ML Informed Drug Resistance in MTB

Authors

Gauranga Kumar Baishya - MDS202325

Siddhesh Maheshwari - MDS202347

Instructor

Dr. Bikul Das

Principal Investigator,
Kavikrishna Laboratory,
IIT Guwahati

February 11, 2025

Acknowledgement

This work is based on the project titled *ML Informed Drug Resistance in MTB*, supervised by *Dr. Bikul Das* and *Shirsajit Mitra*, as part of the Industry Project undertaken in January 2025.

Our sincere gratitude goes to *Dr. Bikul Das* and *Shirsajit Mitra* for their invaluable guidance and support throughout the project, shaping its direction and focus.

Our gratitude also goes to all the participants who generously shared their time and insights, making this collaborative endeavor possible. Their contributions enriched our research and facilitated meaningful conclusions.

Contents

1 Introduction

1.1 Background on Tuberculosis (TB)

Tuberculosis (TB) is a highly infectious disease caused by the bacterium *Mycobacterium tuberculosis*. It primarily affects the lungs, though it can also spread to other parts of the body, including the brain, kidneys, and spine. TB spreads through the air when an infected person coughs, sneezes, or talks, making it a significant public health concern worldwide. According to the World Health Organization (WHO), TB remains one of the top ten causes of death globally and is particularly prevalent in low- and middle-income countries.

Despite being a preventable and curable disease, TB's persistence stems from challenges such as delayed diagnosis, incomplete treatment, and the emergence of drug-resistant strains. Multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB) have further complicated the treatment landscape, making the disease harder to manage and increasing the need for innovative diagnostic and therapeutic approaches.

The WHO's End TB Strategy, launched in 2014, aims to reduce TB deaths by 90% and cases by 80% by 2030. Achieving these ambitious goals requires a comprehensive understanding of the disease's biology, rapid diagnostic techniques, effective treatments, and public health interventions. This project focuses on addressing one critical aspect of the TB challenge: understanding and predicting drug resistance mechanisms through the application of machine learning techniques to genomic data.

1.2 Basics of Genetics

Genetics is the study of heredity and the variation of traits in living organisms, primarily determined by the molecular structure and function of DNA (Deoxyribonucleic Acid). DNA serves as the fundamental blueprint for life, containing the genetic instructions necessary for the growth, development, functioning, and reproduction of all known organisms.

DNA is composed of two long chains of nucleotides twisted into a double helix. Each nucleotide consists of three components: a sugar molecule (deoxyribose), a phosphate group, and a nitrogenous base. The sequence of nitrogenous bases—adenine (A), thymine (T), cytosine (C), and guanine (G)—encodes genetic information. Adenine pairs with thymine, and cytosine pairs with guanine, forming the base-pairing rules essential for DNA replication and transcription.

Within DNA, specific sequences known as *genes* act as instructions for synthesizing proteins, which are essential macromolecules that perform a vast array of functions within cells. Proteins serve as structural components, enzymes that catalyze biochemical reactions, and signaling molecules that regulate biological processes. Each gene encodes a specific protein, and variations in gene sequences can lead to differences in protein function, ultimately influencing traits and susceptibility to diseases.

The concept of the genetic code underpins the translation of genetic information from DNA to functional proteins. Through a process called transcription, the genetic information in DNA is copied into messenger RNA (mRNA). This mRNA is then translated into proteins in a process called translation, wherein ribosomes assemble amino acids in the sequence specified by the mRNA template.

In the context of *Mycobacterium tuberculosis* (MTB), mutations in specific genes can lead to significant changes in protein function, often resulting in drug resistance. For instance, mutations in the *rpoB* gene, which encodes the RNA polymerase β-subunit, can disrupt the binding of the antibiotic rifampicin, rendering it ineffective. Similarly, compensatory mutations in other genes, such as *rpoC*, can offset the fitness costs associated with drug resistance mutations, allowing the bacterium to survive and thrive despite antibiotic treatment.

Advances in genetics, particularly in the field of genomics, have enabled researchers to analyze entire genomes of organisms like MTB. Techniques such as whole-genome sequencing (WGS) provide a comprehensive view of an organism's genetic makeup, allowing the identification of mutations associated with drug resistance. By leveraging this genetic information, machine learning algorithms can be applied to predict resistance patterns, develop diagnostic tools, and design effective treatment regimens.

Understanding the basics of genetics is fundamental to addressing the challenge of drug-resistant TB, as it forms the foundation for analyzing genetic mutations, understanding their biological consequences, and devising innovative strategies to combat the disease.

1.3 Importance of Addressing Drug Resistance in MTB

Drug resistance in *Mycobacterium tuberculosis* (MTB) poses a significant global health challenge, particularly in the fight against tuberculosis (TB). TB is one of the top ten causes of death worldwide, and the emergence of drug-resistant strains has made it increasingly difficult to manage and control the disease. Multidrug-resistant TB (MDR-TB), defined as resistance to at least isoniazid and rifampicin—the two most effective first-line antibiotics—accounts for a growing percentage of TB cases. Additionally, extensively drug-resistant TB (XDR-TB), which exhibits resistance to even more drugs,

has exacerbated the crisis.

Drug resistance arises due to genetic mutations within the bacterial genome, often as a result of incomplete or improper antibiotic use. For instance, mutations in the *rpoB* gene confer resistance to rifampicin, a cornerstone of TB treatment. Such resistance not only renders existing treatments ineffective but also complicates the management of TB by requiring longer, more toxic, and more expensive treatment regimens.

The implications of drug-resistant TB are far-reaching:

- **Increased Mortality and Morbidity:** Drug-resistant TB is associated with higher death rates and prolonged illness compared to drug-susceptible TB.
- **Economic Burden:** Treating drug-resistant TB is significantly costlier, both for healthcare systems and affected individuals. It often involves lengthy hospital stays, expensive second-line drugs, and extensive diagnostic testing.
- **Public Health Threat:** Drug-resistant TB is a highly transmissible disease, posing a significant risk to communities and undermining global TB control efforts.

Addressing drug resistance in MTB is critical to achieving the World Health Organization's (WHO) End TB Strategy, which aims to reduce TB deaths by 90% and cases by 80% by 2030. This goal necessitates significant advancements in understanding drug resistance mechanisms, improving diagnostic capabilities, and developing effective treatment strategies. Early detection of resistant strains through molecular diagnostics, coupled with personalized treatment approaches, can help prevent the further spread of resistance.

Moreover, understanding compensatory mutations, such as those in the *rpoC* gene that mitigate the fitness cost of *rpoB* mutations, is vital for designing better therapeutic interventions. Research into these genetic mechanisms, particularly through advanced tools like machine learning and whole-genome sequencing, offers a promising pathway to combat drug-resistant TB.

The importance of addressing drug resistance in MTB extends beyond the immediate challenges of treatment. It represents a critical step toward safeguarding global health, reducing the burden of TB, and achieving the long-term goal of eradicating the disease.

2 Drug Resistance in MTB

Drug resistance in *Mycobacterium tuberculosis* (MTB) is a significant global health challenge that complicates the treatment and control of tuberculosis (TB). It occurs when the bacteria develop the ability to survive despite the administration of antibiotics. This resistance is primarily caused by genetic mutations or, in rare cases, the acquisition of resistance genes from other microbes. The emergence of multidrug-resistant TB (MDR-TB), resistant to at least isoniazid and rifampicin (two key first-line antibiotics), and extensively drug-resistant TB (XDR-TB), which is resistant to additional drugs, has further exacerbated the TB crisis.

The mechanisms of drug resistance often involve mutations in specific genes that affect the targets of antibiotics. For example, resistance to rifampicin, one of the most effective first-line drugs for tuberculosis (TB), is a prime example. Rifampicin targets the bacterial RNA polymerase by binding to its β-subunit, encoded by the *rpoB* gene. Mutations in the *rpoB* gene prevent rifampicin from binding effectively, rendering the drug unable

to inhibit RNA synthesis. These mutations are a hallmark of multidrug-resistant TB (MDR-TB), as they often occur alongside mutations conferring resistance to other drugs, such as isoniazid.

Isoniazid resistance is another major contributor to MDR-TB. Isoniazid requires activation by the bacterial enzyme catalase-peroxidase, encoded by the *katG* gene, to exert its antibacterial effects. Mutations in *katG*, particularly at codon 315, reduce the enzyme's ability to activate the drug, thereby conferring resistance. Additionally, mutations in the promoter region of the *inhA* gene, which encodes the target of isoniazid, can further impair the drug's efficacy.

In addition to these specific mechanisms, MTB can develop resistance to other drugs through mutations in genes encoding drug targets or enzymes involved in drug metabolism. For instance, mutations in the *pncA* gene confer resistance to pyrazinamide, while mutations in *embB* are associated with ethambutol resistance.

Beyond target-based mechanisms, MTB also exhibits intrinsic resistance mechanisms that contribute to its drug-resistant phenotype. The bacterial cell wall, which is rich in mycolic acids, acts as a physical barrier that limits the penetration of antibiotics. Efflux pumps, which actively expel antibiotics from the bacterial cell, further reduce the intracellular concentration of drugs, decreasing their efficacy.

The interplay of resistance mutations with compensatory mutations further complicates the treatment of drug-resistant TB. Resistance mutations often impose a fitness cost on the bacteria, reducing their growth or survival. Compensatory mutations, such as those in the *rpoC* gene, restore bacterial fitness while maintaining resistance. This dual adaptation allows drug-resistant strains to persist and propagate even in the absence of selective pressure from antibiotics.

Understanding the genetic and molecular mechanisms of drug resistance in MTB is critical for developing effective diagnostic tools and therapeutic strategies. Molecular diagnostics, such as whole-genome sequencing, can rapidly identify resistance-associated mutations, enabling tailored treatment regimens. Additionally, targeting compensatory mechanisms offers a potential avenue for combating resistant strains and enhancing the efficacy of existing drugs.

2.1 Mechanisms of Drug Resistance in MTB

Drug resistance in *Mycobacterium tuberculosis* (MTB) is primarily driven by genetic mutations that enable the bacteria to survive and grow despite exposure to antibiotics. These mutations can alter the structure or function of the antibiotic's target, reduce drug efficacy, or enhance bacterial survival mechanisms. The following outlines the key mechanisms of drug resistance in MTB:

Resistance to rifampicin, one of the most effective first-line drugs for tuberculosis (TB), is a prime example. Rifampicin targets the bacterial RNA polymerase by binding to its β-subunit, encoded by the *rpoB* gene. Mutations in the *rpoB* gene prevent rifampicin from binding effectively, rendering the drug unable to inhibit RNA synthesis.

Isoniazid resistance is another major contributor to MDR-TB. Isoniazid requires activation by the bacterial enzyme catalase-peroxidase (encoded by the *katG* gene) to exert its antibacterial effects. Mutations in *katG*—especially at codon 315—reduce the enzyme's ability to activate the drug. In addition, mutations in the promoter region of the *inhA* gene, which encodes the drug's target, can further impair its efficacy.

MTB may also develop resistance to other drugs through mutations in genes encoding

drug targets or enzymes involved in drug metabolism. For example, mutations in the *pncA* gene confer resistance to pyrazinamide, while mutations in *embB* are associated with resistance to ethambutol.

2.2 Compensatory Mutations and Their Significance

Compensatory mutations play a critical role in the evolution and persistence of drug-resistant *Mycobacterium tuberculosis* (MTB). While resistance mutations allow MTB to survive antibiotic treatment, they often come at a significant cost to bacterial fitness. This fitness cost manifests as reduced growth rates, impaired survival, or diminished competitive ability when compared to drug-susceptible strains. Compensatory mutations serve to offset these disadvantages, enabling drug-resistant strains to thrive and spread more effectively.

One of the most well-studied examples involves resistance to rifampicin, a cornerstone drug for tuberculosis (TB) treatment. Resistance to rifampicin arises from mutations in the *rpoB* gene, which encodes the RNA polymerase β-subunit. These mutations prevent rifampicin from binding effectively to its target but simultaneously reduce the efficiency of the bacterial RNA polymerase, leading to a fitness cost. To mitigate this, compensatory mutations often occur in the *rpoC* gene, which encodes the RNA polymerase β'-subunit. These mutations restore the functionality of RNA polymerase, allowing rifampicin-resistant strains to maintain high levels of growth and survival while preserving their resistance.

The presence of compensatory mutations is significant for several reasons:

- **Persistence of Drug-Resistant Strains:** Compensatory mutations enhance the fitness of drug-resistant strains, enabling them to persist and propagate even in the absence of antibiotic pressure. This contributes to the long-term maintenance of resistant strains in populations and complicates TB control efforts.

- **Increased Transmission Risk:** By restoring bacterial fitness, compensatory mutations make drug-resistant strains more transmissible, increasing the risk of community-wide outbreaks of resistant TB.

- **Diagnostic Challenges:** Compensatory mutations may obscure the detection of drug resistance, as traditional diagnostic methods may not account for the interplay between resistance and compensatory mechanisms. This highlights the need for comprehensive molecular diagnostics.

- **Therapeutic Implications:** Understanding compensatory mutations provides insights into the biology of drug-resistant MTB, offering opportunities to develop targeted therapies. By addressing both resistance and compensatory pathways, novel treatment strategies could be designed to reduce bacterial viability and combat resistant strains effectively.

From an evolutionary perspective, compensatory mutations underscore the adaptability of MTB in response to selective pressures. They illustrate how drug resistance is not an isolated phenomenon but a dynamic process involving a balance between resistance, fitness costs, and compensatory mechanisms. This adaptability highlights the need for robust public health interventions, such as improved diagnostic tools and more effective treatment regimens, to counteract the spread of resistant TB.

Research into compensatory mutations, particularly through whole-genome sequencing and computational analysis, has shed light on the genetic interactions that underpin drug resistance. Advanced tools like machine learning can analyze genomic data to predict the emergence of compensatory mutations, guiding the development of precision treatments. Ultimately, addressing compensatory mutations is crucial for tackling drug-resistant TB and achieving global TB control goals, such as the World Health Organization's End TB Strategy.

Compensatory mutations play a critical role in the evolution and persistence of drug-resistant *Mycobacterium tuberculosis* (MTB). While resistance mutations allow MTB to survive antibiotic treatment, they often come at a significant cost to bacterial fitness. This fitness cost manifests as reduced growth rates, impaired survival, or diminished competitive ability when compared to drug-susceptible strains. Compensatory mutations serve to offset these disadvantages, enabling drug-resistant strains to thrive and spread more effectively.

One of the most well-studied examples involves resistance to rifampicin, a cornerstone drug for tuberculosis (TB) treatment. Resistance to rifampicin arises from mutations in the *rpoB* gene, which encodes the RNA polymerase β-subunit. These mutations prevent rifampicin from binding effectively to its target but simultaneously reduce the efficiency of the bacterial RNA polymerase, leading to a fitness cost. To mitigate this, compensatory mutations often occur in the *rpoC* gene, which encodes the RNA polymerase β'-subunit. These mutations restore the functionality of RNA polymerase, allowing rifampicin-resistant strains to maintain high levels of growth and survival while preserving their resistance.

The presence of compensatory mutations is significant for several reasons:

- **Persistence of Drug-Resistant Strains:** Compensatory mutations enhance the fitness of drug-resistant strains, enabling them to persist and propagate even in the absence of antibiotic pressure. This contributes to the long-term maintenance of resistant strains in populations and complicates TB control efforts.

- **Increased Transmission Risk:** By restoring bacterial fitness, compensatory mutations make drug-resistant strains more transmissible, increasing the risk of community-wide outbreaks of resistant TB.

- **Diagnostic Challenges:** Compensatory mutations may obscure the detection of drug resistance, as traditional diagnostic methods may not account for the interplay between resistance and compensatory mechanisms. This highlights the need for comprehensive molecular diagnostics.

- **Therapeutic Implications:** Understanding compensatory mutations provides insights into the biology of drug-resistant MTB, offering opportunities to develop targeted therapies. By addressing both resistance and compensatory pathways, novel treatment strategies could be designed to reduce bacterial viability and combat resistant strains effectively.

From an evolutionary perspective, compensatory mutations underscore the adaptability of MTB in response to selective pressures. They illustrate how drug resistance is not an isolated phenomenon but a dynamic process involving a balance between resistance, fitness costs, and compensatory mechanisms. This adaptability highlights the need for robust public health interventions, such as improved diagnostic tools and more effective treatment regimens, to counteract the spread of resistant TB.

Research into compensatory mutations, particularly through whole-genome sequencing and computational analysis, has shed light on the genetic interactions that underpin drug resistance. Advanced tools like machine learning can analyze genomic data to predict the emergence of compensatory mutations, guiding the development of precision treatments. Ultimately, addressing compensatory mutations is crucial for tackling drug-resistant TB and achieving global TB control goals, such as the World Health Organization's End TB Strategy.

3 Methodology

3.1 Research Goal and Data Collection

The primary goal of this research is to investigate the interrelationship between mutations in the *rpoB* and *rpoC* genes of *Mycobacterium tuberculosis* (MTB) and to develop predictive models to identify compensatory mutations that enhance the fitness of drug-resistant strains. These mutations are critical in understanding the evolution and persistence of multidrug-resistant tuberculosis (MDR-TB), especially in the context of rifampicin resistance. By uncovering patterns in genetic mutations, this research aims to contribute to improved diagnostic and therapeutic approaches for combating drug-resistant TB.

Rifampicin resistance in MTB is primarily caused by mutations in the *rpoB* gene, which encodes the RNA polymerase β-subunit. These mutations allow the bacteria to evade the inhibitory effects of rifampicin but often come at a cost to bacterial fitness. Compensatory mutations in the *rpoC* gene, encoding the RNA polymerase β'-subunit, can mitigate this fitness cost, allowing resistant strains to survive and propagate more effectively. This study focuses on understanding the genetic interactions between specific mutations in *rpoB* and *rpoC* and developing machine learning models to predict these compensatory relationships.

Data Collection: The success of this research hinges on the availability of high-quality genomic data from diverse populations. Whole-genome sequencing (WGS) data from MTB isolates was obtained from two key sources:

- **PHLTA, Israel:** This dataset includes 233 whole-genome sequences of MTB strains, representing a wide range of resistance profiles and genetic variations.
- **Argentina:** Another 117 whole-genome sequences were collected from MTB strains in Argentina, providing additional diversity in resistance patterns and mutations.

These datasets provide comprehensive information on mutations across the MTB genome, including those in *rpoB* and *rpoC*. The combination of data from different geographic regions ensures that the analysis accounts for genetic diversity and variability in resistance mechanisms across populations.

The raw sequencing data underwent preprocessing to prepare it for analysis. This involved quality control, alignment of sequencing reads to a reference genome, removal of duplicates, and base recalibration. The resulting high-quality data was used to identify single nucleotide polymorphisms (SNPs) and insertions/deletions (Indels) associated with resistance and compensatory mutations.

By leveraging this robust dataset and advanced computational tools, this research aims to:

1. Identify and characterize specific *rpoB* mutations associated with rifampicin resistance.

2. Explore the role of *rpoC* mutations in compensating for the fitness cost of *rpoB* resistance mutations.

3. Develop predictive models to understand the relationship between resistance and compensatory mutations using machine learning techniques.

The insights gained from this research are expected to inform the development of improved diagnostic tools and targeted treatment strategies for MDR-TB, addressing a critical challenge in global tuberculosis control.

3.2 Analysis Pipeline and Tools

The analysis pipeline for this study was designed to process whole-genome sequencing (WGS) data of *Mycobacterium tuberculosis* (MTB) isolates and identify mutations associated with drug resistance and compensatory mechanisms. This multi-step workflow ensures high-quality genomic data, enabling accurate identification of genetic variants and downstream analyses. The pipeline includes the following steps:

1. Raw Data Processing: The WGS data, consisting of raw sequencing reads in FASTQ format, was subjected to initial quality control checks to ensure the integrity of the data. Poor-quality reads and adapter sequences were removed using tools such as `FastQC` and `Trimmomatic`, resulting in cleaned reads suitable for downstream analysis.

2. Mapping to the Reference Genome: The cleaned reads were aligned to the *H37Rv* reference genome of MTB using `bwa mem`, a widely used alignment algorithm. This step determines the exact genomic locations of the sequencing reads, enabling the identification of variations. The resulting Sequence Alignment/Map (SAM) files were converted into the more compact Binary Alignment/Map (BAM) format using `SAMtools` for efficient storage and processing.

3. Sorting and Duplicate Marking: The BAM files were sorted by genomic coordinates using `SAMtools`, ensuring that reads are organized for subsequent analyses. To eliminate potential biases introduced during library preparation and PCR amplification, duplicate reads were identified and marked using the `Picard` toolkit. Removing duplicates prevents overrepresentation of certain regions in the genome, improving the accuracy of variant calling.

4. Base Recalibration: Base quality scores, assigned by sequencing machines, are often prone to errors. To correct these inaccuracies, base quality recalibration was performed using the `GATK (Genome Analysis Toolkit)`. This step adjusts base quality scores based on known variants in the reference genome, improving the confidence and accuracy of variant detection.

5. Variant Calling: Variant calling was conducted using the `GATK HaplotypeCaller`, which identifies single nucleotide polymorphisms (SNPs) and insertions/deletions (Indels) in the genomic data. The resulting Variant Call Format (VCF) files contain detailed information about the genetic variations present in each MTB isolate.

6. Refinement of Variants: To ensure the reliability of the identified variants, a series of filtering steps were applied to remove low-quality or false-positive variants. Filters were based on parameters such as read depth, variant quality, and allelic frequency, resulting in a high-confidence dataset of SNPs and Indels.

7. Analysis-Ready Data: The filtered variants were compiled into final analysis-ready datasets. These datasets were used to identify resistance-associated mutations in the *rpoB* gene and compensatory mutations in the *rpoC* gene. The analysis-ready data was also used as input for machine learning models to predict mutation relationships.

8. Data Analysis and Visualization: Genomic data was analyzed to explore the interrelationships between mutations in the *rpoB* and *rpoC* genes. Visualization tools such as `R` and `Python` libraries (`matplotlib`, `pandas`, and `seaborn`) were used to generate mutation frequency plots, co-occurrence matrices, and other visualizations to identify patterns and trends.

Key Tools Used in the Pipeline: The analysis was supported by a combination of specialized bioinformatics tools and computational frameworks:

- `FastQC` and `Trimmomatic`: Quality control and preprocessing of raw sequencing reads.
- `bwa mem`: Alignment of sequencing reads to the reference genome.
- `SAMtools` and `Picard`: BAM file processing, sorting, and duplicate marking.
- `GATK`: Base recalibration, variant calling, and variant filtering.
- `Python` and `R`: Statistical analysis, data visualization, and model implementation.

Significance of the Pipeline: This comprehensive pipeline ensures the generation of high-quality, analysis-ready genomic data, enabling robust identification of resistance and compensatory mutations. The use of advanced bioinformatics tools and rigorous quality control measures enhances the reliability of the findings, providing a solid foundation for downstream machine learning and predictive modeling. By identifying key genetic variations and understanding their interactions, this pipeline supports the overarching goal of developing targeted interventions for drug-resistant TB.

3.3 Predictive Models and Rationale

The prediction of compensatory mutations and their relationship with resistance mutations in *Mycobacterium tuberculosis* (MTB) is a critical component of this research. Understanding these genetic interactions can improve diagnostic accuracy and guide personalized treatment for drug-resistant tuberculosis (TB). To achieve this, several machine learning models were explored for their ability to analyze the genomic data and predict the relationships between mutations in the *rpoB* and *rpoC* genes.

Predictive Models: The following predictive models were selected based on their suitability for analyzing genomic data and their performance in identifying complex patterns:

- **Association Rules:** This method is used to uncover relationships between mutations by identifying patterns of co-occurrence in the dataset. Association rules are particularly effective in determining if certain mutations in *rpoB* consistently appear alongside mutations in *rpoC*.
- **Logistic Regression:** A widely used statistical model for binary classification tasks, logistic regression was applied to predict the likelihood of compensatory mutations in *rpoC* given specific mutations in *rpoB*. This model provides interpretable results, making it suitable for understanding key mutation relationships.

- **Random Forest:** A robust ensemble learning method, random forest was utilized to capture non-linear relationships between mutations. By constructing multiple decision trees and aggregating their outputs, this model improves prediction accuracy and reduces the risk of overfitting.

- **Gradient Boosting:** This model builds sequential decision trees, each correcting the errors of its predecessor. Gradient boosting is effective for high-dimensional datasets, such as genomic data, and was applied to improve the precision of mutation predictions.

- **Neural Networks:** To capture complex, non-linear patterns in the genomic data, a neural network model was implemented. Neural networks are particularly powerful for analyzing large-scale datasets and identifying subtle interactions between mutations. A fully connected network was designed to predict compensatory mutations based on mutation profiles.

Rationale for Model Selection: The choice of models was guided by the unique characteristics of the genomic dataset and the objectives of the study:

- **Complex Interactions:** Genomic data is highly complex, with numerous possible interactions between mutations. Models such as random forest, gradient boosting, and neural networks were selected for their ability to capture non-linear relationships and intricate patterns in the data.

- **Interpretability:** While complex models offer high accuracy, interpretability is essential for understanding the biological relevance of predictions. Logistic regression and association rules provide interpretable insights, allowing researchers to identify specific mutation pairs and their effects.

- **Scalability:** The dataset, comprising genomic information from hundreds of MTB isolates, required scalable models capable of handling large volumes of data efficiently. Gradient boosting and neural networks were chosen for their scalability and performance on high-dimensional datasets.

- **Generalizability:** To ensure the models perform well on unseen data, techniques such as cross-validation and regularization were applied to prevent overfitting. Ensemble methods like random forest and gradient boosting were particularly valuable for improving model generalizability.

Evaluation Metrics: To assess the performance of the predictive models, the following metrics were employed:

- **Accuracy:** The proportion of correct predictions made by the model.

- **Precision and Recall:** Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the ability of the model to identify all true positives.

- **F1 Score:** A harmonic mean of precision and recall, providing a balanced evaluation of model performance.

- **ROC-AUC:** The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) evaluates the trade-off between true positive and false positive rates, offering a comprehensive measure of classification performance.

Significance of Predictive Models: The use of predictive models enables the identification of compensatory mutations that restore fitness in drug-resistant MTB strains. By combining advanced machine learning techniques with genomic data, this research provides valuable insights into the genetic basis of drug resistance and its compensatory mechanisms. These insights are critical for developing targeted diagnostic tools and therapeutic interventions to combat multidrug-resistant TB.

4 Results and Discussion

The results of this research provide valuable insights into the genetic mechanisms underlying drug resistance and compensatory mutations in *Mycobacterium tuberculosis* (MTB). The analysis leveraged whole-genome sequencing (WGS) data, advanced bioinformatics pipelines, and machine learning models to uncover patterns and relationships between mutations in the *rpoB* and *rpoC* genes.

1. Identification of Resistance Mutations: The variant calling process identified multiple mutations in the *rpoB* gene associated with rifampicin resistance. Among these, mutations at codons 516, 526, and 531 were the most prevalent, confirming their role as primary drivers of resistance. These mutations disrupt rifampicin binding to the RNA polymerase β-subunit, rendering the drug ineffective.

2. Compensatory Mutations in *rpoC*: Compensatory mutations in the *rpoC* gene were also identified, with significant occurrences at codons 332, 483, 491, and 525. These mutations mitigate the fitness costs associated with *rpoB* resistance mutations, enabling drug-resistant strains to thrive and propagate. The co-occurrence of *rpoB* and *rpoC* mutations highlights their functional interplay, with compensatory mutations restoring bacterial fitness while maintaining resistance.

3. Predictive Model Performance: The machine learning models used to predict the relationship between *rpoB* and *rpoC* mutations demonstrated varying levels of accuracy:

- Logistic regression provided interpretable results, identifying key mutation pairs and their likelihood of co-occurrence.
- Random forest and gradient boosting models achieved high accuracy and recall, effectively capturing non-linear interactions between mutations.
- Neural networks were particularly effective in analyzing large-scale genomic data, revealing complex patterns that were not apparent with simpler models.

The ensemble approaches, such as random forest and gradient boosting, outperformed other models in terms of overall prediction accuracy, achieving an F1 score above 0.9 and an ROC-AUC close to 0.95. These results demonstrate the utility of machine learning in analyzing genomic data and predicting compensatory mechanisms.

4. Biological Implications: The findings have significant implications for understanding drug resistance in MTB:

- The identification of compensatory mutations underscores the adaptability of MTB and highlights the importance of targeting these mutations in therapeutic strategies.

- The co-occurrence patterns of *rpoB* and *rpoC* mutations provide insights into the evolutionary pressures faced by MTB during antibiotic treatment.
- These insights can inform the design of molecular diagnostics that detect both resistance and compensatory mutations, enabling more effective treatment regimens.

5. Challenges and Limitations: Despite the promising results, several challenges were encountered:

- The limited sample size may affect the generalizability of the findings. Future studies with larger and more diverse datasets are needed to validate the results.
- While machine learning models performed well, the "black box" nature of neural networks limits the interpretability of their predictions. Combining neural networks with explainability techniques could address this issue.
- Compensatory mechanisms beyond *rpoC* mutations, such as those involving other genes or regulatory pathways, were not explored in depth and warrant further investigation.

6. Future Directions: Building on these results, the following directions are proposed for future research:

- Expanding the dataset to include additional geographic regions and resistance profiles to ensure a more comprehensive understanding of global resistance mechanisms.
- Investigating other potential compensatory mutations across the MTB genome and their role in restoring fitness.
- Integrating additional omics data, such as transcriptomics and proteomics, to provide a multi-layered perspective on resistance and compensation.
- Developing user-friendly software tools that incorporate machine learning models to aid clinicians in diagnosing and treating drug-resistant TB.

5 Conclusion

This research explored the genetic mechanisms underlying drug resistance and compensatory mutations in *Mycobacterium tuberculosis* (MTB), focusing on the interplay between mutations in the *rpoB* and *rpoC* genes. Using whole-genome sequencing (WGS) data and advanced bioinformatics tools, we identified key mutations associated with rifampicin resistance and their compensatory counterparts, which restore bacterial fitness.

The findings underscore the complexity of drug resistance in MTB. Mutations in the *rpoB* gene disrupt the binding of rifampicin to RNA polymerase, conferring resistance but imposing a fitness cost on the bacteria. Compensatory mutations in the *rpoC* gene mitigate this cost, enabling resistant strains to survive, propagate, and pose a greater public health challenge. By characterizing these genetic interactions, this study provides valuable insights into the evolutionary dynamics of drug-resistant TB.

Machine learning models played a pivotal role in uncovering patterns and predicting relationships between resistance and compensatory mutations. Models such as random forest and gradient boosting demonstrated high accuracy and reliability, highlighting the

potential of data-driven approaches in genomic research. These tools not only enhance our understanding of drug resistance mechanisms but also pave the way for the development of personalized diagnostic and therapeutic strategies.

Despite the promising results, several challenges remain. The limited dataset size and geographic scope may affect the generalizability of the findings, and further studies with larger, more diverse samples are needed. Additionally, the role of compensatory mutations beyond the *rpoC* gene warrants further exploration to gain a comprehensive understanding of fitness restoration mechanisms in drug-resistant MTB.

In conclusion, this research contributes to the global effort to combat multidrug-resistant TB by advancing our understanding of the genetic basis of resistance and compensation. The integration of genomic data with machine learning techniques offers a powerful framework for studying drug resistance, improving diagnostic tools, and informing treatment strategies. These findings align with the objectives of the World Health Organization's End TB Strategy, bringing us closer to the ultimate goal of eradicating tuberculosis.