](https://www.nature.com/articles/s41598-022-06449-4#Fig1)) 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.

This passage explains how the researchers collected and prepared the **training data** and **labels** for their machine learning (ML) model to predict drug resistance in **Mycobacterium tuberculosis (MTB)**. Here’s a step-by-step breakdown:

### **1. Data Sources:**

* **Whole-genome sequencing (WGS) data**: This is the genetic data of the MTB samples, downloaded from the **Sequence Read Archive (SRA) database**, which stores large amounts of genetic information.
* **Drug Susceptibility Test (DST) data**: These are the lab results that show whether each MTB sample is resistant or susceptible to certain drugs. The DST data was collected from the **CRyPTIC Consortium** and the **100,000 Genomes Project**.
  + The **lineage data** refers to the genetic background of each MTB sample, helping understand the genetic differences among them.
  + The **phenotypic DST results** were used as **labels** for the ML model. In this context, "labels" are the known outcomes (resistant or susceptible) that the model will learn to predict.

### **2. Dataset Overview:**

* The dataset consists of **10,575 MTB isolates** (individual bacterial samples).
* The **CRyPTIC Consortium** and **100,000 Genomes Project** provided both the genetic data and the drug resistance information for these isolates.

### **3. Imbalanced Dataset:**

* The dataset is **imbalanced**, meaning most of the MTB isolates are **susceptible** to the drugs, and only a small portion are **resistant**. This makes it harder for the model to learn about the resistant cases because they are in the minority.

### **4. Drug Susceptibility Data:**

* The **four first-line drugs** tested are:
  + **Ethambutol (EMB)**
  + **Isoniazid (INH)**
  + **Pyrazinamide (PZA)**
  + **Rifampicin (RIF)**
* For these first-line drugs, the number of isolates with drug susceptibility test results is as follows:
  + **EMB**: 7,138 isolates
  + **INH**: 7,137 isolates
  + **PZA**: 6,347 isolates
  + **RIF**: 7,081 isolates
* There are **6,291 shared isolates** that have results for all four first-line drugs.

### **5. Second-line Drugs:**

* The dataset also includes data for **four second-line drugs** (used when MTB is resistant to first-line drugs).
* **6,820** of the isolates have drug susceptibility test results available for all four second-line drugs.

### **Summary:**

The researchers collected **genomic data** from 10,575 MTB samples and corresponding **drug susceptibility test (DST) data** from major sources. The DST data was used as **labels** to train and evaluate machine learning models, but the dataset is imbalanced, meaning most samples are susceptible to the drugs, and only a few are resistant. The analysis focused on the four **first-line TB drugs** (EMB, INH, PZA, and RIF), with additional data for **second-line drugs**. This dataset allows the model to learn how to predict drug resistance based on genetic information.

To detect the potential genetic features that could contribute to MTB drug resistance classification, we used a command-line tool called ARIBA[18](https://www.nature.com/articles/s41598-022-06449-4#ref-CR18). 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.

This passage explains how the researchers used a tool called **ARIBA** to detect potential genetic features that could contribute to **drug resistance classification** in **Mycobacterium tuberculosis (MTB)**. Here’s a detailed breakdown of the steps involved:

### **1. What is ARIBA?**

* **ARIBA** is a command-line tool used to quickly and accurately analyze genetic data for **antimicrobial resistance (AMR)**. It’s designed to process genomic data and identify **genetic variants** (mutations) that are linked to resistance to antibiotics.
* ARIBA is flexible, meaning it can be customized to detect specific genetic features, and it produces detailed outputs, making it suitable for MTB drug resistance studies.

### **2. Steps in the Process Using ARIBA:**

#### **Step 1: Downloading Reference Data**

* **Reference Data from CARD**: The researchers started by downloading reference data from the **Comprehensive Antibiotic Resistance Database (CARD)**.
  + CARD includes known genetic sequences linked to **antibiotic resistance** from various bacterial species, not just **MTB**. For example, the data includes references from other bacteria like **Staphylococcus aureus**.
  + This reference data contains information about **AMR-associated genes** and their known mutations, which are used to identify resistance patterns in MTB.

#### **Step 2: Clustering Reference Sequences**

* After downloading the reference data, the researchers **clustered the reference sequences** based on their **genetic similarity**.
  + Clustering means grouping similar sequences together. This helps reduce the complexity of the data by creating **reference clusters** of related sequences rather than treating each individual sequence separately.
  + The resulting **clusters** represent similar genetic sequences from multiple bacterial strains, including MTB and other bacteria.

#### **Step 3: Using the Pan-Genome Reference**

* The **pan-genome reference** was built using these clustered sequences. A **pan-genome reference** is a comprehensive collection of all possible genetic variations found across different bacterial strains.
  + In this case, it includes sequences from MTB strains as well as other bacteria. By using this pan-genome reference, the researchers can better detect mutations and variations in the MTB genome that might not be visible using a single reference genome.

#### **Step 4: Aligning Read Pairs**

* The next step was to take the **read pairs** (genomic data from the WGS of each MTB isolate) and align them to the **pan-genome reference clusters**.
  + **Read pairs** are sequences generated from the DNA of an isolate (a single MTB sample). By aligning these reads to the reference clusters, the researchers can see which part of the genome the sequences match.

#### **Step 5: Local Assemblies and Variant Identification**

* For each reference cluster that had reads mapped to it, the researchers ran **local assemblies**. This means they assembled the aligned reads into a sequence to see how closely they match the reference cluster.
  + **Identifying Variants**: They then compared the assembled reads to the closest reference sequence in that cluster and looked for **genetic variants** (mutations). These variants are the differences between the isolate’s genetic data and the reference sequence.

### **3. Output Files Generated by ARIBA**

ARIBA generates several useful output files after running the alignment and analysis:

#### **a. Summary File for Alignment Quality**

* This file contains information about the **quality of the alignment** between the read pairs and the reference clusters. It helps assess how well the data matched the reference sequences.

#### **b. Report File for Detected Variants and AMR Genes**

* This file includes detailed information about the **detected genetic variants** and the **antimicrobial resistance (AMR) genes** found in the samples. It is a key output used to identify mutations linked to drug resistance in MTB.

#### **c. Read Depth File**

* For each reference cluster where reads were mapped, the **read depth file** provides counts of how many times each of the **four DNA bases (A, T, C, G)** was observed at every position (locus) of the closest reference.
  + **Read depth** refers to how many times a particular part of the genome was sequenced. Higher read depth means more confidence in the accuracy of the sequence at that position.

### **Why is This Process Important?**

* This entire process helps the researchers identify specific **genetic features (mutations or variants)** that are associated with drug resistance in MTB. By aligning the data to a pan-genome reference and clustering similar sequences, they can detect even small variations that could influence resistance to antibiotics.
* The use of ARIBA allows for **rapid** and **flexible** detection of these variants, making it easier to analyze large amounts of genetic data and find meaningful patterns that can improve drug resistance classification models.

### **Summary:**

The researchers used the ARIBA tool to analyze genetic data from MTB samples. They downloaded reference data from the CARD database, clustered the sequences, and created a pan-genome reference. They then aligned the genomic data from the MTB isolates to these reference clusters, performed local assemblies, and identified genetic variants. ARIBA produced several output files, including a summary of alignment quality, a report on the detected variants and AMR genes, and read depth information. This process helps identify genetic mutations that contribute to MTB drug resistance.

Next, we filtered out low-quality mappings that did not pass the ‘match’ criteria defined in ARIBA’s GitHub wiki[18](https://www.nature.com/articles/s41598-022-06449-4#ref-CR18). 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.

This passage describes how the researchers **refined their data** by filtering and selecting important **genetic features** that will be used in the machine learning model to predict drug resistance in **Mycobacterium tuberculosis (MTB)**. Here’s a detailed explanation:

### **1. Filtering Low-Quality Mappings**

* **Mappings** refer to how the genetic sequences (reads) from the MTB samples align to the reference data (genomic sequences from the CARD database).
* The researchers used **ARIBA** to align the sequences, but some alignments (mappings) might be of low quality.
* They **filtered out** those low-quality mappings that did not meet specific **‘match’ criteria**, which are defined in **ARIBA’s guidelines** on its GitHub page. This step ensures that only **reliable** and **accurate mappings** are used in the analysis.

### **2. Collecting High-Quality Genetic Features**

* From the remaining **high-quality mappings**, the researchers extracted **genetic features** that are linked to drug resistance. They focused on the following types of features:
  + **Novel Variants in Coding Regions**: These are **newly identified mutations** (variants) in the parts of the genome that encode proteins (coding regions). These variants might influence drug resistance.
  + **Well-Studied Resistance-Causing Variants**: These are **previously known mutations** that are already well understood to cause drug resistance. For example, specific mutations in certain genes are known to confer resistance to drugs like rifampicin or isoniazid.
  + **AMR-Associated Gene Presences**: The **AMR-associated genes** (antimicrobial resistance genes) are specific genes in **Mycobacterium tuberculosis (MTB)** and other bacteria that, when present or mutated, can make the bacteria resistant to certain antibiotics. These genes encode proteins that can either prevent the antibiotic from working or modify bacterial functions in a way that the drug becomes ineffective.

### **3. Selection of Genetic Features**

* The researchers collected **263 genetic features** from the data. These features are important because they represent the variations in the MTB genome that could potentially be linked to **drug resistance**.

### **4. Lineage Indicator Variables**

* In addition to the genetic features, the researchers also included **indicator variables** for each of the **19 lineages** of MTB.
  + **Lineages** are genetic subgroups or strains of MTB that differ from each other. Including **lineage information** as a feature helps the model learn how different lineages might behave differently in terms of drug resistance.
  + **Indicator variables** are used to represent the presence or absence of each lineage in the data. This way, the model can use lineage information as an additional factor in its predictions.

### **5. Total Features**

* By combining the **263 genetic features** with the **19 lineage indicators**, the researchers created a **feature vector** (a list of features used to train the model) with a total of **282 features**.
  + These **features** are the key pieces of information that the machine learning model will use to predict whether an MTB isolate is resistant or susceptible to a specific drug.

### **Summary:**

The researchers filtered out low-quality genetic mappings and focused on high-quality genetic features, including **novel mutations**, **known resistance-causing variants**, and **AMR-associated genes**. They identified 263 important genetic features from the MTB samples. Additionally, they included 19 **lineage indicator variables** to account for differences between genetic subgroups of MTB. Altogether, they created a **feature set** of 282 features, which will be used to train the machine learning model to predict **drug resistance**.

### Feature selection and 1D CNN models

CNN is a class of deep neural networks that takes multi-dimensional data as input[23](https://www.nature.com/articles/s41598-022-06449-4#ref-CR23). When we say CNN, generally, we refer to a 2-dimensional CNN, which is often used for image classification. However, there are two other types of CNN used in practice: 1-dimensional and 3-dimensional CNNs. Conv1D is generally used for time-series data where the kernel moves on one dimension and the input and output data are 2-dimensional. Conv2d and 3D kernels move on two dimensions and three dimensions, respectively.

(Explanation not required)

Because deep learning algorithms require substantial computational power, we performed feature selection to only keep relevant features as input for deep learning algorithms. First, we randomly selected 80 percent of samples to calculate the importance of each feature by using the scikit-learn RF feature importance function that averages the impurity decrease from each feature across the trees to determine the final importance of each variable[24](https://www.nature.com/articles/s41598-022-06449-4#ref-CR24). Then, we tuned the feature importance cutoff to find the one that maximizes the F1-score of an RF model trained on the remaining 20 percent of samples. For each of the eight drugs, features were selected when their feature importance scores were bigger than the optimal cutoff. The tuning processes for first-line drugs are visualized in Fig. [2](https://www.nature.com/articles/s41598-022-06449-4#Fig2).

This passage explains how the researchers used a process called **feature selection** to narrow down the number of features (genetic characteristics) that would be used as input for their **deep learning models**. Deep learning algorithms require a lot of computational power, so reducing the number of features helps make the process more efficient while keeping the most important information. Here's a detailed breakdown of what was done:

### **1. Why Feature Selection?**

* **Deep learning algorithms** are computationally expensive, especially when working with large datasets that have many features.
* To improve efficiency, the researchers performed **feature selection**, which means they kept only the most important features while removing less relevant ones.

### **2. Random Selection of Samples (80% Training, 20% Testing)**

* First, they divided the dataset into two parts:
  + **80% of the samples** were used to calculate the importance of each feature.
  + The remaining **20% of samples** were set aside to test the model.

### **3. How Feature Importance Was Calculated:**

* The researchers used a method from the **scikit-learn** library called **Random Forest (RF) feature importance**.
  + **Random Forest** is a machine learning algorithm that works by creating a group of decision trees and averaging their predictions.
  + **Feature importance** in RF is calculated by measuring how much a feature contributes to reducing "impurity" (or uncertainty) when making decisions in each tree. If a feature consistently helps in making good decisions, its importance score will be higher.
  + The feature importance is calculated by averaging the **impurity decrease** (how much uncertainty is reduced) from each feature across all the trees in the forest.

### **4. Tuning the Feature Importance Cutoff:**

* After calculating the importance scores for each feature, the researchers needed to decide which features to keep and which to discard. To do this:
  + They **tuned the feature importance cutoff**, which is the threshold that determines which features are kept based on their importance score.
  + The goal was to find the cutoff that **maximized the F1-score** of the **Random Forest (RF) model** when tested on the remaining 20% of the samples.
    - **F1-score** is a measure of a model's accuracy, balancing precision (how many selected features are relevant) and recall (how many relevant features are selected).

### **5. Applying Feature Selection for Each Drug:**

* The researchers repeated this process for each of the **eight drugs** they were studying.
  + They kept features whose **importance scores** were higher than the **optimal cutoff** (the threshold that gave the best F1-score).
  + By selecting only the most important features, they improved the performance of the model while reducing computational cost.

### **6. Visualization of the Tuning Process (Fig. 2):**

* For the **first-line drugs**, the researchers created visualizations (shown in **Figure 2**) that illustrate how they adjusted the feature importance cutoff and how this affected the F1-score of the RF model. This helps show the relationship between the number of selected features and the model’s performance.

### **Summary:**

The researchers reduced the number of features (genetic characteristics) by calculating **feature importance** using the **Random Forest (RF) algorithm** from the scikit-learn library. They first used 80% of the samples to determine the importance of each feature, and then fine-tuned the threshold for keeping features by maximizing the **F1-score** on the remaining 20% of the samples. This process was repeated for each of the eight drugs, and the results were visualized for the first-line drugs to show how feature selection improved model performance while reducing computational demand.

These **cutoff values** represent the **thresholds for feature importance** used to determine which features (genetic variants or other factors) are considered relevant for predicting **drug resistance** for each of the four first-line drugs.

* The **cutoff** is a minimum importance score that a feature must have to be included in the final model.
* Features with an importance score **above the cutoff** are considered important for predicting resistance, while those **below** are excluded.
* Each drug has its own optimal cutoff, meaning the importance of genetic features varies by drug:
  + **EMB**: Features with an importance score greater than 0.0004 are used.
  + **INH**: The cutoff is 0.0006.
  + **PZA**: The cutoff is 0.0008.
  + **RIF**: The cutoff is 0.0016.

These cutoffs help improve model performance by selecting the most relevant features for each drug.

After the relevant features were selected, we designed and built a multi-input CNN architecture with TensorFlow Keras[25](https://www.nature.com/articles/s41598-022-06449-4#ref-CR25) that took N inputs of 4 × 21 matrices representing N selected SNP features into the first layer. Each 4 × 21 matrix consists of normalized DNA base counts for each locus within a 21-base reference sequence window centered on the focal SNP (Fig. [3](https://www.nature.com/articles/s41598-022-06449-4#Fig3)). We generated normalized counts based on the raw base counts extracted from the read depth file mentioned in “[Genetic feature extraction](https://www.nature.com/articles/s41598-022-06449-4#Sec4)” section. Our convolutional architecture starts with two 1D convolutional layers followed by a flattening layer for each SNP input. Then, it concatenates the N flattening layers with the inputs of AMR-associated gene presence and lineage features. Finally, we added three fully connected layers to complete the deep neural network architecture (Fig. [4](https://www.nature.com/articles/s41598-022-06449-4#Fig4)). It smoothly integrates sequential and non-sequential features.

Single nucleotide polymorphism, or SNP. If you are reading a news story where it says, for example, scientists find the genetic contributors to diabetes or some other condition or trait, you're probably reading about SNPS. A SNP is a one-letter place where your genome varies from another genome sequence. Thanks to the Human Genome Project, we have found that these single letter changes in our genetic code are placed all across our genomes. We can see that the patterns vary between people and even between populations. If we want to identify genetic contributors to a common complex disease like diabetes, we can group together thousands of people who have diabetes and compare their SNP patterns to thousands of people who do not have diabetes. With enough people in our study, we can use the SNPs as markers to see that certain areas of the genome appear to be the same in people who have diabetes, and that tells us where we should look in more detail for a genetic cause.

This passage explains how the researchers designed a **multi-input Convolutional Neural Network (CNN)** to predict drug resistance based on **selected SNP features** and other genetic data. Here’s a step-by-step breakdown:

### **1. Input Structure:**

* The model takes **N inputs**, where each input represents an **SNP feature** in the form of a **4 × 21 matrix**.
  + **4 × 21 matrix**: This matrix contains the **normalized DNA base counts** (A, T, C, G) for a **21-base sequence** centered around each **SNP**. The 21-base window helps the model capture the surrounding genetic context of the SNP.
  + **Normalized base counts**: These counts are generated from the **raw DNA base counts** using the **read depth file** from previous genetic feature extraction, which shows how many times each base (A, T, C, G) was observed.

### **2. CNN Architecture:**

* The model starts with two **1D convolutional layers** for each SNP input. These layers help detect patterns in the DNA sequence around the SNPs, which could be important for identifying **drug resistance**.
  + **1D Convolutional Layers**: These layers extract features from the 21-base sequence around each SNP.
* After the convolutional layers, a **flattening layer** is used to convert the 2D matrix (SNP feature) into a 1D vector, making it easier to combine with other features later.

### **3. Combining Features:**

* The model then **concatenates** (combines) the outputs from the **N flattening layers** (representing SNP features) with other important inputs:
  + **AMR-associated gene presence**: This input represents whether key **antimicrobial resistance (AMR) genes** are present in the sample.
  + **Lineage features**: These inputs indicate the bacterial lineage, which can affect resistance.

### **4. Fully Connected Layers:**

* After combining the SNP features, AMR-associated gene presence, and lineage features, the model passes this combined input through three **fully connected layers**. These layers help the model learn complex relationships between the features and predict drug resistance.

### **5. Purpose of the Architecture:**

* The architecture is designed to handle both **sequential features** (SNPs in a DNA sequence) using the CNN layers and **non-sequential features** (like AMR genes and lineage) using the fully connected layers. This combination allows the model to integrate different types of genetic information and make better predictions.

### **Summary:**

The researchers built a **multi-input CNN** that takes in SNP data (as 4×21 matrices of DNA base counts), processes it through 1D convolutional layers to extract patterns, and combines this information with the presence of **AMR genes** and **lineage data**. The final part of the model consists of fully connected layers that make the drug resistance predictions by integrating all these features.

## **Results**

### Isolate identification and DST phenotype

### To explore genetic information obtained by running the ARIBA steps listed in “[Genetic feature extraction](https://www.nature.com/articles/s41598-022-06449-4#Sec4)” section, we calculated the numbers of isolates matched on different reference clusters (Fig. [5](https://www.nature.com/articles/s41598-022-06449-4#Fig5)a) and generated a circular phylogenetic tree with lineage and phenotypic DST data annotations (Fig. [5](https://www.nature.com/articles/s41598-022-06449-4#Fig5)b).

### This passage describes how the researchers explored the genetic data obtained from the ARIBA tool. They used two different methods to visualize and interpret the results:

### **1. Isolate Identification via Reference Clusters (Fig. 5a):**

### **Figure 5a** shows the number of isolates (MTB samples) that matched different **reference clusters**. Each reference cluster represents a group of similar genetic sequences (from both MTB and other bacteria) used to detect drug resistance-related mutations.

### The **x-axis** in the figure represents the number of isolates, while the **y-axis** lists the different clusters (e.g., genes or loci associated with drug resistance).

### **Higher counts** indicate that many isolates matched a particular reference cluster, implying that these clusters might be important in predicting drug resistance.

### **2. Circular Phylogenetic Tree (Fig. 5b):**

### **Figure 5b** shows a **circular phylogenetic tree**, which is a visual representation of the evolutionary relationships between different MTB isolates based on their genetic data.

### Each branch represents a different **lineage** (subgroup) of MTB, and the colors correspond to various **lineages** (e.g., Beijing, lineage4, Ghana) as shown in the legend.

### The outer layer of the tree is annotated with **phenotypic Drug Susceptibility Test (DST) data**, which shows resistance to different drugs:

### **Pink**: Ethambutol (EMB) resistance

### **Yellow**: Isoniazid (INH) resistance

### **Green**: Pyrazinamide (PZA) resistance

### **Purple**: Rifampicin (RIF) resistance

### This visualization allows the researchers to see how **lineage** and **drug resistance phenotypes** are related and identify clusters of isolates that are more likely to be resistant to specific drugs.

### **Summary:**

### **Figure 5a** provides an overview of how many isolates matched specific reference clusters, highlighting potentially important genetic features related to drug resistance.

### **Figure 5b** shows the evolutionary relationships between isolates and their drug resistance profiles, using a circular phylogenetic tree to link MTB lineages with resistance to specific drugs.

### \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SOME NOT SO IMP STUFF RN\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **Conclusions**

**AMR infection is one of the major threats to human health. In silico methods are effective to predict drug resistance and a reliable alternative to in vitro assay that is much slower and more expensive. Statistical association rule and ML are two main types of in silico approaches. We developed ML models for first-line TB drug resistance classification on a large and diverse MTB isolate cohort to compare to a statistical rule-based method. The result shows our ML models are more accurate and stable for TB drug resistance prediction across the four first-line drugs than the rule-based method Mykrobe predictor. We designed and developed a customized 1D CNN architecture to adapt and combine sequential and non-sequential features. Even though our deep CNN models haven’t taken advantage of any optimization strategies (e.g., hyperparameter tuning), our CNN architecture slightly outperformed the other two traditional ML algorithms. As a result of variant analysis, 78.8% of variant features selected for our CNN model training are also identified as TB drug resistance-associated ones by WHO.**

*Subject: Summary and Key Insights on MTB Drug Resistance Study*

*Dear [Recipient's Name],*

*I hope this email finds you well. I wanted to share a summary of the recent study we discussed, which focuses on predicting Mycobacterium tuberculosis (MTB) drug resistance using a multi-input deep learning model.*

*The study involves the collection of whole-genome sequencing (WGS) data for 10,575 MTB isolates, alongside phenotypic Drug Susceptibility Test (DST) results from the CRyPTIC Consortium and the 100,000 Genomes Project. The dataset, although imbalanced with more drug-susceptible isolates, provided a robust basis for training various machine learning models.*

*To detect genetic features relevant for predicting drug resistance, the team employed ARIBA, a genotyping tool, which allowed the extraction of SNP features and AMR-associated gene presences. These genetic features, along with lineage data, were used as inputs for a multi-input CNN architecture designed with TensorFlow Keras. This architecture integrates sequential features (SNP data) and non-sequential features (lineage and gene presence) to create a comprehensive model.*

*An essential part of the study was the feature selection process, where feature importance was calculated using random forest models. The team tuned cutoffs for feature selection, optimizing the F1-score for different drugs: ethambutol (EMB), isoniazid (INH), pyrazinamide (PZA), and rifampicin (RIF).*

*The analysis also included the construction of a circular phylogenetic tree to visualize the relationships between MTB lineages and their drug resistance phenotypes. This tree highlighted how certain lineages (e.g., Beijing and lineage4) were associated with multi-drug resistance, reinforcing the importance of combining lineage data with genetic features in the model.*

*In summary, the study's findings contribute significantly to understanding the genetic factors behind drug resistance in MTB and show the potential of using deep learning models for better prediction. The next steps will likely involve optimizing the model further and exploring its application in clinical settings.*

*I’d be happy to discuss this in more detail and explore any potential collaboration opportunities. Looking forward to your thoughts!*

*Best regards,  
Gauranga*