Lincoln University Digital Thesis

Designing and testing holistic computational frameworks for identification of the most effective vaccine and drug targets against human and bovine tuberculosis

_____

A thesis

submitted in partial fulfilment

of the requirements for the Degree of

Doctor of Philosophy

at

Lincoln University

by

Pooja Pawar

_____

Lincoln University

2021

Abstract of a thesis submitted in partial fulfilment of the

requirements for the Degree of Doctor of Philosophy.

Designing and testing holistic computational frameworks for identification of the most effective vaccine and drug targets against human and bovine tuberculosis

by

Pooja Pawar

Abstract

The World Health Organization has considered tuberculosis (TB) a threat with a significant mortality and morbidity rate worldwide. TB is caused by notorious *Mycobacterium tuberculosis*, which has evolved with successful survival strategies leading to the emergence of drug-resistant TB strains making drugs (first-line TB drugs) and vaccine (BCG) ineffective. The global emergence of tuberculosis is threatening to make one of humankind's most lethal infectious diseases incurable, with an estimated 10.0 million new TB cases and 1.4 million deaths in 2019. Further, TB affects animals too; bovine tuberculosis primarily affects cattle and it is caused by the etiological agent *Mycobacterium bovis.* Twenty to thirty per cent of the global livestock population is potentially affected by bovine TB, leading to annual economic losses of more than USD 3 billion globally (Kuria, 2019). This study conducts an in-depth investigation into pathogen-human interactions to gain deeper insights into the evolution of pathogen and their drug resistance mechanisms and uses this understanding to provide potential solutions for effective vaccine and drug development for human and bovine tuberculosis.

Our research begins with gaining an understanding of the pathogenesis of human and bovine TB, the interaction of TB bacteria with its host, the host defence mechanism, bacterial survival strategies in evading the host immune response, and in-depth knowledge of the mechanisms of TB drug resistance. The current drug treatment regimen has not changed in nearly 40 years. Although the first-line drugs play a pivotal role in combating TB, the emergence of resistant TB strains due to different survival mechanisms of TB bacteria such as reduced permeability of cell wall preventing drug entry into the cells, mutations in the drug target protein (major hurdle in TB treatment), inactivation of drug molecules with the help of bacterial enzymes, and a transmembrane drug efflux system to expel the drug out from the bacterial cell has heightened the burden of TB globally. BCG is the only licensed

vaccine available and has been around for almost a hundred years. BCG (Bacillus Calmette-Guérin) is prepared from a live-attenuated strain of *Mycobacterium bovis* and it has shown protection in babies and young children. The inefficiency of BCG in not reducing the prevalence of disease and not protecting adults is so far not understood. Some of the crucial factors might include *Mycobacterium bovis* is less virulent and not a primary causative agent of TB, diversity in TB strains and over-attenuation of presently used BCG strain. The low efficacy of BCG, the emergence of the drug-resistant *Mycobacterium tuberculosis* strains, and challenges in developing drugs and vaccines have generated an urgent requirement for a powerful and effective therapeutic approach for TB treatment. This study introduces three holistic strategies/frameworks for developing new and effective therapeutic methods for fighting TB.

Firstly, the complex diversity of evolution of mutations demands a conceptual approach to tackling the problem at its root and eliminate it there. In this regard, a potential cue would be that fundamentally bacteria favour mutations that are harmless to them and provide maximum protection from the drug while making it ineffective or efflux it from the system. Therefore, when trying to map the pathogenicity of drug resistant bacteria, a promising line of attack is to assess the impact of the mutation on *Mycobacterium tuberculosis,* i.e., whether the mutation is neutral or harmful, and how effectively it weakens the efficacy of the drug binding to the target. In particular, *Mycobacterium tuberculosis* could favour mutations that afford them structural stability and flexibility and functional advantage to disempower the drug from its entry right up to reaching the target. Our study develops this conceptual foundation through a comprehensive and systematic in-depth analysis of drug resistance mechanisms from global mutation data for *Mycobacterium tuberculosis* reported over the last 30 years. We collected data on 821 non-synonymous mutations corresponding to the four first-line drugs isoniazid (n= 202), pyrazinamidase (n=273), rifampicin (n=120) and ethambutol (n=226) from 149 published studies. We first investigated single mutation frequency for understanding the prevalence and diversity of mutations in first-line TB drug targets across the globe. Our research then focussed on comprehensive coverage of mutations to identify their impact on TB bacteria and drug binding with a detailed bioinformatics analysis for understanding crucial changes at the molecular level affecting function, structural stability and sequence conservation and influence of mutation position on drug binding affinity. Our study introduced a new concept of ranking drug-resistant TB mutations into lethal, moderate, mild, and neutral. Out of 821 non-synonymous mutations, we identified 340 'lethal,' 284 'moderate,' 185 'mild' and 12 'neutral' mutations. We observed that the frequently occurring drug resistant mutations had mild to moderate impacts suggesting that *Mycobacterium tuberculosis* favours less harmful mutations suggesting that it tries to strike a balance between its fitness and drug resistance in evolving its survival strategy. The bacteria have followed this mutation

pattern over the last three decades globally and could continue to follow the same generic pattern across the globe in the future.

Vaccines are cost-effective pharmaceutical products that have played an essential role in eliminating and eradicating infectious diseases. The conceptual framework developed in this study uses different bioinformatics approaches, such as comparative proteome analysis, reverse vaccinology, immunoinformatics, and structural vaccinology to identify potential vaccine candidates and construct an in-silico epitope-based vaccine. Our epitope-based TB vaccine development focused on pathogen polymorphism, broad coverage of host population and enhancing humoral and cell-mediated immunity that would significantly reduce TB globally or drastically minimise its prevalence. Our holistic framework identified potential therapeutic candidates by directly analysing the proteome of TB bacterial strains. We performed a comparative proteomic analysis of 159 strains of *Mycobacterium tuberculosis* and 11 *Mycobacterium bovis* strains to cover the diversity and identify conserved proteins among those strains for developing human and bovine TB vaccines. An extensive reverse vaccinology and immunoinformatics analysis provided highly immunogenic, non-toxic and non-allergenic 27 epitopes (CTL epitopes-14, HTL epitopes-5 and B-cell epitopes-8) for *Mycobacterium tuberculosis* and 26 epitopes (CTL epitopes-8, HTL epitopes-2 and B-cell epitopes-16) for *Mycobacterium bovis* required for three-dimensional structure construction of TB vaccine constructs based on a new concept introduced in this research. The constructed epitope-based human and bovine TB vaccines had a strong interaction inside the host, thus activating the macrophages, further leading to the production of B-cells, T-cells and cytokines and generating efficient cell-mediated and humoral immune responses.

Next, a novel subtractive proteomic approach was developed for identifying bovine TB drug targets. We performed a subtractive proteomics approach on the 11 *Mycobacterium bovis* strains to identify drug targets that could further help investigate therapeutic drugs for the treatment of bovine TB. This approach helped identify nine drug targets that are conserved, essential, antigenic and have unique metabolic pathways in *Mycobacterium bovis*.

Our research makes novel contributions to the field of vaccine and drug development for tuberculosis. This research is intended to address the challenges of TB vaccine development that include: expensive, time-consuming and arduous experimental testing; safety concerns while culturing the pathogen in a laboratory; global coverage, identification of immunodominant and conserved epitopes in highly variable or drug-resistant *Mycobacterium tuberculosis* and *Mycobacterium bovis* for inducing potent humoral and cell-mediated immune responses and elimination of cross-reactive epitopes. It identified new drug targets for bovine TB. Better insight into drug resistance mechanisms will aid the development of novel diagnostic tools, design of new drugs and inhibitors, and help plan proper

treatment for the patients. Therefore, the outcome of this study's investigation into the global pattern of drug target mutations and drug resistance mechanisms can help significantly improve current treatment plans and develop new diagnostic techniques for treating TB in humans. As we used several bioinformatics prediction tools together to ensure checks and balances, aiming to reduce the chance of errors and provide accurate results, the vaccines and drug targets developed in this study can be tested experimentally with confidence for further validation as therapeutics with the potential to eradicate human and bovine TB globally.

Keywords: Tuberculosis, *Mycobacterium tuberculosis*, drug resistance mechanisms, bacterial survival strategy, reverse vaccinology, conceptual framework for vaccine development, epitope-based vaccine for human and bovine TB, *Mycobacterium bovis*, improved drug treatment of human TB, new drug targets for bovine TB.

Manuscripts under Preparation

Pawar, P., Samarasinghe, S. and Kulasiri, Don (2022). Towards eradication of TB globally: Comprehensive computational analysis for investigating drug resistance mechanisms of first-line TB drug resistant mutations. Scientific Reports.

Pawar, P., Samarasinghe, S. and Kulasiri, Don (2022). Designing and testing a new approach using computational vaccinology to identify the most lethal vaccine targets against tuberculosis. International Journal of Biological Macromolecules.

Pawar, P., Samarasinghe, S. and Kulasiri, Don (2022). Designing holistic frameworks for identification of therapeutic vaccine and drug targets against bovine tuberculosis. Genomics.

Acknowledgements

Table of Contents

List of Tables

List of Figures

Abbreviations

AAS	amino acid substitutions
AG	arabinogalactan
AMR	antimicrobial resistance
ANN	artificial neural network
APC	antigen-presenting cells
BCG	Bacillus Calmette-Guérin
BCR	B-cell receptor
BLAST	basic local alignment search tool
CAI	codon adaptation index
CMI	cell-mediated immunity
CNS	central nervous system
CRV	computational reverse vaccinology
CSM	cutoff scanning matrix
CTL	cytotoxic T-lymphocyte
CV	computational vaccinology
DBR	drug binding region
DEG	database of essential genes
DOPE	discrete optimized protein energy
DOTS	directly observed treatment short-course
DST	drug susceptibility testing
EMB	ethambutol
EPI	Expanded Program of Immunization
ESR	erythrocyte sedimentation rate
FAS	fatty acid synthetase
FDA	Food and Drug Administration
GDT-HA	global distance test- high accuracy
GRAVY	grand average of hydropathicity
GROMACS	Groningen Machine for Chemical Simulation
HLA	human leukocyte antigen
HMM	hidden Markov model
HP	hypothetical protein
HTL	helper T-lymphocyte
IEDB	immune epitope database
IFN-ɣ	interferon-gamma
IGRA	interferon gamma release assay
IL	interleukin
INH	isoniazid
KAAS	KEGG automatic annotation server
KEGG	Kyoto Encyclopaedia of Gene and Genome
LOMETS	local meta threading server
Man-LAM	mannose-capped lipoarabinomannan
MDR-TB	multidrug-resistant tuberculosis
MHC	major histocompatibility complex
MODS	microscopic observation of drug susceptibility
MSP	maximum segment pair
MTBC	*Mycobacterium tuberculosis* complex

MW	molecular weight
NAG	N-acetylglucosamines
NAM	N-acetylmuramic acid
NMA	normal mode analysis
NO	nitric oxide
ORF	open reading frame
PAMP	pathogen-associated molecular pattern
PDB	protein databank
PE	proline-glutamate
PG	peptidoglycan
PIM	phosphatidyl-myo-inositol mannoside
POA	pyrazinoic acid
PPD	purified protein derivative
PPE	proline-proline-glutamate
PSIC	position-specific indecent count
PSSM	position-specific scoring matrix
PZA	pyrazinamide
RIF	rifampicin
RNAP	DNA-dependent RNA polymerase
ROS	reactive oxygen species
RV	reverse vaccinology
SCL	subcellular localisation
SIFT	sorting intolerant from tolerant
SMART	simple modular architecture research tool
SPAAN	software for predicting adhesins and adhesin-like proteins using neural networks
STP	streptomycin
SVM	support vector machine
TACO	tryptophan-aspartate containing coat
TAT	twin-arginine translocase
TB	tuberculosis
TCR	T-cell receptors
TLR	toll-like receptor
TNF-α	tumour necrosis factor-alpha
WHO	World Health Organization
WT	wild-type
XDR-TB	extensively drug-resistant tuberculosis

Chapter 1

Introduction

Advancements in medical science and health care systems have doubled the life expectancy of humans and reduced the mortality rates of children. The death rate from infectious diseases has also been reduced over the past decade. But the emergence of new, and re-emergence of already existing infectious diseases and antimicrobial resistance (AMR), has become a global threat. In 2019, more than seven million people died of infectious or communicable diseases. There were also 2.5 million deaths caused by lower respiratory infections, including pneumonia, influenza and bronchitis that year. Tuberculosis (TB) has remained among the top causes of death in the past 20 years.

The emergence of rapidly evolving and lethal tuberculosis disease threatens to make one of humankind's most important infectious diseases incurable. Vaccine and drug therapy are the two most important countermeasures that humans have developed against TB. The complex evolution has led to the expansion of survival strategies in TB bacteria, and the emergence of drug-resistant TB strains made drugs and vaccine ineffective. In-depth knowledge of the pathogenesis of infection, antimicrobial resistance and challenges in vaccine and drug development will help ease the burden of global infectious diseases. Our study conducts an in-depth investigation into pathogen-human interactions to gain deeper insights into evolution of pathogen and their drug resistance mechanisms and use this understanding to provide potential solutions for effective vaccine and drug development for tuberculosis. This chapter discusses the research problem, the research objectives and gives an overview of the thesis chapters.

1.1 Tuberculosis - a deadly infectious disease

In 2020, TB was the 13th leading cause of death and the second leading cause of death from a single infectious agent after COVID-19. Regardless of the advancements made in the field of medical science, TB remains the cause of death for more than 1.4 million people every year (WHO Global Tuberculosis Report, 2020). TB is the leading cause of mortality in low- and middle-income countries and is considered a major health concern worldwide (World Health Statistics, 2021). Tuberculosis is an infection of the lungs caused by closely related bacterial species called the *Mycobacterium tuberculosis* complex (MTBC). The primary causative bacterium is *Mycobacterium tuberculosis.* Tuberculosis is a highly contagious disease that spreads from an infected person to a healthy person through air and inhaled airborne droplet nuclei. When another uninfected individual inhales these droplets, they travel down the person's trachea to the lungs. The TB bacteria multiply and activate the body's immune

response resulting in macrophages surrounding the bacteria to form a granuloma, thereby containing the spread of TB infection in the lungs (Hong et al., 2016).

In the early 19th century, TB was a widespread infection. First-line TB drugs for treating drug-susceptible TB patients are isoniazid (INH), ethambutol (EMB), pyrazinamide (PZA) and rifampicin (RIF) (Joshi, 2011). However, the evolution of TB bacteria has led to the emergence of drug-resistant TB strains that make drugs ineffective. Bacteria do this through various mechanisms including by making the cell wall impermeable preventing drug entry into the cells, mutations in the drug target protein, inactivation of drug molecules with the help of bacterial enzymes, and a transmembrane drug efflux system to expel the drug out from the bacterial cell. In 2019, an 85% success rate was observed in drug-susceptible TB patients, and the remaining 15% was accounted for by drug resistance and deaths (WHO Global Tuberculosis Report, 2020). Therefore, a better understanding of resistance mechanisms is essential for increasing the success rate of TB drugs to 98-100% to save lives. Furthermore, this understanding is key to improving current treatment strategies, designing new drugs and inhibitors, developing new diagnostic techniques and developing new vaccines.

Vaccines have been considered crucial in combating deadly and drug-resistant infectious diseases, such as TB, by reducing the prevalance of tuberculosis and eradicating TB from its roots. Currently, there is only one licensed vaccine available for TB, called BCG. This is prepared from a live-attenuated strain of *Mycobacterium bovis*, the causative agent of TB in cattle (Brosch et al., 2007). BCG has been around for almost a hundred years and was first used in 1921 (Dockrell & Smith, 2017). BCG is given to infants after birth in most parts of the world. This vaccine protects babies and young children; however, it fails to protect adults against TB (McShane, 2014). The inefficiency of BCG in reducing the prevalence and emergence of disease and in protecting adults is currently not well understood. Some critical factors might include *Mycobacterium bovis* being less virulent and not the principal causative agent of human TB, diversity in TB strains, and over-attenuation of the currently used BCG strain.

The widespread emergence of TB resistant strains has challenged the view of tuberculosis as a treatable disease. As of August 2020, 22 drugs and 14 vaccines were in different phases of clinical trials (WHO Global Tuberculosis Report, 2020). To accomplish the WHO's goal of eliminating TB with its End-TB strategy, there is a need for a vaccine that can stimulate a specific cell-mediated and humoral immune response in the host to eliminate the chances of infection and re-infection from tuberculosis. Vaccines in clinical trials are mainly dependent on inducing humoral immune responses using either weak mycobacterial strains or two to three antigens for producing recombinant subunits and viral vectored vaccines. Therefore, these vaccines do not guarantee yielding of a broad coverage of immune response. The drug-resistant issue is also not addressed adequately by the vaccines in clinical trials.

The low efficacy of BCG, re-emergence of the disease in immunocompetent individuals and the emergence of drug-resistant *Mycobacterium tuberculosis* strains have, together, generated an urgent requirement for a powerful and effective treatment against TB.

Bovine tuberculosis, caused by *Mycobacterium bovis*, has also become a global concern over the last two decades. Bovine TB primarily affects cattle, but other domestic livestock, such as deer, goats, horses, sheep, etc., are also affected. It is a zoonotic disease that can spread to humans by the inhalation of aerosols or by ingesting unpasteurized milk. Bovine TB is more common in less-developed and developing countries. Twenty to thirty per cent of the global livestock population is potentially affected by bovine TB. The disease can lead to an economic crisis because of the significant loss of livestock and trade restrictions. For example, in New Zealand, the beef and dairy industries are at potential risk from TB. Thus, bovine tuberculosis has a zoonotic potential to raise health concerns for the public (Cosivi, 1998; Renwick et al., 2007) Out of 10 million cases in humans, in 2019, WHO estimated that 0.14 million cases were zoonotic TB, caused by *Mycobacterium bovis*, with 11,400 human deaths (WHO Global Tuberculosis Report, 2020). Animal test-and-slaughter schemes implemented in several countries have successfully reduced the prevalence of bovine tuberculosis. However, such expensive control programmes have increased economic burdens and opposition by farmers (Bennett, 2009; Torgerson & Torgerson, 2010). There is no effective treatment available for bovine TB due to its infectious nature and the drug resistance of *Mycobacterium bovis*. Antibiotic therapy can be used on animals living in captivity, but it is not so reliable for herd or free-grazing animals. The BCG vaccine is another option available for treating the disease, but it shows limited efficacy in cattle. The prevention of bovine TB is a long-term goal that can only be accomplished by developing more effective vaccine than BCG and designing new drugs.

1.2   Formulating the research problem and solution

Global eradication of TB at its roots, requires a deep understanding of: (i) pathogenesis of TB, (ii) interaction of TB with its host, (iii) the host defence mechanism, (iv) bacterial survival strategies in evading the human immune response, (v) in-depth knowledge of the mechanisms of TB drug resistance, and (vi) challenges in developing drugs and vaccines for TB. Thus, for reducing the burden of TB globally, we developed holistic strategies/frameworks for accomplishing three main goals of this study:

(1) Understanding the impact of drug resistant mutations to gain an in-depth knowledge of the survival strategies of *Mycobacterium tuberculosis* and resulting drug resistance mechanisms

The main barrier to an effective TB treatment is the evolution of *Mycobacterium tuberculosis*, resulting in resistant strains. The available TB drugs are successfully evaded by the genetically-encoded resistance mechanisms in *Mycobacterium tuberculosis*. The rapidly increasing rates of mono-resistant (resistance to only one first-line TB drug), multi-drug resistant (resistance to at least isoniazid and rifampicin, two most important first-line TB drugs) and poly-resistant (resistant to more than one first-line TB drug, other than isoniazid and rifampicin), with 0.5 million new cases and 0.2 million deaths, in 2019, due to MDR-TB (multi-drug resistance), is a frightening situation resulting in treatment failures (WHO Global Tuberculosis Report, 2020) (Almeida Da Silva & Palomino, 2011). To understand *Mycobacterium tuberculosis* survival mechanisms that help it escape first-line TB drug actions, it is crucial to answer the following research questions:

(i) What are the effects of mutations on first-line TB drug targets and how serious is the problem?

Prolonged exposure to TB drugs helps *Mycobacterium tuberculosis* acquire spontaneous chromosomal mutations in the first-line drug target proteins, *katG*, *rpoB*, *pncA* and *embCAB,* that are targeted by the drugs, INH, RIF, PZA and EMB, respectively. Acquired resistance due to chromosomal mutations (mostly point mutations in the coding region of the target proteins (Robert & Pelletier, 2018) is the main reason for drug resistance in TB and is achieved through: altering the drug binding site of first-line TB target proteins, whole target modifications, drug inactivation and mutations in the conserved regions of the targets that makes the drugs ineffective but does not completely disrupt the biological function of the target (Alcaide et al., 1997; Almeida Da Silva & Palomino, 2011; Nguyen, 2016; Riccardi et al., 2009; Zhang & Yew, 2009). In light of the inadequate protection from the BCG vaccine, improvement in the success rate of drug therapy has become crucial for tackling TB. However, due to drug resistance, only an 85% success rate was observed in 2019 in the treatment of drug-susceptible TB patients. Therefore, a better understanding of drug resistance mechanisms is essential for increasing its success rate to 98-100% to save lives.

(ii) What are the gaps in drug resistance research?

Several studies have been conducted to understand the mechanism of drug resistance and the evolution of *Mycobacterium tuberculosis* (Telenti et al., 1993; Heym et al., 1995; da Cunha et al., 2007; Lakshmipathy et al., 2013). Most of the studies were conducted to determine the changes in drug binding affinity by constructing a 3D structure of wild-type target proteins and mutants, performing molecular docking using different tools, and comparing the wild-type target proteins' binding energy with the mutant proteins. However, these studies focussed on studying only a few mutations although, over the years, a large number of mutations have been reported. Further, by focusing only on drug

binding affinity, they have not provided significant knowledge of the evolution of drug resistant mutations and survival strategies including drug resistance mechanisms of *Mycobacterium tuberculosis.*

(iii) What can be done to reduce the problem of drug resistance?

Our study follows the process of *Mycobacterium tuberculosis* and TB drug interactions to probe deeply into how TB bacteria evolve drug resistance mechanisms for survival. In this regard, a potential clue would be that fundamentally bacteria favour mutations that help them survive by providing maximum protection from the drug while making the drug either ineffective or efflux it from the system. For a mutation to be a drug-resistant mutation, it should reduce the binding affinity of a drug without hampering the natural affinity of its substrate. Normal functioning and the structural stability of the target protein is needed for the survival of the disease-causing organism. Thus, it is essential to elucidate the effect of the first-line drug-resistant mutations in *Mycobacterium tuberculosis* on: (i) functional changes (leading to modification of target), (ii) stability changes leading to destabilizing or stabilizing effect on the target protein, (iii) irregular binding (reduced affinity for drug resulting in its inactivation) or tighter binding with the drug (the prodrug (an inactive biological compound which is metabolized inside the host body to make an active drug (Rautio et al., 2018)) activated but not released), (iv) altered functionality of the relevant amino acid residues (changes in residues in the binding site or residues directly interacting with an active site impact the binding of drug with its target), (v) altered conserved protein sequences (whether bacteria mutate conserved regions for survival), and (vi) hotspot sites within the drug target (some mutations are more prevalent at specific positions or regions (Zhang & Yew, 2009). Therefore, when trying to map the pathogenicity of drug resistant bacteria, a promising line of attack is to assess the impact of the mutation on *Mycobacterium tuberculosis,* i.e., whether the mutation is neutral or harmful, and how effectively it weakens the efficacy of the drug binding to the target. These two perspectives will help unravel drug resistance mechanisms employed by bacteria, that will shed light on potential avenues for improving the efficacy of existing drugs, developing new drugs and developing improved diagnostic methods and treatment strategies.

(2) Developing an epitope-based *in-silico* human TB vaccine for greatly improved efficacy

Vaccines are designed using biological agents (attenuated, killed or toxoids) resembling the disease-causing pathogen. Vaccines work by mimicking disease-causing agents to stimulate the host immune system to provide defences against them when infected or re-infected (Figure 1.1). Cell-mediated and humoral immune responses are the two main forms of immune response. T-cells and B-cells,

respectively, are the principal constituents of these two forms of response of the immune system. The T-cells play a significant role in cell-mediated immunity (CMI). CMI is the predominant form of the immune response. In CMI, the pathogen is ingested by the antigen-presenting cells (APCs), such as macrophages, dendritic cells, etc. The pathogen is then fragmented into smaller antigenic peptides and later presented to T-cell receptors (TCR) present on the T-cell surface, through the major histocompatibility complex (MHC) molecule attached to antigen presenting cell surface, to produce memory T-cells. The B-cells are essential in providing a humoral or antibody-mediated response.

The antigens recognised by the antibody paratope (antigen-recognition site in antibodies) or by the B-cell receptor (BCR) produce memory B-cells. Antigenic determinants, specific segments of an antigen, are known as epitopes. Epitopes comprise a short stretch of amino acids that are recognised by B-cell and T-cell receptors and specific antibodies. Chapter2 provides details of the immune system and immune response generated by humans against *Mycobacterium tuberculosis*.

Figure 1. 1: Immune response generated in the host body after delivery of a vaccine (Image source: https://www.historyofvaccines.org/content/how-vaccines-work)

Vaccination is administering a vaccine into a host body to elicit an artificial active immune response against infectious diseases. An antigen is a molecule that stimulates an immune response inside the host body and that is recognised by APCs, T-cells or antibodies. After vaccination, the APCs and B-cells recognise, ingest and breakdown the foreign substance into smaller antigens (Figure 1.1). The antigens fragmented inside APCs are then presented to T-helper cells, leading to activation of the immune response. Vaccination helps in the production of memory cells to remember a specific disease agent. The immune response thus produced is called the primary response to a pathogen. Figure 1.1 shows the activation of B-cells and T-cells, the response of B-cells by producing antibodies that later bind to specific antigens and the T-cell response by fragmenting the antigens and production of B-cell and T-

cell memory cells. When the pathogen attacks the host body again, the memory cells produced by the vaccination process will recognize the specific antigens and alert the immune cells to generate a specific immune response to kill the pathogen. This response to a pathogen is called a secondary response. The secondary immune response is greater in magnitude and faster than the primary response. The main goals of developing a new TB vaccine are to: (i) enhance the speed and strength of the host immune response against *Mycobacterium tuberculosis*, (ii) provide long-lasting protection, (iii) enhance herd immunity, (iv) reduce mortality, (v) cost-effective vaccine, and (vi) prevent antimicrobial resistance (Kennedy & Read, 2017; Lipsitch & Siber, 2016; Rodrigues & Plotkin, 2020).

(i) Why is the current TB vaccine not effective?

Although BCG has shown its effectiveness by reducing the incidence of tuberculosis in children, miliary TB and tubercular meningitis, there are main concerns for adult TB (Brewer, 2000). BCG trials have shown that it does not induce the same immune response as when a vaccine is given to premature babies. The safety of BCG remains a significant concern in immunocompetent persons (Hesseling et al., 2007). The protection by BCG is highly variable in adults and the prevention of chronic infection is not high. The reasons for such high variability of BCG vaccine include differences in clinical assays, genetic variability in a sample population, different levels of protection against the clinical forms of tuberculosis, malnutrition and variability in *Mycobacterium tuberculosis* strains (Barreto et al., 2006; Doherty & Andersen, 2005; Skeiky & Sadoff, 2006).

(ii) What are the challenges in developing a new TB vaccine?

The first challenge that needs to be addressed is pathogen polymorphism, i.e., the vaccine's effectiveness against the many strains of *Mycobacterium tuberculosis* (Gulukota, 2008). A vaccine prepared using a single strain or single antigen of a pathogen does not provide effective immunity to the host. The genetic variability among different strains of an immunoevasive pathogen increases the chances of drug resistance. The second critical challenge is triggering a B-cell and T-cell immune responses in the host. The third challenge faced in developing a vaccine is autoimmunity or hypersensitive reactions. Sometimes there is a homology that exists between the *Mycobacterium tuberculosis* genome and the host genome. Suppose this homologous part of the genome is used in a vaccine? In that case, the host body may trigger a suppressive immune response due to the host's immune system being tolerant or by triggering autoimmunity against itself (de Groot & Martin, 2009). Another challenge in developing an effective TB vaccine is the broad coverage of the population needed.

In most cases, a single strain or single antigen vaccine helps treat a small subset of patients in a specific region. The vaccine is then sometimes considered ethnicity-biased in terms of protection. The other obstacles in vaccine design are safety (Movahedi & Hampson, 2008), expensive experimental testing, and time (Naveen et al., 2014). The culturing of *Mycobacterium tuberculosis* to identify vaccine targets in the laboratory is costly and time-consuming. The purification and detoxification of vaccine products are expensive. Leakage of *Mycobacterium tuberculosis* is always a risk in the laboratory.

(iii) What are the gaps in vaccine research?

Many studies have been performed to address the issues of TB. Some studies provided satisfactory results, but they did not significantly contribute to the End-TB strategy. The H4:IC31 recombinant subunit vaccine, in phase II of the clinical trial of 2020, contained the fusion protein of Ag85A and TB10.4 with an IC31 adjuvant (Ahsan, 2015). *In-vitro* studies were also performed to evaluate the T-cell immune response of the H4:IC31 vaccine using data from clinical trials. These studies suggested the need for a new vaccine with a more specific response (Luabeya et al., 2015; Penn-Nicholson et al., 2018; Rodo et al., 2019). The subunit vaccine developed using the conventional approach contained one or more antigenic proteins (an antigen is a molecule that stimulates an immune response inside the host body) for initiating the immune response. The traditional method requires cultivating pathogens in the laboratory and then performing different biochemical, microbiological, immunological tests to detect the vaccine candidates. This approach is laborious, requires expensive experimental testing and sometimes fails to reveal suitable antigens. The research by Monterrubio-López et al. (2015) did not provide a solution to pathogen polymorphism as they applied a reverse vaccinology approach using only a single strain of *Mycobacterium tuberculosis*. In 2017, Hossain et al. (2017) identified T-cell epitopes present on the extracellular protein, 85B, of *Mycobacterium* spp. They used one protein for recognising promiscuous T-cell epitope and did not focus on the antigen variability issue.

(iv) How can we develop a new TB vaccine with a strong and specific immune response?

An effective solution for pathogen polymorphism is using highly conserved vaccine targets from genome sequences of different *Mycobacterium tuberculosis* strains for developing a TB vaccine. Targeting highly conserved regions for their significant structural and functional roles in the *Mycobacterium tuberculosis* life cycle would provide broad-spectrum protection against *Mycobacterium tuberculosis* strains and also protection against drug resistance. A successful vaccine development approach must address the following challenges of conventional vaccine development:
- reducing the cost, time and arduous experimental testing
- overcoming safety concerns while culturing the entire pathogen in the laboratory
- narrowing the research to the selected vaccine targets

- identifying the surface-exposed (Rappuoli, 2000), secreted, adhesin proteins
- conserved epitopes in highly variable (Gershoni et al., 2007) or drug-resistant mycobacteria
- identifying immunodominant epitopes for inducing a potent humoral and cell-mediated immune response
- identifying of highly immunogenic and non-toxic targets
- eliminating cross-reactive epitopes

(3) Identify drug targets and develop an epitope-based vaccine for the treatment of bovine TB

Bovine TB is a chronic infectious disease caused by *Mycobacterium bovis*. Cattle are considered the main reservoir of bovine TB compared to other domestic livestock. Isolating the infected animals and slaughtering them are undertaken to reduce transmission among other animals in the herd. The isolation of infected animals is not an option in hugely populated or low-income countries. Srinivasan et al., 2018 estimated that 21.8 million cattle are infected with bovine TB in India (Srinivasan et al., 2018). The pasteurization of milk is not compulsory in India. Thus, bovine TB in cattle also impacts human health. It is also important to remember that getting infected from eating the meat of the infected animal is less likely, but the risk is still there. While accomplishing this objective the following research questions need answering:

(i) What are prominent issues in bovine tuberculosis treatment?

Currently, there is no effective treatment available for bovine TB. *Mycobacterium bovis* is considered naturally resistant to pyrazinamidase (first-line TB drug) (Nakajima et al., 2010). First-line human TB drugs for treating livestock are also ineffective and costly as the treatment requires six to nine months of daily doses of medication. The use of BCG vaccine has not provided a sufficient level of protection. Several attempts were made to develop a live-attenuated or heat-killed vaccine against bovine TB in cattle, but none were successful (Buddle, 2010; Buddle et al., 2018; Palmer & Thacker, 2018; Parlane & Buddle, 2015).

(ii) What approach can be used against *Mycobacterium bovis* for treating bovine TB?

The development of a bovine TB epitope-based vaccine containing B-cells and T-cells (MHC-I and MHC-II restricted) epitopes could elicit a humoral and cell-mediated immune response. Thus, protective immunity is required for lowering the transmission of infection. Generating memory cells against bovine TB is an essential step towards tackling the disease worldwide. A new approach for the treatment of bovine TB can be developed by using or modifying the available veterinary drugs to target *Mycobacterium bovis* drug candidates to answer pathogenicity and drug resistance against bovine TB.

1.3   Research objectives and approaches

Over the recent years, the development of vaccines and drugs has gained the attention of researchers to develop effective therapeutic tools against tuberculosis. The current explosion in bioinformatics has revolutionized the field of vaccine and drug development. Bioinformatics provides new tools to identify potential vaccine and drug targets without culturing the pathogens in the laboratory. These tools help identify potential therapeutic candidates by directly analysing the proteome of a bacterial strain. First, however, it is essential to understand why the existing drugs and vaccines are ineffective before developing new therapeutic tools for TB. Chapter- 2 attempts to understand in depth the reasons behind the ineffectiveness of current TB treatments.

The primary purpose of this research is to understand the interaction of hosts and TB bacteria and to acquire in-depth knowledge of the survival strategies of TB bacteria against host immune responses and TB drugs and developing an effective therapeutic technique for fighting TB.

There are three overall objectives for this research that are further divided into sub-objectives. The main objectives of the study are to:

(1)   **Undertake a comprehensive analysis of TB drug resistant mutations to understand the survival strategy of *Mycobacterium tuberculosis***

Eradicating TB globally demands a fundamental understanding of the emergence and evolution MDR-TB as the foundation for a strategic approach to combat the threat of MDR to human life for good. Currently, such a foundation is lacking and is urgently needed. Our study develops this foundation through a comprehensive and systematic in-depth analysis of drug resistance mechanisms from global mutation data for *Mycobacterium tuberculosis* reported over the last 30 years. For this extensive investigation of drug resistance, an in-depth look into global mutation patterns to explore the global evolution of drug resistance is performed using the following sub-objectives:

(1.1)   Creating a catalogue of non-synonymous first-line TB drug mutations.

(1.2)   Calculating the single mutation frequencies for each mutation and identifying hotspot residue sites and areas in the target proteins.

(1.3)   Predicting the impact of drug-resistant mutations at the molecular level affecting the function, structural stability, and sequence conservation in the first-line Tb drug targets to understand the survival of *Mycobacterium tuberculosis.*

(1.4)   Categorizing drug resistant mutations into lethal, moderate, mild and neutral in relation to impact on bacterial survival.
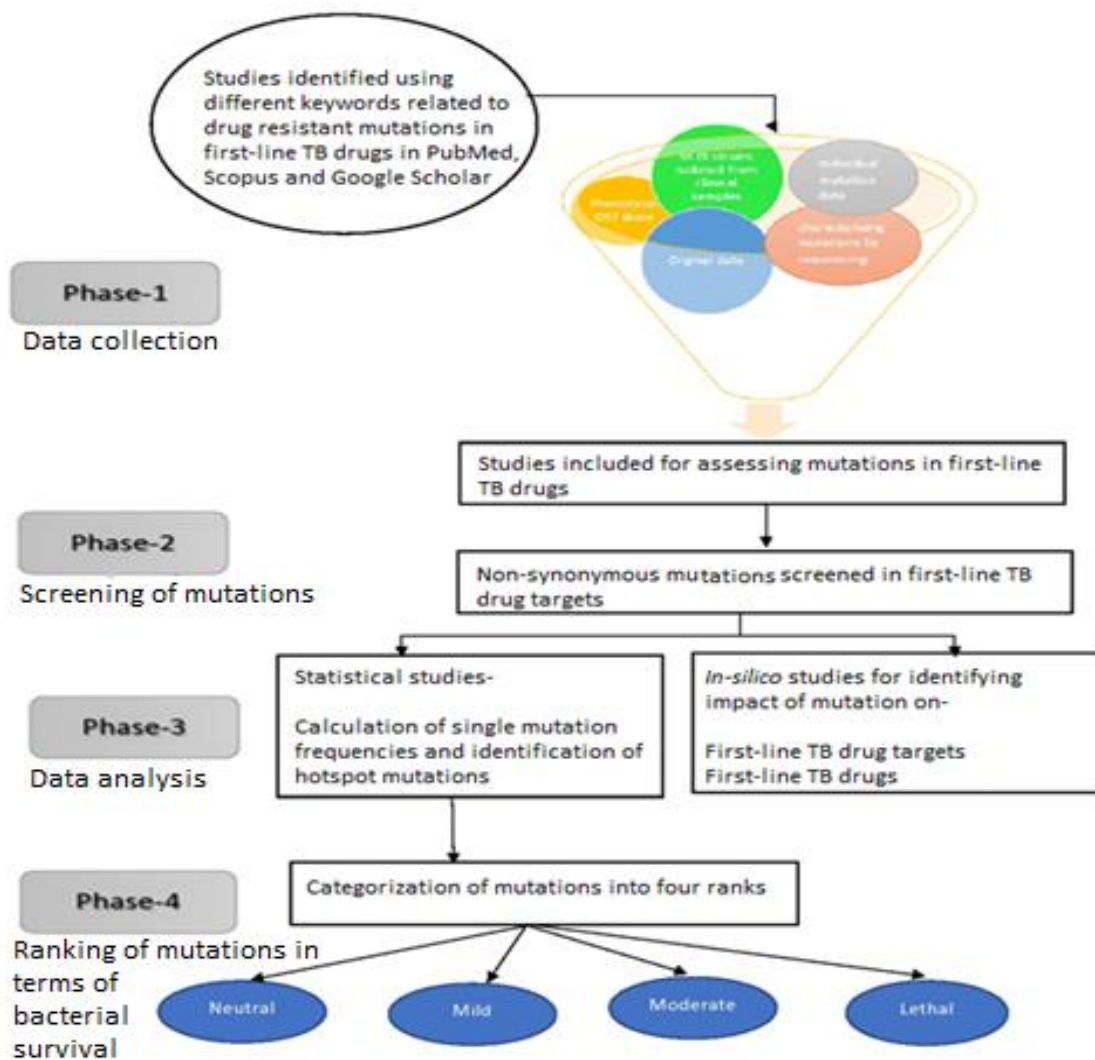
Figure 1. 2: Framework designed for understanding the impact of mutations on *Mycobacterium tuberculosis* survival

The essential things needed to address drug resistance are collecting first-line mutation data from different sources and using that data to improve our knowledge of the underlying drug resistance mechanisms to achieve the long-term goal of eradicating TB. Figure 1.2 shows that the proposed framework has four phases: collecting mutational data, screening of non-synonymous mutations, analysing mutational data in two stages: (i) mutational statistics, and (ii) the impact of mutations on the targets and drugs, and categorizing mutations into lethal, moderate, mild and neutral in relation to impact on bacterial survival to evaluate the total drug resistance strategy of *Mycobacterium tuberculosis*. The single mutation frequency was calculated to understand the prevalence and diversity

of mutations in first-line TB drug targets across the globe. Our research focused on providing a comprehensive coverage of mutations and identifying their impact on TB bacteria and drug binding, using detailed bioinformatics analysis to understand crucial changes at the molecular level of the target affecting its function, structural stability and sequence conservation. The influence of mutation position on drug binding affinity is determined by comparing the binding energies of mutant TB drug target proteins with the wild-type target protein. The coverage of a large number of strains (mutations) is novel in this study as well as the study of mutational impact on targets and drugs. Further, this research introduced a new concept of ranking drug-resistant TB mutations into lethal, moderate, mild and neutral for bacterial survival. When designing new drugs, this method can help predict the impact of the mutation on the respective drug targets to develop better drugs for TB treatment. Furthermore, the ranking of mutations into four different categories can assist in developing inhibitors for a specific mutation or group of mutations and help develop personalised treatment plans for TB patients.

(2) Designing a conceptual and computational framework for developing an effective epitope-based human TB vaccine

An epitope-based vaccine can stimulate a specific and swift adaptive and humoral immune response. Thus, a conceptual framework is developed that uses different bioinformatics approaches, such as comparative proteome analysis, reverse vaccinology, immunoinformatics, and structural vaccinology to identify potential vaccine candidates and construct an *in-silico* vaccine. Advancements in genomics have introduced various sequencing techniques, like shotgun sequencing, which provides information about the whole genome sequence of an organism. Information about the genome, transcriptome or proteome of a pathogen (Brusic & Flower, 2004) can help identify novel vaccine candidates that are important for developing effective vaccines. Computational vaccinology uses computer-aided vaccine design to identify novel vaccine candidates within the genome of the target pathogen before experimental testing and validation (He et al., 2010). Computational vaccinology allows researchers to identify vaccine targets that might be missed in the conventional approach because a pathogen cannot be successfully and safely cultured in the laboratory (Flower et al., 2010). Currently, computational vaccinology aims to identify suitable vaccine targets using reverse vaccinology (RV) and immuno-informatics approaches. Suitable vaccine candidates can be identified using a computational pipeline that allows analysis of the genome or proteome of a pathogen. This approach is known as 'reverse vaccinology' RV provides the repertoire of antigenic proteins present in a pathogen, that are not able to be understood in the conventional approach (Rappuoli, 2000). The antigenic proteins identified by the RV approach can be analysed further to predict their B-cell and T-cell epitopes. The field of immunoinformatics provides many tools for epitope mapping (Vivona et al., 2008).

Crucial features of an epitope-based vaccine include sequence conservancy, antigenicity, exclusion of self-peptides and multiple allelic interactions. For achieving this objective, several sub-objectives are required to be accomplished:

(2.1)   Undertake comparative proteomic analysis of the 159 strains of *Mycobacterium tuberculosis* for identification of conserved proteins.

(2.2)   Use functional classification of the conserved proteins identified.

(2.3)   Identify surface-exposed antigenic proteins that are virulent and do not cause autoimmunity in the host using the reverse vaccinology pipeline.

(2.4)   Perform immunoinformatics analysis for identifying T-cell and B-cell epitopes from the antigens and filter potential epitopes.

(2.5)   Construct an *in-silico* vaccine and evaluate immune response.
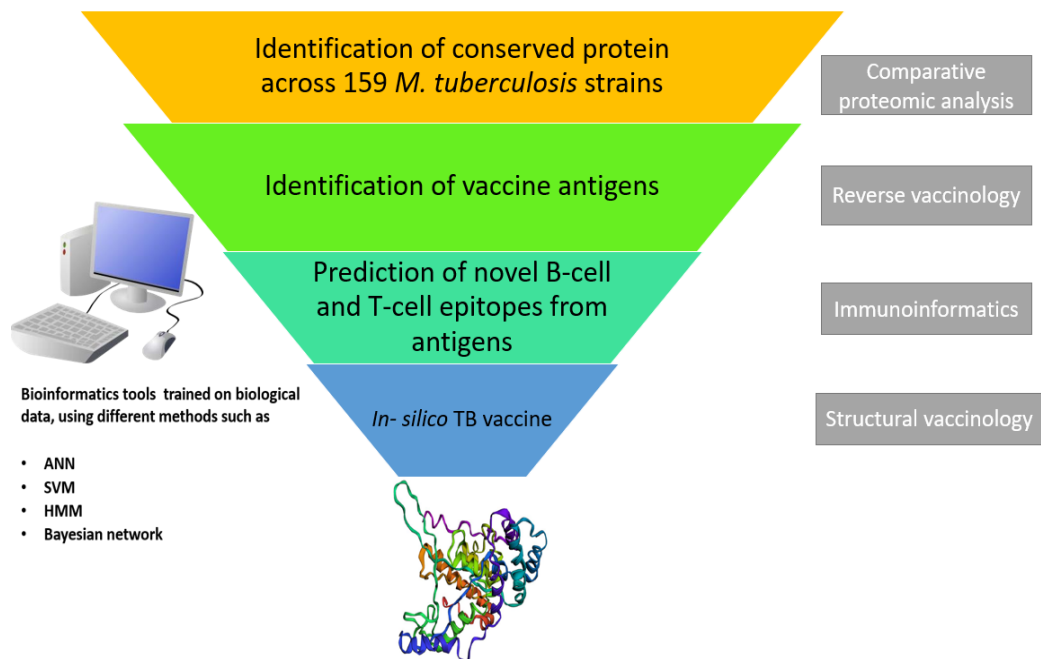
Figure 1. 3: Conceptual framework for developing *in-silico* vaccine against *Mycobacterium tuberculosis*

Figure 1.3 shows the conceptual framework created for designing the TB vaccine. The first sub-objective, comparative proteomic analysis, must identify conserved proteins across 159 *Mycobacterium tuberculosis* strains to address pathogen polymorphism. The resultant conserved proteins need to be categorised into eleven distinct functional categories, as given by Smith (2003). After a detailed study of the etiology of tuberculosis, reverse vaccinology methods are then used to perform several essential steps that were not carried out in previous research undertaken on TB. These

steps identifies the outer-membrane antigenic proteins with unique characteristics, such as, like signal peptides, membrane-spanning regions, lipoprotein signatures, adhesion probabilities, and motif attachment from the genome or proteome of the pathogen. After completing this, immunoinformatics analysis for the prediction of T-cells (MHC-I and MHC-II restricted epitopes) and B-cell epitopes is performed. Filtering of the epitopes is undertaken to identify most antigenic, non-allergic, non-toxic and excluding self-peptides. The docking analysis of the filtered epitopes with every other epitope creates unique combinations of epitopes. Combinations having a strong binding affinity to one another are used for constructing vaccines. Structural vaccinology is then used to help develop an *in-silico* vaccine and elucidate its tertiary structure. This extensive study analysing strains reported in the last 30 years will identify the broadest coverage of effective epitopes to date. The construction of the three-dimensional structure of the vaccine is based on a new concept introduced in this research.

The benefit of an epitope-based vaccine is removing deleterious epitopes that can cause cross-reactive reactions or autoimmunity in the host. Evaluation of the immune response is essential for predicting vaccine efficacy in generating a strong and specific humoral and cell-mediated immune response. This research is intended to address the challenges of TB vaccine development that include: expensive, time-consuming and arduous experimental testing; safety concerns while culturing the pathogen in a laboratory; identification of surface exposed, secreted and adhesin proteins; conserved epitopes in highly variable or drug-resistant *Mycobacterium tuberculosis*; identifying immunodominant epitopes for inducing potent humoral and cell-mediated immune responses; elimination of cross-reactive epitopes; and immunogenicity assessment of selected epitopes. The development of an *in-silico* epitope-based subunit vaccine using computational vaccinology is expected to be highly effective and safer against tuberculosis.

(3) Developing effective vaccines and identifying drug targets against bovine TB

The third objective of the thesis is to study the comparative proteomic analysis of *Mycobacterium bovis* to discover an *in-silico* vaccine and drug targets for treating bovine tuberculosis. The objective was achieved using the following sub-objectives:

(3.1) Undertaking comparative proteomic analysis of 11 strains of *Mycobacterium bovis* for the identification of conserved proteins.

(3.2) Designing an *in-silico* vaccine against bovine TB using the same framework developed in objective 2 for human TB.

(3.3) Identifying potential drug targets against *Mycobacterium bovis*.

Currently, no effective treatment is available for treating bovine tuberculosis and animal slaughtering is usually undertaken to reduce the burden of bovine tuberculosis in the environment. In this research, we propose therapeutic drug targets and vaccines for treating bovine TB. The emergence of antimicrobial resistance in bacteria has reduced the efficacy of antibiotics in treating the disease. To overcome the antimicrobial resistance issue, we used a novel approach to identify conserved and pathogenic drug targets to design better drug therapeutics against bovine TB. For designing an *in-silico* vaccine for bovine tuberculosis, the framework created in objective 2 was used. A novel subtractive genomic approach was developed for identifying bovine TB drug targets. This approach was used to determine the drugs targets that are conserved, essential, antigenic and have unique metabolic pathways in *Mycobacterium bovis*. The detailed method for drug target identification is explained in chapter 5. The strategy developed for identifying drug targets is generic and can be used for other zoonotic infectious diseases.

1.4 Organization of the thesis

This research provides an in-depth exploration of the TB bacteria's distinctive set of strategies for survival and the different defence mechanisms it uses against drugs and vaccines. It provides a potential solution to each research question. This research makes novel contributions to the field of vaccine development for tuberculosis. It also makes an effort to understand the global pattern of drug mutations for improving current treatment plans and developing new diagnostic techniques. Conducting laboratory or *in-vitro* studies on *Mycobacterium tuberculosis* and *Mycobacterium bovis* would be time-consuming and expensive. In this research, various bioinformatics tools and software have been used to computationally analyse the TB bacterial genome to provide a potential solution to reduce the burden from TB disease through vaccine and drug development. Using a number of prediction tools together to achieve checks and balances, our research framework aimed to reduce the chance of errors and provide accurate results.

The thesis is divided into six chapters. Figure 1.4 explains the overall thesis organization.

Chapter 1- Introduction

Chapter 2- Why are the existing TB vaccine and drugs not effective? A potential solution to the problem

Chapter 3- A comprehensive analysis of first-line tuberculosis drug resistant mutations to understand the survival strategy of *Mycobacterium tuberculosis*

Chapter 4- Developing and testing a conceptual and computational framework towards an effective human TB vaccine

Chapter-5 Identification of therapeutic vaccine and drug candidates against *Mycobacterium bovis*

Chapter 6- Summary and conclusion

Figure 1. 4: Overall thesis organization in six chapters

Chapter 1 provides a brief introduction to infectious diseases and TB and the emergence of antimicrobial resistance. This chapter highlights the importance of vaccines in reducing the burden of infectious diseases and TB. The research problem, three research objectives and approaches used in accomplishing each objective are summarised in this chapter. **Chapter 2** provides a detailed review of why the existing drugs and vaccines against tuberculosis are not adequate. It explains the pathogenesis of TB, the human immune response and evasion strategies of *Mycobacterium tuberculosis* from an immune response perspective. The challenges in developing new drugs and vaccines for TB treatment and potential solutions to address the challenges are discussed. **Chapter 3** corresponds to the first objective, which deals with understanding the impact of mutations on the first-line TB drug targets. This chapter provides an in-depth understanding of the impact of mutations at the evolutionary, functional and structural levels. The method developed in this chapter can also help when studying future mutations and there is scope to introduce new steps in the method to improve it. **Chapter 4** focuses on the second objective, which is developing an *in-silico* vaccine against human tuberculosis. The chapter describes the method proposed for developing a potentially safe, conserved and immunogenic epitope-based *in-silico* vaccine using reverse vaccinology, immunoinformatics and structural vaccinology. The evaluation of the humoral and cell-mediated immune responses generated by the TB vaccine is also described in this chapter. The vaccine identified is expected to evoke a specific immune response to provide broad immune protection against many *Mycobacterium tuberculosis* strains. **Chapter 5** corresponds to the third objective, which deals with developing a method for

proposing a potential therapeutic vaccine and drug targets for bovine tuberculosis. The vaccine is produced using the same method as objective 2. A conceptual method to identify drug targets against *Mycobacterium bovis* is described in the chapter. Finally, a summary of the entire thesis is given in **chapter 6**, which provides an overview of the research with important highlights, conclusions, and suggestions for future.

Chapter 2

Why are the existing TB vaccine and drugs not effective? A potential solution to the problem

The available anti-TB therapies have led to a 50% reduction in mortality rate since 1990. Still, millions of people are getting infected and dying from the TB disease. The low efficacy of BCG and the emergence of drug resistance are the main challenges to eliminating TB globally. This research is designed to provide potential solutions by developing efficient therapeutic tools (vaccine and drugs) for treating TB through a deep understanding of the disease. Thus, it is crucial to understand host-pathogen interactions and the underlying mechanisms *Mycobacterium tuberculosis* uses to evade the vaccine and drug treatment. Section 2.1 discusses the history of TB, the epidemiology, cell wall structure and genome of *Mycobacterium tuberculosis* and the pathogenesis of TB. This section explains TB disease progression, the multitude of immune responses generated by the host's immune system and mechanisms used by *Mycobacterium tuberculosis* to evade those responses. Section 2.2 reviews the current treatments (vaccine and drugs) available and highlights the crucial factors that make the treatments ineffective. Section 2.3 describes the challenges for developing new drugs and vaccines. Section 2.4, in the latter part of the chapter, provides knowledge about what efforts can be made in TB vaccine and drug development to provide potential therapeutic tools.

2.1 Tuberculosis- a persisting infectious disease

TB is an evolving deadly disease caused by one of the world's most infectious bacteria, *Mycobacterium tuberculosis.* According to the World Health Organization (WHO), tuberculosis is a global threat with significant mortality and morbidity rates (Pieters, 2008; Wlodarska et al., 2015) (WHO Global tuberculosis Report, 2020). Despite the advances in medical sciences, TB remained the cause of death for 1.4 million people in 2019 and was responsible for 10.0 million new cases worldwide (WHO Global Tuberculosis Report, 2020). According to the world's oldest literature, TB has plagued some of humanity's earliest civilizations (Zimmerman, 1979). Today, vaccine and drug therapy are the two most important human countermeasures against TB. However, *Mycobacterium tuberculosis* has developed several ways to destabilize the human immune response by adapting to the host's changing environment and spreading the infection to other parts of the body. This high level of adaptation has led to the evolution of *Mycobacterium tuberculosis* and the resulting rise in resistant strains. The widespread emergence of TB resistant strains has challenged the view of tuberculosis as a treatable disease.

2.1.1   History of tuberculosis

Mycobacterium tuberculosis has plagued some of the earliest civilizations. DNA findings in ancient Egyptian remains have provided evidence that tuberculosis is well over 3000-5000 years old (Zimmerman, 1979). Since ancient times, tuberculosis has been an epidemic and had been called by many names, such as, white plague, phthisis, scrofula, Pott's disease. Hippocrates was the first to describe phthisis as a pulmonary infection manifesting as weight loss, cough, and blood in the sputum, which gradually led to death. In the 16th century, Italian physician, Girolamo Fracastoro, proposed a theory that the disease was transmitted by microorganisms that are invisible to the naked eye. Benjamin Marten, in 1720, published information about the infectious origin of tuberculosis in his publication "A new Theory of Consumption." He stated that TB could be caused by tiny living creatures that could lead to the symptoms and lesions of tuberculosis (Hardy, 1999). In 1793, Mathew Baille, a Scottish pathologist, named phthisic abscesses "tubercules" (Houston, 1999). In 1854, Hermann Brehmer, a botany student suffering from tuberculosis, submitted his doctoral dissertation "Tuberculosis is a curable disease," describing the first successful remedy to cure tuberculosis. As suggested by his doctors, he travelled to a healthier climate, the Himalayan mountains, to get rid of TB (Daniel, 2011). In 1895, Jean-Antoine Villemin, a French military surgeon, demonstrated the infectious nature of tuberculosis by inoculating a rabbit with purulent liquid obtained from a tuberculous cavity of a patient who died from TB (Daniel, 2006). The history of TB dramatically changed when Robert Koch, a German physician, discovered *Mycobacterium tuberculosis,* in 1882. He named it Koch's bacillus. Robert Koch presented his findings on the infectious cause of tuberculosis on the evening of 24 March 1882, a day commemorated as World TB Day. In his presentation, he not only demonstrated the identification of tubercule bacteria but also provided his famous "Koch postulates." Koch postulates are as follows: (i) the microbes should be present in abundance in organisms suffering from the disease and should not be found in healthy organisms, (ii) the microbes can be taken from an infected host and grown independently in pure culture, (iii) the cultured microbe should cause disease when introduced into a healthy host and (iv) the microbe isolated and identified from the host (healthy host who was infected by the microbe in (iii)). Koch was awarded the Noble Prize in Physiology and Medicine in 1905 for identifying the TB bacterium.

2.1.2   Epidemiology of tuberculosis

As stated earlier, in 2019, about 10.0 million new TB cases were reported worldwide (Figure 2.1), and approximately 5.8 million (56%) were men, 3.2 million (32%) women and 1.0 million (12%) were children (WHO Global Tuberculosis Report, 2020). HIV-infected patients accounted for 2.7 million (11%) of all new TB cases. The prevalence of TB is the highest in South-East Asia (44%), Africa (25%) and the Western Pacific (18%). India, Indonesia, China, Nigeria, Philippines, Pakistan, Bangladesh and South Africa are the top eight countries, accounting for two-thirds of the total global cases. Therefore,

global progress on eradicating TB depends on significant advances in TB prevention and care in these countries (WHO Global tuberculosis Report, 2020). Further, as stated earlier, there were 1.2 million deaths from TB in 2019 and an additional 0.2 million deaths from tuberculosis in HIV-positive people. Resembling prevalence of TB, low-income and low-middle income countries in the world find TB to be one of the leading causes of death.
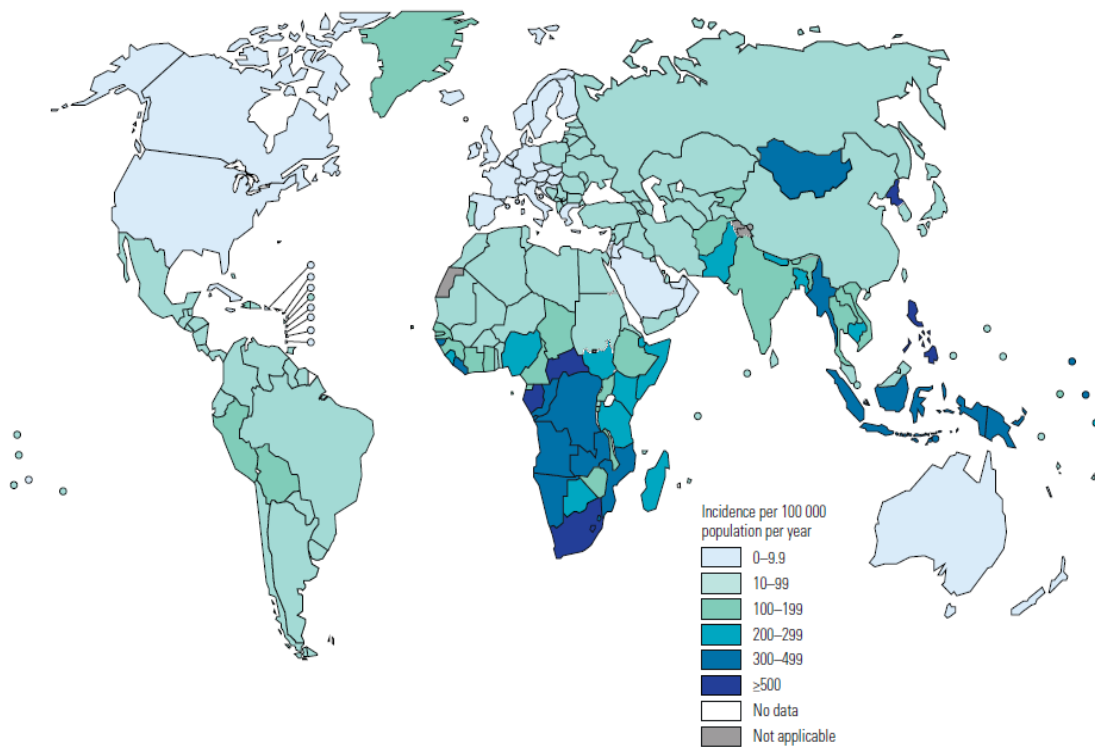
Figure 2. 1: Estimated TB incidence rates in 2019 (Sourced from WHO Global TB Report, 2020)

The burden of tuberculosis can be measured in terms of:

1. Incidence (number of new and relapse cases at a specific point in time)
2. Prevalence (number of cases at a given point in time)
3. Mortality (number of deaths at a given point in time)
4. Fatality rate (number of persons dying from tuberculosis among all persons with the disease)
5. Attack rate (number of cases developing tuberculosis among all persons who were exposed to the disease)

The emergence and spread of multidrug-resistant tuberculosis (MDR-TB) and extensively drug-resistant tuberculosis (XDR-TB) is of great concern when treating infected patients. MDR-TB is defined as tuberculosis that shows resistance to at least two crucial first-line anti-TB drugs, isoniazid and rifampicin, with or without resistance shown to other first-line drugs (WHO Global tuberculosis Report,

2020). MDR-TB is a serious global concern and outbreaks have been reported in many countries. In 2019, 0.5 million new MDR-TB cases were reported. The incidence rate has reduced compared to the incidence in 2019; however, 0.2 million still died from MDR-TB in 2019 (WHO Global tuberculosis Report, 2020). The majority of new cases of MDR-TB was seen in South-East Asia, Western Pacific and Africa (Figure 2.2). MDR-TB requires prolonged treatment using expensive and highly toxic, second-line anti-TB drugs. Extensively drug-resistant TB is a form of TB where *Mycobacterium tuberculosis* shows resistance to isoniazid and rifampicin (MDR-TB) as well as fluoroquinolone and all of the second-line anti-TB injectable drugs (amikacin, kanamycin or capreomycin) (WHO Global tuberculosis Report, 2020). A total of 12,350 new XDR-TB cases were notified in 2019.
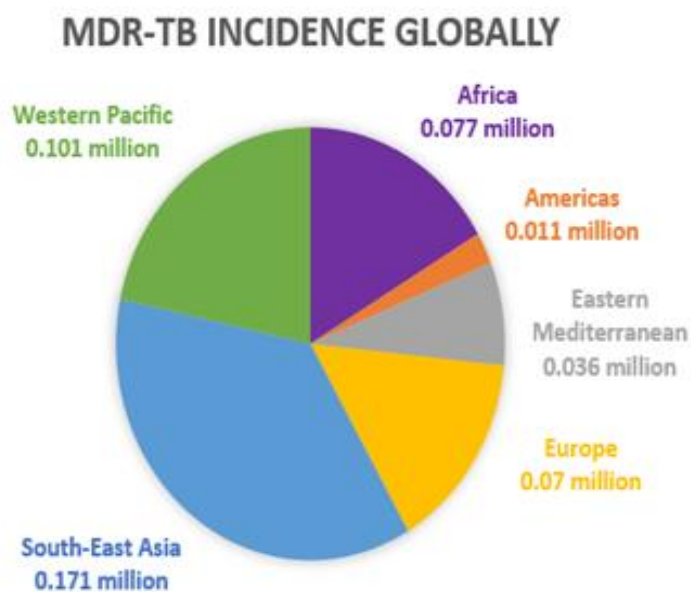
Figure 2. 2: Incidence of new cases of MDR-TB reported in 2019 by the World Health Organization (WHO). The majority of new cases of MDR-TB occurred in South-East Asia, Western Pacific and Africa

2.1.3 *Mycobacterium tuberculosis*- the causative agent of tuberculosis

2.1.3.1 Taxonomy

Kingdom:	Bacteria
Phylum:	Actinobacteria
Order:	Actinomycetales
Suborder:	Corneybacterineae
Family:	Mycobacteriaceae
Genus:	Mycobacterium
Species:	*tuberculosis*

2.1.3.2 Morphology

Mycobacterium tuberculosis is a rod-shaped, acid-fast bacillus measuring around 0.2-0.6 µm in width and 1.5-4 µm in length. It is a weakly gram-positive, slow-growing and an obligate aerobe. Under an electron microscope, tubercle bacilli appear as straight or slightly curved rods (Figure 2.3). According to growth conditions and the age of the culture, bacilli may vary in size and shape. Some of the bacilli may have coccobacillus, filamentous and branched growth in the form of long rods. Irregular staining and mechanical damage to the cell wall during smear preparation may cause a beaded appearance in some stained bacilli. Cell division by *Mycobacterium tuberculosis* usually takes 20-24 hours and it can be cultured in three to five weeks using Lowenstein-Jensen medium.

Image removed for Copyright compliance

Figure 2. 3: *Mycobacterium tuberculosis* **scanning electron micrograph (Image source: http://textbookofbacteriology.net/tuberculosis.html)**

2.1.3.3 Cell wall structure

The cell wall of *Mycobacterium tuberculosis* has a complex structure that provides specific characteristics such as virulence and resistance to antibiotics. It mainly contains mycolic acid (MA), which makes up to 50% of the dry weight of the cell (Kleinnijenhuis et al., 2011). In principle, the cell wall consists of an inner layer and an outer layer (Figure 2.4). The inner layer consists of peptidoglycan (PG), phosphatidyl-myo-inositol mannoside (PIM) and arabinogalactan (AG) (Torrelles & Schlesinger, 2010). Peptidoglycan is composed of peptides and glycan. The glycan strand consists of repeating units of N-acetylglucosamines (NAG) attached to N-acetylmuramic acid (NAM) (Trias et al., 1992). The peptidoglycan helps in preventing osmotic lysis in the *Mycobacterium tuberculosis* cell. The highly cross-linked network of peptidoglycan maintains the shape of the bacteria. Arabinogalactan is the most important polymer present next to peptidoglycan in the cell wall. It is linked to peptidoglycan by a unique diglycosylphosphoryl bridge, containing NAG and rhamnose. Arabinogalactan comprises a

galactose backbone with arabinose branches and provides strength to the *Mycobacterium tuberculosis* cell wall (Trias et al., 1992; Trias & Benz, 1994).

Image removed for Copyright compliance

Figure 2. 4: Structure of cell wall of *Mycobacterium tuberculosis* showing the components of the outer and inner layers of the cell wall and their distributions (Kleinnijenhuis et al., 2011)

The outer compartment consists of mycolic acid, mannose-capped lipoarabinomannan (Man-LAM) and mannoglycoproteins (Torrelles & Schlesinger, 2010). Mycolic acids are the most significant components of a cell wall and are unique α-branched lipids linked to a hexa- arabinose motif by ester linkages at the terminus of the branched arabinogalactan (Trias & Benz, 1994). Mycolic acid molecules are responsible for providing resistance to the lysosomal enzymes of the host cell (Brennan & Nikaido, 1995). A thick layer of mycolic acid impairs the entry of nutrients into the mycobacteria and is the main reason behind the slow growth of *Mycobacterium tuberculosis.* InhA, involved in synthesising mycolic acid, is an important drug target for isoniazid and ethionamide (Ahmad & Mokaddas, 2014). Man-LAM, present on the surface of *Mycobacterium tuberculosis*, is an important virulence factor (Strohmeier & Fenton, 1999; Torrelles & Schlesinger, 2010). It binds to the mannose receptor present on the surface of the alveolar macrophage of the host, leading to its entry into the host cell (Torrelles & Schlesinger, 2010). LAM also helps in resisting the host's cellular oxidative response. Mannoglycoproteins are secreted by mycobacteria that help in the growth of the cell. The porin present in the cell wall facilitates the transport of substances. The cell wall is hydrophobic and has a strong, permeable barrier, making it naturally resistant to host defence mechanisms and antibiotics.

2.1.4 Transmission of tuberculosis

Tuberculosis is a contagious disease and persons having active pulmonary tuberculosis are the sources of transmission for the disease. *Mycobacterium tuberculosis* can be released into the environment by the infected person through coughing, sneezing, speaking loudly or singing, in the form of aerosol droplets. The infection begins when healthy individuals inhale these aerosol droplets. The transmission process of tuberculosis is very efficient as these droplets can persist in the environment for several

hours. The infectious dose is very low, i.e., fewer than ten bacilli are needed to start an infection. The risk of transmission is dependent on numerous factors, such as the number of organisms being expelled into the air, the concentration of microorganisms in the air determined by the volume of the space and its ventilation, the duration of time an exposed person breathes the contaminated air and the immune status of the exposed person (Churchyard et al., 2017). An infected person takes three to four weeks to transmit the disease to another healthy person. Other factors influencing tuberculosis transmission include poor nutrition, HIV infection, close contact within highly populated areas, and intravenous drug use (Jerant et al., 2000).

2.1.5 Genome of *Mycobacterium tuberculosis*

The genome of the strain *Mycobacterium tuberculosis* H37Rv was first published in 1998. The circular genome consists of 4,411,529 base pairs with around 4000 genes and contains a high G+C content of about 65.6% (Cole et al., 1998). The genome also comprises six pseudogenes. Table 1 shows the classification of *Mycobacterium tuberculosis* H37Rv genes and their function.

Table 2. 1: General classification of *Mycobacterium tuberculosis* H37Rv genes (Smith, 2003)

Function	No. of genes	% of total genes	% of total coding capacity of total genes in genome
Lipid metabolism	225	5.7	9.3
Information pathways	207	5.2	6.1
Cell wall and cell processes	517	13.0	15.5
Stable RNAs	50	1.3	0.2
IS elements and bacteriophages	137	3.4	2.5
PE and PPE proteins	167	4.2	7.1
Intermediary metabolism and respiration	877	22.0	24.6
Regulatory proteins	188	4.7	4.0
Virulence, detoxification and adaptation	91	2.3	2.4
Conserved hypothetical function	911	22.9	18.4
Proteins of unknown function	607	15.3	9.9

About 6% of the genome is involved in lipid metabolism and this is considered an essential part of the *Mycobacterium tuberculosis* genome (Cole et al., 1998; Smith, 2003). Out of the 225 genes involved in lipid metabolism, 100 genes, required for β-oxidation of fatty acids, are necessary for the pathogen's survival inside the host body. In addition, fatty acids are thought to be a significant carbon source essential for the growth of *Mycobacterium tuberculosis* inside macrophages (Smith, 2003).

One of the distinctive features of the *Mycobacterium tuberculosis* genome is the presence of 167 genes belonging to acidic, glycine-rich protein families named the PE and PPE protein families. PE (proline-glutamate) and PPE (proline-proline-glutamate) sequences, 110 and 180 amino acids in length, respectively, are present in the conserved N-terminal regions of each of the PE and PPE protein families (Brosch et al., 2000; Smith, 2003). Out of 167 genes, 104 genes belong to the PE family and 68 belong to the PPE family (Brosch et al., 2000). Proteins belonging to these protein families, present in the cell wall and cell membrane (Banu et al., 2002), are the main reason for the antigen variability in *Mycobacterium tuberculosis* (Banu et al., 2002; Smith, 2003). These proteins can play an important role in vaccine design and development.

The 188 regulatory genes present in the genome include 13 sigma factors required for transcription regulation and 13 two-component regulatory genes necessary for signal transduction (Smith, 2003). Out of 517 genes involved in cell wall and cell processes, 125 genes are needed for transportation processes. Most of the genes present in the cell walls are involved in intermediary metabolism, and cell wall and cell processes. Fifteen per cent of the genes in the cell wall are required for lipid metabolism and 5% for virulence, detoxification and adaptation. *KatG*, *rpoB*, *inh-A*, *ahpC*, *gyrA*, *pncA*, *rpsL* and *rrs* are examples of genes involved in TB drug resistance. In addition, 19-kDa is an essential lipoprotein engaged in suppressing the human immune response (Hestvik et al., 2005). The two genes, *sodA* and *sodC,* encode the superoxide dismutase required for protecting *Mycobacterium tuberculosis* against reactive oxygen species (ROS) inside the alveolar macrophages (Piddington et al., 2001).

Table 2.2 shows the size and number of genes of some sequenced genomes. Accordingly, the genome's size varies from 4.33-4.41 million base pairs (MB) and the number of genes varies from 4008-4846.

Table 2. 2: List of some completely sequenced *Mycobacterium tuberculosis* members

Organism	Genome size (MB)	No. of genes
Mycobacterium tuberculosis H37Rv	4.41	4008
Mycobacterium tuberculosis CDC1551	4.40	4282
Mycobacterium tuberculosis H37Ra	4.41	4296
Mycobacterium tuberculosis F11	4.42	4288
Mycobacterium tuberculosis KZN1435	4.39	4277
Mycobacterium tuberculosis str. Haarlem	4.40	4271
Mycobacterium tuberculosis KZN4207	4.39	4271
Mycobacterium tuberculosis KZN605	4.39	4276
Mycobacterium tuberculosis CCDC5180	4.40	4284
Mycobacterium tuberculosis str. Erdman= ATCC 35801	4.39	4284
Mycobacterium tuberculosis str. Beijing/NITR203	4.41	4294
Mycobacterium tuberculosis EAI5/NITR206	4.39	4272
Mycobacterium tuberculosis CCDC5079	4.41	4285

2.1.6   Immune response against tuberculosis

Before understanding how the immune system responds to TB infection, it is essential to have a basic understanding of the immune system.

2.1.6.1 Overview of the immune system

The immune system comprises immune cells that undertake functions to protect the host body in response to an infection. The mammalian immune system includes the innate (natural) and adaptive immune systems (Marshall et al., 2018). The innate immune system provides the first line of defence by eliminating an infection in its initial stages. The innate immune system also plays a vital role in the activation of the adaptive immune system. The innate immune system recognizes conserved microbial structures called pathogen-associated molecular patterns (PAMPs) (Hoffmann, 1999; Medzhitov & Janeway, 2000). The PAMPs interact with the toll-like receptors (TLR) present on the surface of

antigen-presenting cells (APCs), causing activation of the adaptive immune system resulting in elimination of the infection (Hoffmann, 1999; Marshall et al., 2018; Medzhitov & Janeway, 2000).

There are two types of responses produced by the adaptive immune system: a humoral immune response (activation of B-lymphocytes) and a cell-mediated immune response (activation of T-lymphocytes) (Tomar & De, 2014). In humoral immunity, extracellular microbes are eliminated by antibodies secreted by B-lymphocytes. B-lymphocytes differentiate into B-effector cells, secreting antibodies, and memory B-cells. In cell-mediated immunity, helper and cytotoxic T-lymphocytes play a role in eradicating the infection. The activation of T-lymphocytes is dependent on the presentation of antigens/epitopes by MHC molecules. MHC molecules are glycoproteins present on the surface of the APCs and their main task is to present epitopes to the T-cell receptors of T-cells (Figure 2.5). MHC class-I molecules present the antigen to the cytotoxic T-lymphocyte (CTL) and MHC class-II molecules present it to the helper T-lymphocyte (HTL) (McMaster et al., 2015; Tomar & De, 2014). CTL can directly kill the pathogenic antigens or infected cells, whereas HTL plays an indirect role in eliminating the infection. HTL instructs the immune cells and releases the cytokines to activate the APCs, CTL and B-cell to kill the pathogen (Medzhitov & Janeway, 2000).
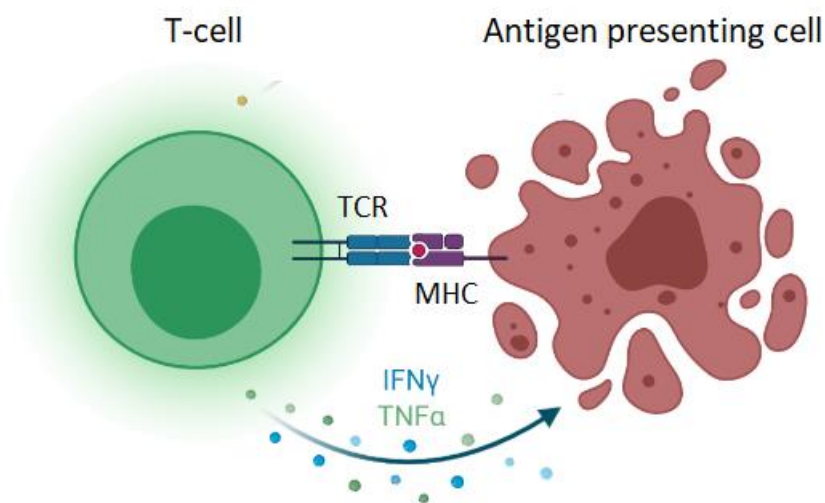
Figure 2. 5: Classical pathway of presentation of antigens/epitopes by MHC present on the surface of APC to TCR of T-cell. Figure created using template from Biorender (https://biorender.com/)

2.1.6.2 Innate immune response to *Mycobacterium tuberculosis*

Mycobacterium tuberculosis is usually transmitted as droplet nuclei by inhalation. Most of the larger droplets are blocked by the epithelium of the upper respiratory tract, where an infection is unlikely to develop (McNerney et al., 2012; Rohde et al., 2007).). The smaller droplet nuclei reach the tiny air sacs

where the actual infection begins. Once in the lung, alveolar macrophages provide the body's first line of defence against *Mycobacterium tuberculosis*. The immune system can clear the infection only if the innate immune system is activated correctly. The toll-like receptors present on the surface macrophage recognizes the foreign proteins (PAMPS) present on *Mycobacterium tuberculosis* and phagocytoses, forming a space called a phagosome (Yuk & Jo, 2014). TLR-2, 4 and 9 are essential immune receptors for TB bacteria (Adami & Cervantes, 2015; Ferluga et al., 2020; Rivas-Santiago et al., 2008). The phagosome fuses with lysosomes within macrophages to form a phagolysosome (Vergne et al., 2003; Yuk & Jo, 2014). After fusion, the hydrolytic molecules of the lysosome inhibit the replication of *Mycobacterium tuberculosis* and lead to its breakdown. The proteolytic activity of the proteasome helps in the fragmentation of whole pathogenic bacterium. However, *Mycobacterium tuberculosis* has developed several ways to subvert the killing mechanisms that allow the replication and proliferation of bacteria. In the first three weeks of infection, most individuals do not show any symptoms of infection (asymptomatic) or may only have a mild flu-like illness.

2.1.6.3 Adaptive immune response to *Mycobacterium tuberculosis*

Approximately three weeks after the initial infection, adaptive immunity attempts to wall off the *Mycobacterium tuberculosis* and prevent it from spreading. Figure 2.6 presents the stages of the adaptive immune response against *Mycobacterium tuberculosis*. First, the alveolar macrophage presents mycobacterium antigens to CD4+ T helper cells through MHC class II and MHC class I to CD8+ cytotoxic T lymphocytes, leading to the activation of the lymphocytes (Dheda et al., 2010). Upon activation, T-helper cells release interferon-gamma (IFN-γ) to perform bactericidal activity and enhance the killing of *Mycobacterium tuberculosis* present within the macrophages by the production of nitric oxide (NO) and reactive oxygen species (Adami & Cervantes, 2015; Rohde et al., 2007). The infected macrophages then release pro-inflammatory cytokines, such as interleukin-12 (IL-12), IL-23 and tumour necrosis factor-alpha (TNF-α), which lead to the recruitment of mononuclear cells from nearby blood vessels (Figure 2.6) (Dheda et al., 2010; Schaible & Kaufmann, 2000). Finally, the release of IL-4, IL-5 IL-10 and IL-13 by T-helper cells promote the activation of B-lymphocytes leading to antibody production. The role of B-lymphocytes in protecting against TB is still not clear. However, researchers are working to provide experimental evidence in support of cellular immunity provided by B-lymphocytes. The immune cells surround the site of primary infection and form a mass of tissue called a granuloma.

The roles of some of the cytokines involved in the immune response are described below:
- Interferon-gamma (IFN-γ): T-helper cells release interferon-gamma (IFN-γ) to enhance the killing of *Mycobacterium tuberculosis* present within the macrophage by the production of nitric oxide (NO) and reactive oxygen species (ROS) (Ferluga et al., 2020; Rohde et al., 2007).

- Tumour necrosis factor-alpha (TNF-α) plays a vital role in forming granuloma, activating alveolar macrophages. In addition, TNF-α has several immunoregulatory properties for the containment of the disease (Ferluga et al., 2020; Pal et al., 2016).

- IL-12 is produced mainly by alveolar macrophages after the phagocytosis of TB bacteria and has a crucial role in the production of IFN-γ. Therefore, the expression of IL-12 receptors is increased at the disease site (Pal et al., 2016).

- IL-4 expression causes suppression of IFN-γ production. It is also involved in the activation of macrophages (van Crevel et al., 2002). In the later stages of infection, the expression of IL-4 is related to progression of disease, reactivation of latent infections and intensified tissue damage (Schindler et al., 2001).

- IL-10: The macrophages produce IL-10 after the phagocytosis of TB bacteria (van Crevel et al., 2002). Expression of IL-10 downregulates the production of IFN-γ, TNF-α and IL-12 (Ferluga et al., 2020).
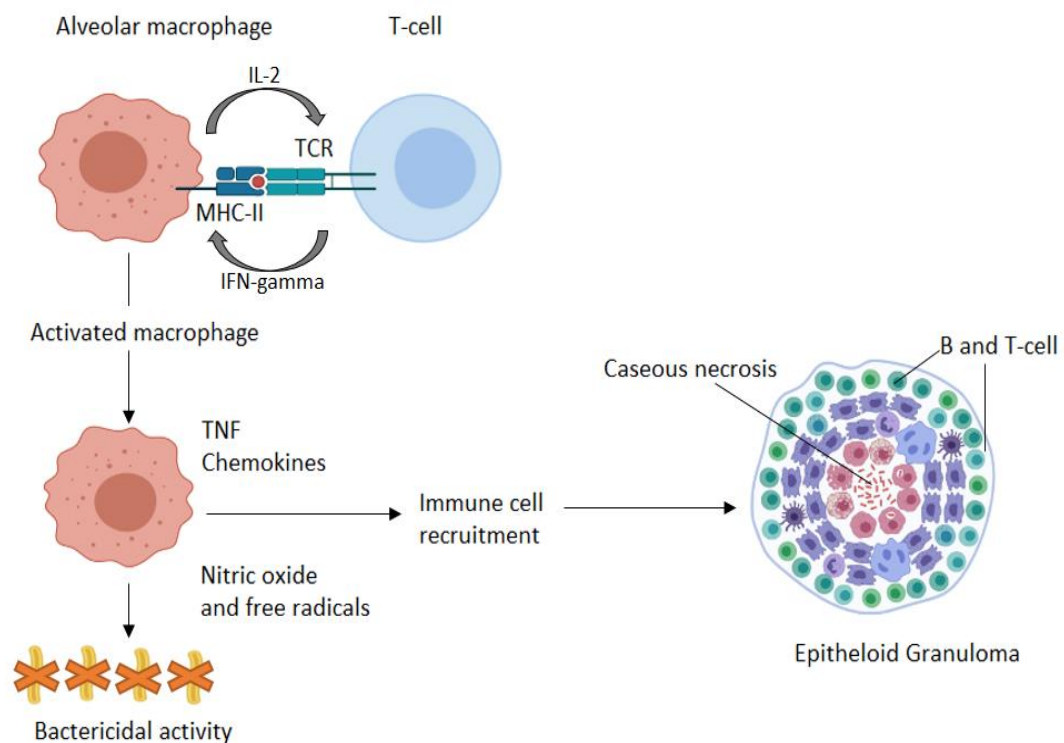
Figure 2. 6: Adaptive immune response against *Mycobacterium tuberculosis.* Figure created using template from Biorender (https://biorender.com/)

2.1.7 Disease progression

The development of a TB infection depends on the patient's immune system. Progression of infection includes primary tuberculosis, secondary tuberculosis, miliary tuberculosis and extrapulmonary

disease (Knechel, 2009; S. Sharma & Mohan, 2004). Primary TB usually occurs in the upper part of the lower lobe and lower part of the upper lobe of the lungs. It is generally asymptomatic and, in most cases, primary TB is contained. In the case of healed TB, the site of infection becomes fibrosed and, later, calcified. The calcification may be seen on a chest X-ray. In other instances, TB infection may progress either to active tuberculosis, i.e., progressive lung disease or miliary tuberculosis. In miliary tuberculosis, the infection spreads from the lung to other parts of the body through lymph nodes or blood vessels. A different fate of a primary infection is latent tuberculosis. In this later case, bacterial growth is arrested but may get activated if the patient's immune system become suppressed. Secondary tuberculosis refers to a pattern of TB that develops in a previously sensitized individual. It can develop from the re-activation of latent TB or by re-infection with *Mycobacterium tuberculosis*. The re-activation of TB can result from alcoholism, smoking, poor nutritional status or reduced immunity. The site of secondary tuberculosis is the apex of the lungs. In extrapulmonary disease, the infection spreads to other parts of the body, such as, the central nervous system (CNS), kidney, liver and the condition becomes fatal.

2.1.8 Survival mechanisms of *Mycobacterium tuberculosis*

Mycobacterium tuberculosis is a pathogen that has developed several mechanisms for subverting the innate and adaptive immune responses.

1. Sulfatides (anionic glycolipids) present in *Mycobacterium tuberculosis*, prevent phagolysosomal fusion facilitating the existence and replication of bacteria (Goren, 1977).

2. *Mycobacterium tuberculosis* can modify the expression of the Rab5 present on the phagosome, thus inhibiting of fusion of phagosome and lysosome (Ankley et al., 2020; Clemens et al., 2000; Vergne et al., 2005).

3. TACO (tryptophan-aspartate containing coat) protein, usually not expressed in macrophages, is present on the phagosome wall. Expression of TACO protein prevents the degradation of *Mycobacterium tuberculosis* by inhibiting the fusion of the phagosome with the lysosome (Ferrari et al., 1999).

4. *Mycobacterium tuberculosis* relies on inhibition of IFN-ɣ by the help of 19-kDa lipoprotein (Hestvik et al., 2005).

5. Two genes, *sodA* and *sodC,* encode the superoxide dismutase required against ROS and convert it to molecular oxygen to protect *Mycobacterium tuberculosis* inside the alveolar macrophages (Piddington et al., 2001).

6. PE11, a PE and PPE protein family member, is involved in the pathogenesis of TB and macrophage persistence (Deng et al., 2015) and PE65 evades the adaptive immune response by obstructing the helper-T cell response (Khubaib et al., 2016).

7. The secretory system of TB bacteria plays a vital role in the pathogenesis of *Mycobacterium tuberculosis*. For example, ESAT-6 regulates macrophage apoptosis and ESX-1 and ESX-5 subvert the defence action of alveolar macrophages (Ates et al., 2016; McLaughlin et al., 2007; Shah & Briken, 2016).

8. Lipoarabinomannan (LAM) in the cell wall plays a vital role in phagosome maturation arrest (Strohmeier & Fenton, 1999).

9. The induction of suppressor of cytokine signalling inhibits the differentiation of helper-T cells, thus, escaping from antigen presentation (Nagabhushanam et al., 2003).

10. *Mycobacterium tuberculosis* can also produce ammonia in large quantities, thereby preventing its acidification.

2.1.9 Signs and Symptoms

Early symptoms of pulmonary tuberculosis can include -

- Cough with sputum for more than three weeks.
- Weight loss, loss of appetite, fatigue and night sweats.
- Fever, with a rising body temperature in the evening.

The infection can progress to a more severe condition of TB with the following symptoms -

- Chest pain.
- Cough with blood-stained sputum.
- Pulmonary shadow in an X-ray.
- Increased erythrocyte sedimentation rate (ESR) that supports the investigation findings.
- Presence of *Mycobacterium tuberculosis* in the clinical sample.
- Additional symptoms relevant to the organ/tissue involved.

2.1.10 Diagnosis of tuberculosis

The World Health Organization stated that millions of people had missed proper diagnosis and care since 2000 (WHO Global Tuberculosis Report, 2020). They need more funding for universal access for diagnosis and testing: "Despite increases in TB notifications, there was still a large gap (2.9 million) between the number of people newly diagnosed and reported compared to the 10 million people estimated to have developed TB in 2019. This gap is due to a combination of the underreporting of people diagnosed with TB and under-diagnosis (if people with TB cannot access health care or are not diagnosed when they do)" (WHO Global Tuberculosis Report, 2020). Efforts are needed to close the gap between the diagnosis and treatment of tuberculosis.

The barriers in controlling TB are early, accurate and rapid diagnosis of the disease. Early detection of the disease helps in preventing transmission and ensures prompt treatment of tuberculosis. The most frequent sample used by a patient with a persistent and productive cough is sputum. Because most mycobacteria grow slowly, three to six weeks may be required to detect them on solid media growth (Knechel, 2009). The diagnosis of TB can be made by two methods: (i) direct detection of actively growing TB bacilli (sputum smear microscopy, culturing TB bacteria and molecular assays), and (ii) indirect detection by diagnosing the immune response of humans against TB bacilli (tuberculin skin test and interferon-gamma release assay). The diagnostic methods used by many countries include chest X-rays, Mantoux test (tuberculin skin test), Interferon Gamma Release assay (IGRA), sputum smear microscopy and biopsies (Katoch, 2004; Mehta et al., 2012). The chest X-ray determines the impact of tuberculosis on the lungs. In the Mantoux test (TST), a tuberculin purified protein derivative (PPD) is injected on the lower part of the arm to detect an active TB infection (Lange & Mori, 2010). The drawback of TST is that it does not detect latent tuberculosis. The diagnostic field in detecting TB has made tremendous advances over the years. New diagnostic methods are molecular-based tests that include amplification of nucleic acid, facilitating the rapid diagnosis of infection. MODS (microscopic observation of drug susceptibility) have been used to detect drug susceptibility and Xpert MTB/RIF is used to detect resistance towards rifampicin (Vadwai et al., 2011).

2.2 Current TB treatment: Are we successfully battling tuberculosis?

This section explains the drug therapy and vaccine (BCG) used for TB treatment and factors impacting the efficacy of current therapies. According to World Health Organization are (WHO, 2010), the aims of a standard treatment regime should be:

- curing patients with active TB to reduce the mortality rate.
- preventing reactivation of TB.
- reducing transmission of tuberculosis.
- preventing the progress of drug resistance.

2.2.1 Drug therapy

In 1943, streptomycin (STP) was the first effective anti-TB drug discovered (Schatz et al., 2005). STP is an aminoglycoside that interferes with mRNA translation by binding with ribosomal proteins and inhibiting protein synthesis of mycobacterial proteins (Flynn & Chan, 2001). At the start, TB patients treated with streptomycin improved, but after a few months of medication, patients' health started to deteriorate. The resistance shown by *Mycobacterium tuberculosis* was identified as the main reason behind the ineffectiveness of streptomycin (Crofton & Mitchison, 1948). Therefore, single anti-mycobacterial drugs were introduced in the 1950s (isoniazid) (Middlebrook, 1954) and 1960s

(rifampicin) (Wehrli et al., 1968). Unfortunately, the use of a single drug for TB treatment saw a significant increase in deaths of TB patients. Between 1970 and 1990, there were several outbreaks of tuberculosis (Keshavjee & Farmer, 2012), causing more people to die. In 1993, tuberculosis was stated as a global public health emergency (Murray & Lopez, 1994). This emergency necessitated the use of multidrug combination therapies with active health programmes - named the directly observed treatment short-course (DOTS) strategy (Grange & Stanford, 1994). The DOTS strategy implemented the supervision of treatment and regular follow-up of patients by community or health care workers to achieve a success rate of treatment of up to 95% (Frieden, 2007). The first-line treatment regimen (DOTS strategy) includes an intensive phase with four drugs (isoniazid (INH), rifampicin (RIF), ethambutol (EMB) and pyrazinamide (PZA)) for two months followed by a continuation phase with two drugs (isoniazid and rifampicin) for next four months. The DOTS strategy involves taking multiple drugs daily for 6-12 months to combat infection and prevent a relapse of TB, depending on the health of the patient and success rate of the treatment (Rook & Hernandez-Pando, 1996).

2.2.1.1 First-line anti-TB drugs

First-line TB treatment drugs comprise isoniazid, rifampicin, pyrazinamide and ethambutol. Figure 2.7 shows that TB drugs' mechanisms are very organized, targeting from unique aspects of cell walls to the genetic machinery of *Mycobacterium tuberculosis.* Most of the first-line drugs have bactericidal activity and are incredibly active against *Mycobacterium tuberculosis.* The most effective TB drug at killing replicating tubercle bacilli is isoniazid, while rifampicin is active against replicating and non-replicating TB bacilli. The current regimen has not changed in nearly 40 years as it is still efficacious in treating drug-sensitive TB patients (Seung et al., 2015). The mechanism of action of each drug is as follows:

(i) Isoniazid (INH)

Isoniazid interferes with the biosynthesis of the cell wall and possesses bactericidal activity. INH is a prodrug that is activated inside isoniazid-susceptible species. Isoniazid passively diffuses through the cell wall and gets activated by the catalase-peroxidase enzyme (*katG*) of the mycobacterial species (Zhang et al., 1992). The activation of isoniazid results in highly reactive oxidants such as superoxide, hydrogen peroxide, alkyl hydroperoxides and highly acylating groups. These oxides attack multiple targets in the cell wall of *Mycobacterium tuberculosis* (Johnsson & Schultz, 1994; Timmins & Deretic, 2006). Specifically, activated INH prevents the action of enoyl-acyl carrier protein reductase (*inhA*) (Quemard et al., 1995), an essential component of the fatty acid synthetase II (FAS-II) complex required for mycolic acid synthesis (Rawat et al., 2003; Vilchèze et al., 2000). Mycolic acids are essential components of the cell wall (Figure 2.4) and inhibition of mycolic acid will disrupt the cell wall, ultimately leading to the death of *Mycobacterium tuberculosis* (Zhang, 2005).

(ii) Pyrazinamide (PZA)

Pyrazinamide is also a prodrug, which is activated into pyrazinoic acid (POA) (Scorpio & Zhang, 1996) with the help of the *Mycobacterium tuberculosis* enzyme, pyrazinamidase (*pncA*) (Salfinger et al., 1990). PZA interferes with cell wall synthesis by targeting the enzyme fatty acid synthase I (FAT-I). PZA has excellent sterilizing activity against semi-dormant *Mycobacterium tuberculosis* bacilli (Sun et al., 2002) and is usually given with rifampicin to shorten the duration of TB treatment (Mitchison, 1985).

Isoniazid (INH) inhibits the action of mycolic acid. INH is activated by catalase-peroxidase (*katG*)

Ethambutol (EMB) inhibits arabinogalactan synthesis. It targets arabinosyl transferase, encoded by *embCAB* operon

Pyrazinamide (PZA) disrupts membrane energetics. PZA is activated by pyrazinamidase (*pncA*)

Mycobacterium tuberculosis

Cell wall

Cell membrane

Rifampicin (RIF) inhibits RNA synthesis by targeting β-subunit of RNA polymerase (*rpoB*)

RNA synthesis

Figure 2. 7: Mechanism of action of first-line TB drugs. The TB drugs target *Mycobacterium tuberculosis* in a very organized manner. First, it weakens the cell wall and cell membrane and targets the *Mycobacterium tuberculosis* genetic machinery. (Figure created using template from Biorender (https://biorender.com/)

(iii) Ethambutol (EMB)

EMB, a bacteriostatic agent, was added to the DOTS treatment regimen to reduce the occurrence of drug-resistant tuberculosis. EMB enters mycobacterial cells through the porin protein channel. It inhibits cell wall synthesis by interacting with the membrane-associated arabinosyl transferase enzyme (*emb CAB*) (Deng et al., 1995). The *embCAB* operon, consisting of three genes *embC*, *embA* and *embB* (Telenti et al., 1997), encodes the protein, arabinofuranosyl transferase. The encoded protein is required to synthesise arabinogalactan, an essential structural component of the cell wall of *Mycobacterium tuberculosis* (Sahu et al., 2015) (Figure 2.4).

(iv) Rifampicin (RIF)

Rifampicin has an excellent bactericidal effect on actively growing *Mycobacterium tuberculosis* cells and a sterilizing action on semi-dormant tubercle bacilli (Somoskovi et al., 2001). Once inside

Mycobacterium tuberculosis, RIF inhibits mRNA transcription by binding to the β-subunit of the DNA-dependent RNA polymerase (RNAP) (Pang et al., 2013). Thus, the inhibition of protein synthesis leads to cell death; however, the actual mechanism of destruction of *Mycobacterium tuberculosis* cell by transcription inhibition is not entirely understood (Koch et al., 2014).

2.2.1.2 Drug resistance

Drug resistance is described as a decrease in the sensitivity of a strain to a satisfactory degree when it encounters a drug. The emergence of drug resistance in tuberculosis is not a new phenomenon. *Mycobacterium tuberculosis* strain*s* showed resistance to streptomycin in 1944 (Zhang & Yew, 2009). The evolution of the *Mycobacterium tuberculosis* complex has led to the development of survival strategies, linked with point mutations in the coding region of first-line TB drug targets (catalase-peroxidase, pyrazinamidase, arabinosyl transferase and DNA-directed RNA polymerase subunit beta), leading to the daunting scenario of drug resistance. Drug-resistant TB bacterial strains have many mechanisms to make drugs ineffective, such as preventing entry of drug molecules into the bacterial cell with the help of impermeable cell walls, alteration of the drug target protein by random, single or multistep chromosomal point mutations and expelling the drugs by the transmembrane drug efflux systems of the bacterial cell wall (Cohen et al., 2014; Nachega & Chaisson, 2003; Peñuelas-Urquides et al., 2018). Other factors associated with the available tuberculosis treatments (Chan & Iseman, 2002; Jasmer et al., 2002; Narita et al., 1998; Szakacs et al., 2006; van den Boogaard et al., 2009; Volmink & Garner, 2007) are:

- inadequate or inefficient administration of TB drugs and ignorance of healthcare workers in the treatment and control of TB
- poor case holding, use of poor quality/sub-standard drugs and insufficient or irregular TB drug supplies
- drug intolerance and toxicity
- poor compliance of patients to the prescribed regimens.
- availability of anti-TB drugs without prescription
- illiteracy and low socio-economic status of patients
- coinfection with HIV.
- delay in the identification and susceptibility testing of *Mycobacterium tuberculosis* strains and failure to identify pre-existing drug resistance.

Tuberculosis bacteria showing resistance to a single first-line drug are known as mono-resistant TB. Resistance to more than one drug, other than isoniazid and rifampicin, is called polyresistant TB. Different levels of drug resistance (monoresistance, polyresistance and multi-drug resistance) and failure to the complete therapy as prescribed can lead to the unfortunate outcomes from of treatment,

including the risk of treatment failure, continued transmission of *Mycobacterium tuberculosis* isolates, disease relapses, drug resistance, and death.

The drug resistance mechanism in tuberculosis can be categorised as follows (Lemos & Matos, 2013):

(i)   **Intrinsic resistance –** this refers to the natural or inborn ability of tubercle bacilli to resist the action of a drug through making changes in its fundamental structural properties or functions, irrespective of previous exposure to the drug. For example, *Mycobacterium tuberculosis* has evolved numerous intrinsic resistance mechanisms for neutralizing the toxicity of chemicals, including anti-TB drugs (Nguyen, 2016). The main factors involved in the intrinsic resistance mechanism are as follows:

- Cell wall permeability: the *Mycobacterium tuberculosis* cell wall is usually thick and hydrophobic due to the presence of mycolic acid and arabinogalactan (Jarlier & Nikaido, 1994). The unusual composition and structure of the cell wall plays a vital role in the resistance mechanism and provides a hydrophobic barrier to hydrophilic substances, including antibiotics (Karakousis et al., 2008).

- Efflux pump: *Mycobacterium tuberculosis* expels drugs or chemical reagents with the help of efflux pumps (Paulsen et al., 2001). The *Mycobacterium tuberculosis* genome encodes 26 ABC transporters and 18 facilitator proteins, which act as drug exporters (Braibant et al., 2000). The membrane-spanning proteins perform vital functions in the TB bacteria's physiology and metabolism, such as transportation of nutrients, toxins, or signalling molecules through the cell wall. Consequently, the antibiotic resistance roles of many carriers could be secondary and attributable to non-specific transport (Nguyen, 2016).

- Drug target mimicry: *Mycobacterium tuberculosis* neutralizes fluoroquinolones by a molecular mimicry mechanism. Overexpression of the *mfpA* gene results in setting DNA gyrase free from drug attack (Ferber, 2005).

(ii)   **Primary resistance -** Is defined as drug resistance in a TB patient who has never received any anti-TB drugs previously (Zhang & Yew, 2009). Primary drug resistance can occur by the transmission of resistant tubercle bacilli from one infected individual to another. The most significant factors contributing to drug resistance are poor treatment regimens, non-compliance with drugs and poor diagnostic techniques.

(iii)   **Acquired resistance -** Acquired resistance is defined as resistance developed in a patient where earlier treatment was inadequate. Acquired drug resistance usually occurs through horizontal gene transfer or mutations (Nguyen, 2016). First-line TB drugs have been

designed for *Mycobacterium tuberculosis* target proteins that play a vital role in biological processes and functions, such as physiology, metabolism and translation of proteins. Thus, a mutation in a drug target protein causes structural or functional changes, eventually leading to a decrease in the drug's binding affinity or preventing the drug from binding to its respective target. A mutation changes the nucleotide base and that may or may not change its expression, resulting in different or similar protein sequences. Mutations can impact the biological function of the resulting protein either by loss of function or gain of a new function. Depending on the resulting protein, mutations can be classified into the following categories:

(i) Nonsense mutation: A mutation in the DNA leads to a premature stop codon in the transcribed mRNA, resulting in an abnormal function or non-functional protein product.

(ii) Silent mutation (synonymous mutation): A mutation occurring within a codon that does not alter the resulting amino acid protein sequence.

(iii) Missense mutation (non-synonymous mutation): A mutation in the codon that results in an amino acid substitution in the final protein product.

There are few other types of mutations: insertion (addition of a new nucleotide base), deletion (removal/deletion of one or more nucleotide bases) and a frameshift mutation (insertion or deletion in a coding frame, resulting in an entirely different protein product).

The compartmentalization of the *Mycobacterium tuberculosis* cell and a thick lipid-rich cell wall prevents the transfer of genetic material across it. Spontaneous chromosomal mutations are the leading cause of acquired resistance in *Mycobacterium tuberculosis*, as horizontal gene transfer has not been reported (Nachega & Chaisson, 2003; Nguyen, 2016; Zhang & Yew, 2009). Almost 20% of all TB strains are resistant to at least one primary TB drug and there is a growing incidence of multi-drug resistant tuberculosis (MDR-TB). Evolution has led to the development of survival strategies in TB bacteria, which are linked with point mutations, insertions or deletions in the first-line drug targets, leading to the daunting scenario of drug resistance (Cohen et al., 2014; Heym et al., 1994; Nachega & Chaisson, 2003; Telenti et al., 1993). Prolonged exposure to drugs helps in acquiring spontaneous chromosomal mutations in *katG* (Heym et al., 1995; Rouse et al., 1996), *rpoB* (Telenti et al., 1993), *pncA* (Scorpio et al., 1997; Sreevatsan et al., 1997) and *emb* (Ramaswamy et al., 2000) for INH, RIF, PZA and EMB resistance, respectively. Some laboratory studies that have been conducted to identify mutations in the first-line targets include: (i) mutations identified in *katG* S315T (Rouse et al., 1996), D419H, M420T, D542H, and R632C (Ando et al., 2010), (ii) studies of *rpoB* identified F505L (Heym et al., 1994), (iii) *pncA* studies predicted H51P, F58L, S67P, T114P, V130G (Hirano et al., 1998), and (iv) mutations in *embB* M306I, M306L, M306V, F330V, T630I (Ramaswamy et al., 2000).

2.2.1.3 Global TB drug pipeline

The World Health Organization has recommended first-line, second-line and third-line drugs to treat drug-susceptible and drug-resistant tuberculosis. The first-line drugs for the treatment of drug-susceptible TB are a combination of the already mentioned INH, RIF, PZA and EMB. Second-line anti-TB drugs are used to treat patients who have shown drug resistance to first-line drugs and the treatment time is extended from six to nine months (Cheng et al., 2004). Fluoroquinolone, amikacin, capreomycin, kanamycin and viomycin are second-line TB drugs for drug resistance (Ahmad & Mokaddas, 2014; Peñuelas-Urquides et al., 2018). Second-line anti-TB drugs target many essential functions in tubercle bacilli, such as DNA synthesis, transcription, translation and energy metabolism pathways (Hoagland et al., 2016; Sun et al., 2002). However, these drugs are less active and have a higher degree of toxicity and intolerability in patients. The third-line anti-tuberculosis drugs, linezolid, amoxicillin, ciofazine and thiacetazone treat multi-drug resistant tuberculosis (MDR-TB). However, the efficacy of these drugs is not yet evident (Dooley et al., 2013). The current treatment regime for TB includes drugs that are around 40 years old (Seung et al., 2015) and the success rate has reduced from 95% to 85% (WHO Global tuberculosis Report, 2020). The emergence of drug resistance has led to the inhibition of important drug targets. Several programmes for drug development are in process to determine novel drugs that target essential mechanisms, such as energy metabolism, virulence, persistence, biosynthesis of the cell wall and signal transduction to ensure a shorter duration and safer TB regime (Lamichhane, 2011; Singh & Mizrahi, 2017). The current global TB drug development pipeline includes 22 drugs in phase-I, phase-II and phase-III of clinical trials, as shown in Figure 2.8 (WHO Global tuberculosis Report, 2020).
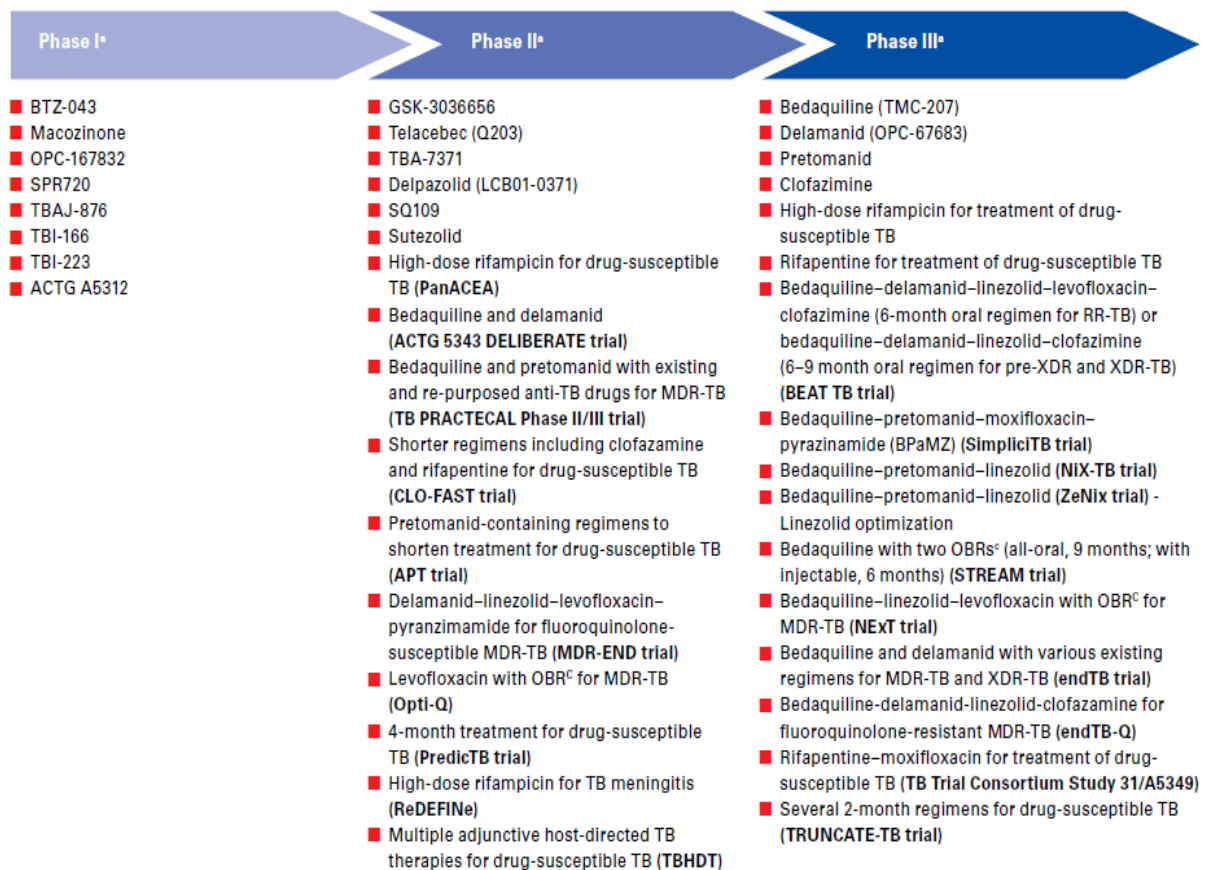
Figure 2. 8: Global drug discovery development and treatment regimen for treating tuberculosis disease. (New drugs are listed followed by treatment regimen) (Sourced from WHO Global tuberculosis Report, 2020)

2.2.2 TB vaccine

A vaccine is considered the most effective method for combating an infectious disease in reducing incidence rates and costs and increasing safety. To achieve the WHO goals of End-TB strategy, an effective vaccination programme is required to prevent TB transmission.

2.2.2.1 BCG

For TB, the only licensed vaccine available is the BCG (Bacillus Calmette-Guérin). BCG was developed by Albert Calmette and Camille Gurein, in 1921, at the Institute Pasteur in France (Herr & Morales, 2008). Calmette and Guerin followed Louis Pasteur's principle for constructing a live-attenuated vaccine against the infectious disease. BCG is a live attenuated vaccine prepared from *Mycobacterium bovis,* the causative agent of TB in cattle (Brosch et al., 2007). *Mycobacterium bovis* has shown more than 95% sequence similarity to *Mycobacterium tuberculosis* (Garnier et al., 2003). The administration of BCG to children began in 1921 with no, or few, significant side effects. In 1960, the World Health

Organization started developing policies for BCG vaccination and in 1974, BCG was part of the Expanded Program of Immunization (EPI) for infants (Lahariya, 2014).

When BCG is injected intracutaneously, the host body initiates the process of an anti-tuberculosis immune response. First, toll-like receptors (TLR-2, 4 and 9) (Bulut et al., 2005; Quesniaux et al., 2004) present on the APCs (macrophages and dendritic cells) recognize, internalize the injected components, and stimulates the production of different costimulatory molecules (innate immune response against BCG) (Moliva et al., 2017). This leads to the activation of the adaptive immune system (helper-T cells and cytotoxic-T cells) (Andersen & Kaufmann, 2014), resulting in cytokine production; for example, interferon- ɣ and granzymes (Covián et al., 2019; Kaufmann, 2013; Moliva et al., 2017). IFN- ɣ activates the macrophages leading to the destruction of vaccine components. The role of humoral immunity is not well understood in response to BCG (Dockrell & Smith, 2017). Memory immune cells (memory B cells, helper-T cells and cytotoxic-T cells) are then generated for eliminating a TB infection in the future. The administration of BCG can cause a mild infection at the site of injection.

Currently, BCG immunization is one of the widely implemented vaccination programmes in the world. The main reason for the success of BCG is its cost-effectiveness (Doherty & Andersen, 2005). The BCG vaccine is given to neonates and infants in countries with a low incidence of TB disease. However, in countries with high incidence rates, the BCG vaccine is also given to young children (WHO Global Tuberculosis Report, 2020).

2.2.2.2 Failure of BCG

Although BCG has shown its effectiveness in reducing the incidence of tuberculosis in children, miliary TB and tubercular meningitis, there are major concerns regarding adult TB (Brewer, 2000). BCG trials have shown that it does not induce the same kind of immune response as when the vaccine is given to premature babies. Disseminated BCG disease has been observed in HIV-infected children following the BCG vaccination. The safety of BCG remains a significant concern in immunocompetent persons (Hesseling et al., 2007). BCG has shown 0-80% efficacy against pulmonary tuberculosis in adults (Doherty & Andersen, 2005). The effectiveness of BCG varies from person to person, country to country and children to adult (Briassoulis et al., 2005; Colditz et al., 1994). The genetic variability among the sample population significantly impacts vaccine efficacy.

BCG generates a higher helper-T cell response than a cytotoxic-T cell response (Lindestam et al., 2015). CTL plays direct role in killing pathogenic bacteria. Despite the efforts of BCG immunization programmes, the incidence rate of tuberculosis has not lowered in recent years. The inefficiency of BCG in reducing the prevalence of disease and not protecting adults is, so far, not well understood.

Some of the crucial factors that contribute to lowering the efficacy of BCG might include (Barreto et al., 2006; Doherty & Andersen, 2005; S. Gupta et al., 2011):

- *Mycobacterium bovis* is less virulent and not a primary causative agent of TB
- diversity in TB strains
- the ability of TB bacteria to evade the immune response
- over-attenuation of the presently used BCG strain
- differences in clinical assay
- different levels of protection against the clinical forms of tuberculosis
- malnutrition
- health condition of an individual
- MHC-I and II polymorphism among the vaccinated population.

Other factors that impact BCG are tobacco and alcohol consumption, pollution, and the presence of other diseases, such as, HIV/AIDS, cancer and diabetes in individuals (Gupta et al., 2011).

2.2.2.3 TB vaccine in clinical trials

TB prevention is a long-term goal. There has been considerable progress made in TB vaccines using different approaches including the use of recombinant protein or DNA attenuated *Mycobacterium tuberculosis*, virus as a vector, novel adjuvants and potent antigen delivery systems. The global effort to eradicate tuberculosis has led to 14 vaccines being in various stages of clinical trials (WHO Global Tuberculosis Report, 2020) (Figure 2.9). There are three vaccines in phase-I, nine in phases-II-a and II-b and two in phase-III of clinical trials. Ad5 Ag85A is a recombinant vectored vaccine using adenovirus serotype 5 that expresses *Mycobacterium tuberculosis* Ag85A. ChAdOx1.85A/MVA85A vaccine uses simian adenovirus as a vector for expressing Ag85A and TB/FLU-04L is a recombinant influenzae vectored vaccine expressing *Mycobacterium tuberculosis* Ag85A and ESAT-6 (Wilkie et al., 2020). MTBVAC is a live-attenuated vaccine produced from a clinical isolate of *Mycobacterium tuberculosis* Mt103 by deleting the *phoP* and *fadD*26 genes (Arbues et al., 2013; Tameris et al., 2019). DAR-901 is a heat-killed vaccine produced using the *Mycobacterium obuense* strain (von Reyn et al., 2017). The vaccine was developed by researchers at Dartmouth University. RUTI is a polyantigenic vaccine made by detoxifying and fragmenting *Mycobacterium tuberculosis* encapsulated in a liposome (Ahsan, 2015). H56:IC31 is a recombinant subunit vaccine containing Ag85A, ESAT-6, Rv2660c with the IC31 adjuvant. M72+AS01E is a recombinant subunit vaccine containing *Mycobacterium tuberculosis* antigens, 32A and 39A, along with the AS01E adjuvant (Brazier & McShane, 2020). The two vaccines in phase-III are VPM 1002 and MIP/Immuvac. MIP is produced by heat-killing *Mycobacterium indicus pranii* to prevent the high risk of causing tuberculosis (Sharma et al., 2017).

The vaccines in the clinical trial mainly depend on inducing an immune response using either weak mycobacterial strains or using two to three antigens to produce recombinant subunits and viral vectored vaccines. These vaccines do not guarantee to yield a broad coverage immune response. The drug-resistant issue is also not adequately addressed by any of the vaccine in the clinical trial. The emergence of drug resistant *Mycobacterium tuberculosis* emphasizes the need for a new and effective TB vaccine.
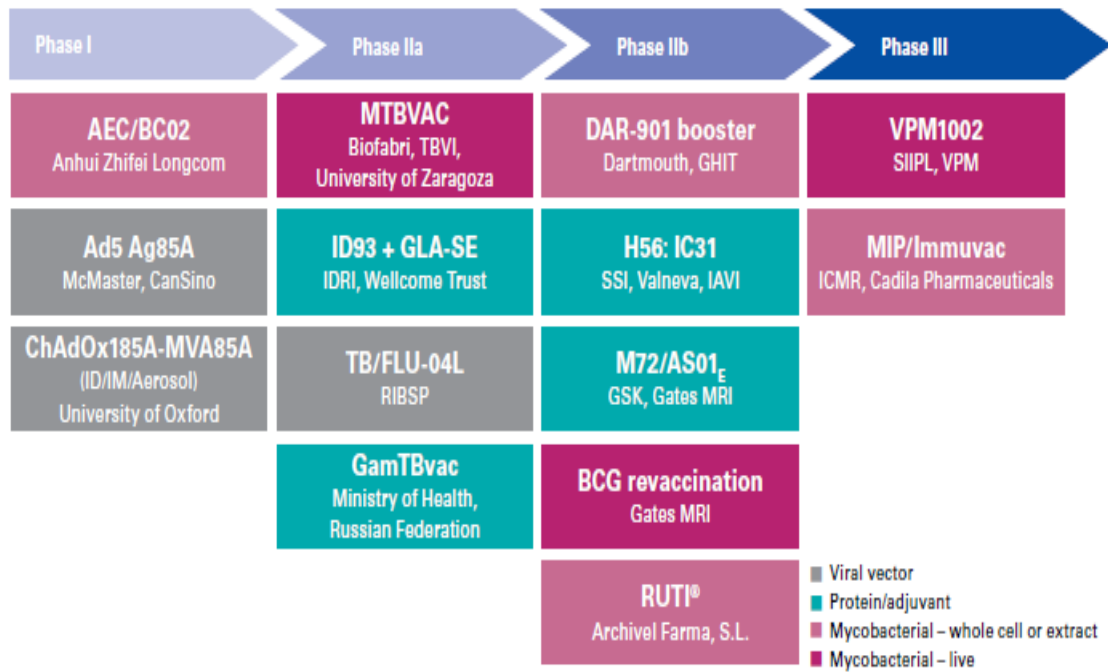
Figure 2. 9: TB vaccine candidates in phases-I, -II and -III of clinical trials in 2020. A total of 14 vaccine candidates are in clinical trials. (Sourced from WHO Global tuberculosis Report, 2020)

2.3   Major obstacles in developing an effective TB treatment

Current therapeutic and diagnostic strategies are inadequate to eliminate TB by 2050 (Ottenhoff & Kaufmann, 2012). The complete eradication and elimination of TB from the world needs more effective and safer therapeutic tools. For this reason, it is crucial to have an in-depth knowledge of mechanism of TB infection when developing an efficient TB treatment.

2.3.1   Challenges to providing an effective drug treatment

First-line drugs have been designed to target proteins that play a vital role in biological processes and functions, such as the physiology, metabolism and translation of proteins within *Mycobacterium tuberculosis*. Figure 2.10 describes the critical challenges in providing efficient drug treatment:
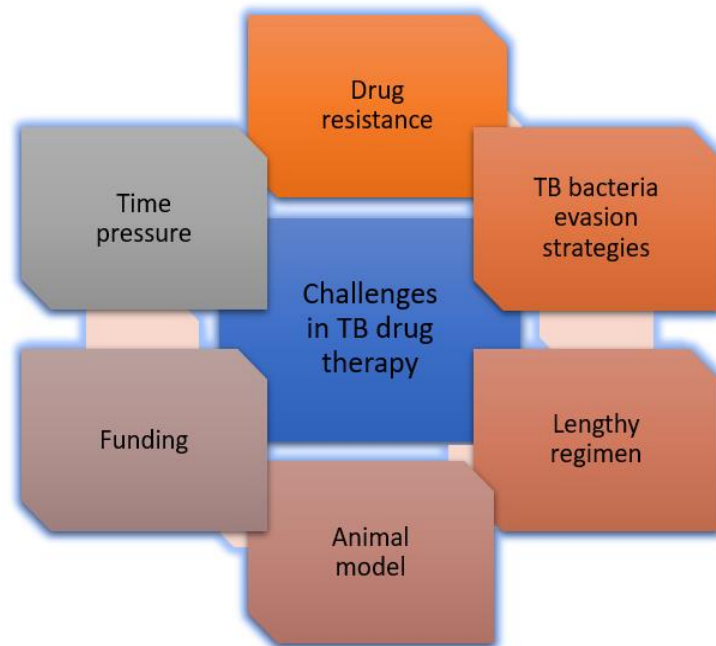
Figure 2. 10: Challenges in providing efficient tuberculosis drug treatment

1. **Drug resistance** - The emergence and transmission of TB drug-resistant strains obstruct the efforts to control and manage TB. Multi-drug-resistant TB (resistant at least to rifampicin and isoniazid) occurs due to the sequential accumulation of mutations in the drug target's genes (Zhang & Yew, 2015).

2. **Selection of bacterial evasion strategies -** *Mycobacterium tuberculosis* must follow some strategic pathway to select a drug-resistant mutation to balance the impact of drug-resistant mutations on the fitness of *Mycobacterium tuberculosis* (Gagneux et al., 2006).

3. **Understanding compensatory mechanisms -** If a drug-resistant mutation alters the function and structure of *Mycobacterium tuberculosis*, there must be a compensatory mechanism to compensate for TB bacteria's fitness cost (Comas et al., 2012).

4. **Delay in detecting drug-resistant TB bacilli -** TB patients are sometimes not assessed correctly as drug-resistant positive, resulting in inappropriate treatment for TB.

5. **Lengthy treatment regimen -** Prolonged exposure to first-line TB drugs can allow *Mycobacterium tuberculosis* to potentially evolve more resilient drug resistance strategies (Ginsberg & Spigelman, 2007).

6. **Animal model** - Most animal models are not able to replicate human TB disease. So, understanding the pathology of TB is quite a difficult task (de Rycker et al., 2018).

7. **Insufficient funding** - Despite the progress in the global drug development pipeline, more funding is needed for developing new drugs and improving the pipeline.

8. **Time** - Drug discovery and development is a time-intensive process. If a drug fails in a clinical trial, it will waste not only money but also time.

9. Some other challenges include:

    - improper use of drugs

    - costly MDR drugs

    - toxic side effects of MDR drugs

2.3.2 Challenges in developing an effective TB vaccine

BCG, a hundred-year-old vaccine, is the only registered vaccine available to prevent tuberculosis, but there is still a high morbidity and mortality rate in various countries from this disease. Moreover, over the previous decades, there have been no significant achievements in vaccine development. Figure 2.11 shows some of the hurdles in TB vaccine development.

- Evolution of resistant *Mycobacterium tuberculosis* strains
- Understanding of host- TB bacteria interaction
- Evasion from immune response

Pathogen diversity

Understanding of immune response

- Inadequate innate immunity
- Poor immunogenicity

- MHC allele variability
- Ethnicities

Host variability

Environmental & geographic factors

- Poor health
- Co-infection
- Dense population
- Pollution

- Autoimmunity
- Toxic

Vaccine safety

Time & cost

- Expensive laboratory testing &validation
- Time in culturing slow growing *Mycobacterium tuberculosis*

Challenges in TB vaccine development

Figure 2. 11: Major obstacles in TB vaccine development. Some crucial challenges include pathogen diversity, host variability and vaccine safety

1. **Pathogen diversity** - The emergence of resistant strains has caused the genetic variability among *Mycobacterium tuberculosis.* The current vaccines in clinical trials mainly use either weak mycobacterial strains or two to three antigens for producing recombinant subunits and viral vectored vaccines. The most significant issue with these vaccines is that they do not guarantee to yield a broad spectrum of immune response. In addition, the knowledge gap in understanding the human-TB bacteria interaction, host evasion strategies and poor understanding of *Mycobacterium tuberculosis* antigens is of significant concern.

2. **Understanding of host immune response** - Understanding the correct functioning of the human immune system is very complex. Therefore, the identification of novel TB antigenic proteins and their immunodominant regions (B-cell and T-cell epitopes) using the conventional approach remain a significant challenge in TB vaccine development. However, an in-depth knowledge of host responses is critical for developing a TB vaccine to generate a strong innate and adaptive (humoral and cell-mediated) immune response.

3. **Host variability** - MHC molecules are highly polymorphic and more than a thousand HLA (human leukocyte antigen) alleles for humans have been identified. Furthermore, the different HLA types are expressed at different rates in different ethnicities worldwide. Thus, developing a TB vaccine needs an analysis of population coverage to minimize the risk of developing an ethnically biased vaccine (Bui et al., 2006).

4. **Vaccine safety** - The reversion to a disease-causing pathogen is a significant risk involved with a viral-vectored vaccine. Therefore, the vaccine should not contain any toxic compounds that can cause severe side effects after vaccination. In addition, the use of antigens that cause autoimmunity or hypersensitive reactions in a host is also a major issue in vaccine development.

5. **Time and cost** - The conventional process of using whole organisms and live-attenuated or heat-killed organisms in vaccine development and production is very costly and it takes many years to develop an effective vaccine. Traditional vaccine development requires cultivating *Mycobacterium tuberculosis* in the laboratory and then performing different biochemical, microbiological, immunological tests to detect the vaccine candidates. This approach is laborious, requires expensive experimental testing and sometimes fails to reveal suitable antigens. In addition, the purification and detoxification of vaccine products are costly.

6. **Environmental and geographic factors** - Poor health, co-occurrence of HIV/AIDS and diabetes, poverty, smoking and alcohol consumption, pollution and the economic status of the country indirectly impact the development of a TB vaccine. The COVID-19 pandemic caused a 50% drop in the correct diagnosis of TB cases, and this will lead to 400,000 more deaths due to TB in the future (WHO Global tuberculosis Report, 2020).

2.4   Potential solution to the TB treatment problem

Mycobacterium tuberculosis has developed several ways to destabilize the human immune response by adapting to the changing environment of the hosts and spreading the infection to various parts of the body. The high level of adaptation has led to the evolution of TB bacteria, resulting in resistant strains. The available TB drugs and vaccine are successfully evaded by genetically encoded resistance mechanisms in TB bacteria. The widespread emergence of TB resistant strains has challenged the view of tuberculosis as a treatable disease. In this research, an attempt is made to provide a solution for

developing an effective TB treatment using an in-depth understanding of host-TB bacteria interaction, survival strategies of the bacteria and challenges in developing TB drugs and vaccines.

2.4.1 TB drug therapy

Drug resistance is emerging at an alarming rate, thus requiring more research to address this situation. The success rate of first-line TB drugs is 85%. Thus, there is a need to understand the underlying drug resistance mechanisms more deeply to help improve the success rate of drugs and achieve the long-term goal of the End-TB strategy. The non-synonymous mutations in *katG*, *pncA, emb* and *rpoB* are the main reason for drug resistance in the first-line TB drugs. The mutations lead to alterations in drug binding sites, modification of the whole drug target, inactivation of drug or preventing the entry of drug inside *Mycobacterium* tuberculosis. Thus, a mutation in a drug target protein causes structural or functional changes, eventually leading to a decrease in drug binding affinity or preventing the drug from binding to its respective target. Therefore, the analysis of first-line mutation data accumulated over several decades is crucial to improving our knowledge of the underlying drug resistance mechanisms. In addition, the whole spectra of mutations need to be studied to understand the *Mycobacterium tuberculosis* survival strategies and the impact of mutations resulting in:

- functional changes (leading to modification of target),
- stability changes with destabilizing or stabilizing target protein,
- irregular binding (reduced affinity for drug resulting in its inactivation) or tighter binding with drug (prodrug is activated but not released),
- altered functionally relevant residues (changes in residues in binding site, catalytic site or residues directly interacting with active site),
- protein sequence conservation and
- hotspot sites within drug target (some mutations are more prevalent at specific positions or regions.

This thesis research proposes a solution for in-depth understanding of the *Mycobacterium tuberculosis* survival strategy against first-line TB drugs through the analysis of the impact of mutations on the structure and function of targets and weakening the effectiveness of drugs. Chapter- 3 of the thesis explains the approach developed in this research based on a comprehensive bioinformatics study.

Examining the impact of mutations on the drug-target protein structure and function in the laboratory can be time consuming and costly. Bioinformatics provides advanced and user-friendly algorithms crucial to managing and analysing the genomic data available from sequencing techniques.

Computational analysis has been an excellent method to understand the impacts of mutations on structure and function as well as the drug binding affinity of the first-line TB drug target proteins. Therefore, different bioinformatics tools have been used in this research to predict the impact of non-synonymous amino acid substitution. Furthermore, using a combination of bioinformatics tools, this research will be able to identify these effects in detail.

Most of the studies conducted in the past determined changes in drug binding affinity by comparing the wild-type TB drug target protein's binding energy with the that of the mutant protein. However, the nature of the drug resistant mutations and the specific changes they cause in first-line TB drug targets and how they afford drug resistance and their impact on *Mycobacterium tuberculosis* fitness is currently not well understood.

The analysis of each mutation will help explore the specific drug resistance mechanisms it confers that would provide deep insights into some of the unnoticed features from the limited studies in the past. A deeper understanding of the survival of *Mycobacterium tuberculosis* may guide us in:

- optimizing the action of currently available drugs by designing new inhibitors for drug resistance
- developing new drugs and designing new TB therapy regimens
- creating an atlas of drug-resistance mutations as a reference in diagnosis
- documenting an in-depth view of bacterial survival strategies and their development for future reference
- developing new tools for proper diagnosis of drug-resistant TB patients.

2.4.2 TB vaccine

The complete eradication and elimination of TB from the world requires developing a more effective and safer vaccine. The generation of a strong and swift immune response that can prevent disease progression and transmission is a prerequisite for a TB vaccine. New approaches for vaccine development must focus on host-pathogen interactions, the limitations of BCG and drug-resistant *Mycobacterium tuberculosis* strains (Ahsan, 2015). Helper (CD4+) and cytotoxic-T (CD8+) cells and cytokines released by T-cells play an essential role in providing optimum protection against a tuberculosis infection (Deenadayalan et al., 2013; Flynn, 2004). As stated earlier, BCG generates a higher helper-T cell response than a cytotoxic-T cell response. Thus, developing a new TB vaccine with improved efficacy must create a pool of effective B-cell, T-helper and T-cytotoxic type memory cells.

In the current research, we developed an epitope-based vaccine using computational vaccinology that can help in treating TB. Chapter-4 of this thesis explains the proposed computational framework for addressing the challenges of TB vaccine development to produce an effective TB vaccine.

The current explosion in bioinformatics has revolutionized the field of vaccine development. Bioinformatics provides new tools that facilitate identification of the potential vaccine targets without the need to culture pathogens in the laboratory. The application of bioinformatics in biotechnology, immunology and vaccinology has allowed the development of the more revolutionary field of computational vaccinology (CV) (Vivona et al., 2008). The present study incorporates different branches of computational vaccinology, such as, comparative analysis, reverse vaccinology, immunoinformatics and structural vaccinology to identify potential vaccine candidates against tuberculosis. Computational vaccinology is a low-cost technique, entirely feasible for use on the plethora of genomic data being generated. Therefore, it is justifiable for use on a broad range of pathogens. The introduction of new steps in reverse vaccinology and immunoinformatics tools introduced in this study, will improve the search for identifying T-cell and B-cell epitopes as potential vaccine candidates for *Mycobacterium tuberculosis* and help in constructing an epitope-based vaccine for TB.

To address issues, such as antigen variability, drug resistance and broad coverage of the immune response, a large number of *Mycobacterium tuberculosis* strains should be used to identify vaccine targets. Thus, to develop a universal vaccine for TB, there is the necessity to include these *Mycobacterium tuberculosis* strains in the vaccine development approach and identify the conserved antigens among them. Selecting novel antigens or vaccine candidates is a stumbling block in the vaccine design process. Advancements in genomics have provided information regarding the genome, transcriptome or proteome of a pathogen (Brusic & Flower, 2004) to help identify novel vaccine candidates.

Computational reverse vaccinology (CRV) allows the analysis of genomic sequences to predict surface-exposed antigenic proteins that can be used as potential vaccine candidates for subunit vaccine preparation (Flower et al., 2010). Many of the essential and virulent TB vaccine candidates are *Mycobacterium tuberculosis* membrane-bound proteins. For developing an effective epitope-based vaccine that can elicit a cell-mediated and humoral response, the epitopes (B-cell, CD4+ and CD8+ T-cell epitopes) identified from these membrane-bound proteins can be promising targets for the vaccine. Compared to the conventional approach, computational vaccinology helps to develop a safer vaccine as careful selection of non-toxic, non-allergenic immunodominant epitopes is undertaken. The

main advantage of an epitope-based vaccine is the removal of deleterious epitopes that can cause cross-reactive reactions or autoimmunity in the host. Epitope-based vaccines are more potent and, when controlled correctly, induce a specific immune response to a broad range of immunodominant epitopes, target multiple-conserved epitopes and break immune tolerance. An epitope-based vaccine helps in the selection of vaccine candidates that are not harmful when following safety measures of the laboratory. The proposed vaccine can be further studied *in vivo* and *in vitro* to prove its effectiveness in mounting an immune response.

Chapter 3

A comprehensive analysis of first-line tuberculosis drug resistant mutations to understand the survival strategy of *Mycobacterium tuberculosis*

Chapter-3 aims to accomplish objective-1 of our study by providing a potential solution to reduce the burden of the emerging global drug resistance problem. In chapter-2, we discussed the available TB treatments and the enormous efforts expended to eradicate this disease. Unfortunately, the emergence of drug resistance has led to a decrease in the success rates of TB drugs from 95% to 85%. This chapter explains the holistic analysis performed in this research to gain deeper insights into drug resistance mechanisms in TB based on a comprehensive bioinformatics study. This chapter comprises six sections. Section 3.1 provides an overview of chapter-3. Section 3.2 discusses the current understanding of drug resistance and what needs to be undertaken to improve our knowledge to unravel drug resistance mechanisms to gain insights into bacterial survival strategy. This section also describes the framework developed in this study for this purpose. The materials and methods used to implement the framework are discussed in section 3.3 and the results are discussed in section 3.4. Finally, a summary of the chapter is given in section 3.5.

3.1   Overview

TB is a deadly disease caused by an infectious bacterium, *Mycobacterium tuberculosis,* which poses a significant threat to global health. According to the world's oldest literature, TB plagued some of humanity's earliest civilisations (Zimmerman, 1979). Regardless of the advancements made in medical sciences, TB remains the cause of death of 1.4 million people, with 10.0 million new cases worldwide in 2019 (WHO Global tuberculosis Report, 2020). TB drug treatment includes the **d**irectly **o**bserved **t**reatment **s**hort-course (DOTS) strategy comprising intensive and continuation phases. The rapidly increasing mono-, poly- and MDR-TB cases with 0.5 million new MDR-TB cases (drug resistance shown to INH and RIF) and 0.3 million deaths are a frightening situation leading to treatment failure (WHO Global tuberculosis Report, 2020). Although the first-line drugs play a pivotal role in combating TB, the emergence of drug resistant TB strains, the improper use of drugs, patients' non-compliance with the treatment, drug intolerance and toxicity, delayed or incorrect diagnosis, and limited access to medicines has heightened the burden of TB globally (Bhat et al., 2018; Portelli et al., 2018). Improving the success rate of the current TB treatment and eliminating TB from its roots requires an in-depth knowledge of the survival strategies of *Mycobacterium tuberculosis* against first-line TB drugs. Our

study aims to understand drug resistance mechanisms through a comprehensive and systematic in-depth analysis of global mutation data for *Mycobacterium tuberculosis* as reported over the last 30 years. Our study investigates drug resistance mutations, their frequency, global spread, and the evolution of mutations over time, to explore the nature and geographic spread of drug-resistant mutations. Importantly, we unravel the mechanisms of drug resistance through an in-depth look into the interplay between *Mycobacterium tuberculosis* and drugs, and how mutations alter it in favour of its survival.

3.2 How we can unravel the TB drug resistance mechanisms

3.2.1 TB Drug resistance - what we know

The emergence of drug resistance in tuberculosis is not a new phenomenon, as *Mycobacterium tuberculosis* strain*s* showed resistance to streptomycin in 1944 (Crofton & Mitchison, 1948). The current DOTS strategy for TB treatment comprises: (i) an intensive phase including a combination of four first-line drugs rifampicin (RIF), isoniazid (INH), pyrazinamide (PZA) and ethambutol (EMB) (two to four months), and (ii) a continuation phase with INH and RIF (four to six months) followed by four months of continuation phase (Grange & Stanford, 1994). The drugs used in the current treatment regime are approximately 40 years old. Unfortunately, the resistance showed by *Mycobacterium tuberculosis* has led to heightened burdens of TB across the globe.

Several *Mycobacterium tuberculosis* drug resistance mechanisms have been put forward (Cohen et al., 2014; Nachega & Chaisson, 2003). For example, preventing the entry of drug molecules into *Mycobacterium tuberculosis* with the help of its impermeable cell wall, expelling of drugs by transmembrane drug efflux systems, alteration of the drug target protein by random, chromosomal point mutations (Sebastian et al., 2017), and altering the bacterial cell wall. In chapter-2, we briefly discussed the intrinsic and acquired drug resistance mechanisms. The presence of a thick lipid cell wall and the action of the efflux pump provide natural resistance to TB treatment.

In contrast, prolonged exposure to drugs helps in acquiring spontaneous chromosomal mutations in the first-line TB drug targets *katG* (Heym et al., 1995), *rpoB* (Telenti et al., 1993), *pncA* (Sreevatsan et al., 1997) and *emb* (Ramaswamy et al., 2000) targeted by INH, RIF, PZA and EMB, respectively. Thus, the main reason for drug resistance in TB is chromosomal mutations (mostly point mutations in the coding region of the target proteins) resulting in the alteration of drug target or prodrug target binding site, whole target modification, drug inactivation and imitation of the drug target by other mycobacterial proteins (Alcaide et al., 1997; Nguyen, 2016; Riccardi et al., 2009; Zhang & Yew, 2009). The mutations resulting in these acquired resistance mechanisms can be due to changes in function,

structure or stability, alteration in conservation of protein sequence or reduced drug binding affinity of first-line TB target proteins. da Cunha et al. (2007) tried to understand the change happening when INH binds to mutant *katG* S315T. They first performed docking and dynamic simulations and later conducted an *in vitro* study to understand the enzymatic activity of *katG*. They found a reduction in enzymatic activity of *katG* due to mutation at position S315T. Lakshmipathy et al. (2013) observed deviations in the interaction of PNZ and mutant *pncA* with an A102P mutation and reduced drug binding as compared to the wild-type *pncA*. Kumar and Jena (2014) performed a computational study to understand the impact of S450L, H445Y mutations on RIF through molecular docking and found that the mutations had impacted the binding of *rpoB* and RIF. Aggarwal et al. (2018) conducted a study to understand the role of W68R and W68G mutations in PNZ resistance. The researchers predicted that changes in structural stability caused by W68R and W68G reduced the drug affinity of PNZ. Singh and Mizrahi (2017) performed a dynamic molecular analysis of mutation G279D in catalase-peroxidase (*katG*). They found changes in structural stability, changes in the binding cavity and reduced affinity for INH with a lower docking score. Unissa et al. (2017) identified different types of mutations at position 315 of catalase-peroxidase (*katG*) showed reduced affinity towards INH. They used five mutants in their study: S315T, S315I, S315R, S315N and S315G. We found no relevant research performed for understanding ethambutol resistance.

Most of the studies discussed above were conducted to determine changes in drug binding affinity by constructing a 3D structure of the wild-type target protein and the mutant, performing molecular docking using different tools, and comparing the wild-type target protein's binding energy with the that of the mutant protein. However, the nature of the mutations and the specific changes they cause in first-line TB drug targets and how they afford drug resistance and their impact on *Mycobacterium tuberculosis* fitness is currently not well understood.

3.2.2 What is needed to know to improve our understanding of TB drug resistance?

Drug resistance is emerging at an alarming rate and requires further research to address this situation. The most urgent and crucial research need is to understand the underlying drug resistance mechanisms more deeply to help achieve the long-term goal of eradication of TB globally through a clearer understanding of the bacterial survival strategy. The most promising avenue for this is to analyse the first-line mutation data accumulated over recent decades to improve our knowledge of the underlying drug resistance mechanisms.

The first-line TB drugs were designed to target proteins that play a vital role in biological processes and functions, such as physiology, metabolism and translation of proteins within *Mycobacterium tuberculosis*. Thus, for a mutation to be drug resistant, it should reduce the binding affinity of a drug

without hampering the natural affinity for its substrate, thus, not jeopardising the normal functioning and structural stability of the target protein needed for the survival of the disease-causing organism.

Specific studies have shown that chromosomal mutations in *Mycobacterium tuberculosis* are frequently associated with a decrease in relative fitness, thereby affecting the growth, stability and development of resistant TB strains (Andersson & Hughes, 2010; Gagneux et al., 2006; Mariam et al., 2004). Therefore, *Mycobacterium tuberculosis* usually has compensatory mutations to counteract the reduction in fitness. *In-vitro* studies and mathematical models have shown that the fitness cost of a drug-resistant isolate could be compensated for by putative compensatory mechanisms that restore the fitness of the mutant strains (Reynolds, 2000). For example, the fitness cost of the RIF resistant mutation, S531L, in the *rpoB* gene is overcome by a compensatory mutation in the *rpoA* and *rpoC* genes (Comas et al., 2012) and the mutation leading to overexpression of *ahpC* gene compensates for the S315T *katG* gene mutation (Sherman et al., 1996).

Previous studies have assumed that only few mutations occur in *Mycobacterium tuberculosis* but our study has revealed the extensive repertoire of these mutations as will be shown in this chapter. Considering the extensive nature of mutations that *Mycobacterium tuberculosis* uses to make the drug ineffective and stay viable, an in-depth exploration of the impact of different mutations at different positions in the first-line TB drug targets of *Mycobacterium tuberculosis* is the key to better understanding the distinctive set of strategies used by a bacterium for its survival and the different defence mechanisms it uses against drugs. Specifically, a deeper understanding of the underlying changes occurring in drug target proteins due to mutations in *Mycobacterium tuberculosis* can provide crucial insights into drug resistance mechanisms. These changes in the target protein correspond to two aspects of the interplay between the drugs and *Mycobacterium tuberculosis* - the effect on the drug and on the survival of *Mycobacterium tuberculosis*. The impact on the drug can be broken down into various defence mechanisms against the drug: preventing the entry of a drug, inactivation of the drug, preventing drug binding by altering drug binding sites or modification of the target. However, changes in target proteins alter the structure and function of the target which, in turn, can affect the bacterial fitness that *Mycobacterium tuberculosis* needs to minimise through various compensatory mechanisms. Investigation into these opposing forces by analysing mutation data could provide a balanced view of drug resistance and the survival of *Mycobacterium tuberculosis*. In particular, understanding where the target protein mutations occur, whether in conserved or non-conserved regions, mutation frequency and the advantage and role of the mutations in these specific locations in maximising the effect on drug and minimising the impact on *Mycobacterium tuberculosis* can provide crucial insights into bacterial strategies leading to acquired resistance.

3.2.2.1 Hypothesis on drug resistance and bacterial fitness

To gain a deeper understanding of TB drug resistance due to mutations, we hypothesize that for its survival, *Mycobacterium tuberculosis* uses an arsenal of different strategies to disable drugs and its survivability. Strategies with the most significant disruption to the drug and the most negligible impact on its survival may be preferable. However, this may not always be realistic; therefore, *Mycobacterium tuberculosis* may instead use compromise strategies that offer it a relative survival advantage over the drugs. This study is designed to test this hypothesis and elucidate the most effective *Mycobacterium tuberculosis* strategies for drug resistance.

Central to unravelling drug resistance mechanisms is the nature and extent of the changes in the drug targets due to mutations. In the drug - *Mycobacterium tuberculosis* interplay, these changes in the target determine the extent of the disabling of drugs and the survival of *Mycobacterium tuberculosis*. Drug resistant mutations may cause modification of the target, leading to structural changes in the target leading to a reduced affinity for drugs to prevent the activation of prodrugs (INH and PZA) or irregular binding with drugs (EMB and RIF) leading to failure in disrupting the cell wall/membrane and RNA polymerase to kill *Mycobacterium tuberculosis*. These changes also affect the function and stability of the target and impact *Mycobacterium tuberculosis* survival. Therefore, drug resistance is an instance of bacterial evolution through natural selection in favour of bacterial survival. How these processes work are reflected in the thriving drug resistance mechanisms achieved through target modification.

The most effective changes in the target are those that produce non-synonymous amino acid substitutions (AAS). Hotspot sites within the target would help develop successful strategies by *Mycobacterium tuberculosis* for drug resistance. These are single positions substituted multiple times by different amino acids (non-synonymous AAS) and these mutations could be more prevalent at specific positions in the target. It is also crucial to find where in the drug target these changes are occurring. Mutations in the conserved regions would have a more significant effect than those in non-conserved regions as they can cause greater functional and stability changes in drug targets, that significantly impact drug binding. However, this can also more severely impact the survivability of *Mycobacterium tuberculosis*. Therefore, these mutations would be much less favourable than those in the variable (non-conserved) regions. Another essential consideration would be the specific region of the mutations, as mutation in some regions can weaken the drug more effectively. Mutations in the drug binding region (DBR) could be favoured more as they can weaken the drug target more; however, those outside this region could also play a supporting role or compensatory role to improve the survival of *Mycobacterium tuberculosis*. To investigate this, we divide the target into three sites: site-1, representing residues in the drug binding site, site-2, having residues directly interacting with the

binding site and, site-3, representing the rest of the target protein. Mutations in site-1 have a more significant impact on drug binding compared to sites-2 and -3. These are expected to be the most favoured by *Mycobacterium tuberculosis*. All these possibilities suggest that *Mycobacterium tuberculosis* has at its disposal an arsenal of strategies to help them strike a favourable balance between weakening the drugs and improving its survival that needs to be elucidated (Zhang & Yew, 2009).

The strategies described above have different effects on both *Mycobacterium tuberculosis* and the drugs. As said previously, a strategy that has the least impact on bacteria and the highest impact on a drug is the best from a bacterial point of view; however, a compromise solution may be most pragmatic for bacterial survivability. *Mycobacterium tuberculosis* may seek a compromise solution by largely favouring mutations with mild to moderate impact on fitness and fewer mutations with neutral fitness effect and even fewer mutations that are lethal to its survival. Therefore, in this study, we aim to uncover the specific strategies used by *Mycobacterium tuberculosis* from the type of mutations (conserved or not), frequency and location of mutations and their relative impact on the structure and function of the target and drug binding from the mutational data found over the last 30 years. We then rank the mutations in terms of their impact on drugs and *Mycobacterium tuberculosis* into lethal, moderate, mild and neutral. Lethal mutations could be considered deleterious to *Mycobacterium tuberculosis* and a drug, and mild to moderate mutations as low-cost mutations favouring the survival of *Mycobacterium tuberculosis* while enabling drug resistance, and neutral as cost-free or beneficial mutations with the potential restoration of fitness by compensatory mutations. From these arsenals of strategies that variously impact drug binding and *Mycobacterium tuberculosis* survival, we could expect a higher occurrence of mild-moderate, much fewer lethal mutations and even fewer neutral mutations. Our study also explores the global and regional character of drug resistance towards developing strategies for the worldwide eradication of TB.

3.2.3 Framework developed for studying drug resistant mechanisms

Conducting laboratory or *in-vitro* studies for determining the effect of different mutations on *Mycobacterium tuberculosis* would be time-consuming and expensive. Our study provides a potential solution to reduce the burden of the emerging drug resistance problem globally. With this aim, we perform a holistic research study to gain deeper insights into drug resistance mechanisms in TB based on a comprehensive bioinformatics study. We investigate the mutation statistics and the impact of different mutations on the survival of TB and drug weakening and then rank them accordingly to uncover the totality of the *Mycobacterium tuberculosis* strategy for drug resistance. This can be used as the foundation for either improving first-line drugs or developing new drug therapeutics or treatment protocols. We also aim to create an atlas of drug resistant-mutations in the targets of first-

line TB drugs for future reference. A reliable catalogue of drug-resistant mutations can be used as a reference standard for validating the mutations identified in the genome of new drug-resistant TB strains when available. The survival strategies unravelled in this study can also be the foundation for continuing the exploration of the evolution of drug resistance into the future.

Accordingly, this study is focused on understanding the nature of drug resistance through an in-depth analysis of drug resistant mutations. We design a complete methodological pipeline to make the discoveries mentioned above. The main advantage of the method developed is that it can be used to predict the impact of any newly discovered mutation. The procedure can also be used for mutational studies in any other organism (not restricted to *Mycobacterium tuberculosis*).

The proposed method involves four phases: collecting mutational data (phase 1) and screening non-synonymous mutations in the drug targets (phase 2). In phase 3, mutations are analysed in two stages: (i) mutational statistics, and (ii) the impact of mutations on the targets. These impacts will then be used in Phase 4 to categorise mutations into lethal, moderate, mild and neutral to evaluate the total drug resistance strategy of *Mycobacterium tuberculosis*. Figure 3.1 shows the workflow for the comprehensive analysis of the impact of TB drug resistant mutations on target function and stability, and drug binding, using various bioinformatics tools. The details of the methods used in the workflow can be found in the Materials and Methods Section.

Following these phases, this study analyses 821 mutations found in *katG*, *pncA*, *rpoB* and *embCAB* of drug resistant TB isolates (approximately 31,073 TB isolates). These 821 mutations were found in this study from an extensive survey of 1489 studies, subsequently reduced to 149 studies that provided the relevant mutational data reported over three decades as discussed in the Methods and Results sections of this chapter. Then we studied mutation statistics covering the prevalence, i.e., mutation frequency, location of all single mutations and hot spot sites in drug resistant TB isolates. We also explored the character of the global (WHO regions) spread of mutations to assess regional variations in mutations and drug resistance. Then we studied the impact of a mutation on structural stability, functionality, target conservation and the altered binding energy of drugs using several bioinformatics tools (shown in capital letters in Fig. 3.1). Furthermore, each of these studies uses up to three different bioinformatics tools for comparison, validation or consensus.

In the analysis of impact of mutation on the drug and target, the first step involves building and validating the protein structure of the wild-type and each mutant-type first-line drug target. In the second step, the identification of evolutionarily conserved amino acid residues in the target proteins and protein domain prediction is undertaken. In the third step, mutated residues are distributed into

three sites: sites -1, -2 and -3. In the fourth step, each of these mutations is analysed for the impact of the mutation on *Mycobacterium tuberculosis* and drugs in terms of functional changes and structural stability, and its drug binding affinity based on molecular docking. In the final step, these impacts are used for a consensus prediction, using several computational methods, of lethal, moderate and neutral mutations in first-line TB drug target proteins, as metrics *Mycobacterium tuberculosis* could use to develop its survival strategy. These results will be synthesised to formulate the holistic drug resistance strategy evolved by *Mycobacterium tuberculosis*. Furthermore, analysis of each mutation in terms of the survivability of *Mycobacterium tuberculosis,* and the impact on each drug will help explore specific drug resistance strategies that would provide deep insights into some of the features that had not been noticed in the limited studies of the past. The new knowledge gained in this research can help develop new strategies or tools to combat drug-resistant TB, including identifying new drugs, designing effective inhibitors to deal with drug-resistant strains, and developing personalized treatments and diagnostics techniques.
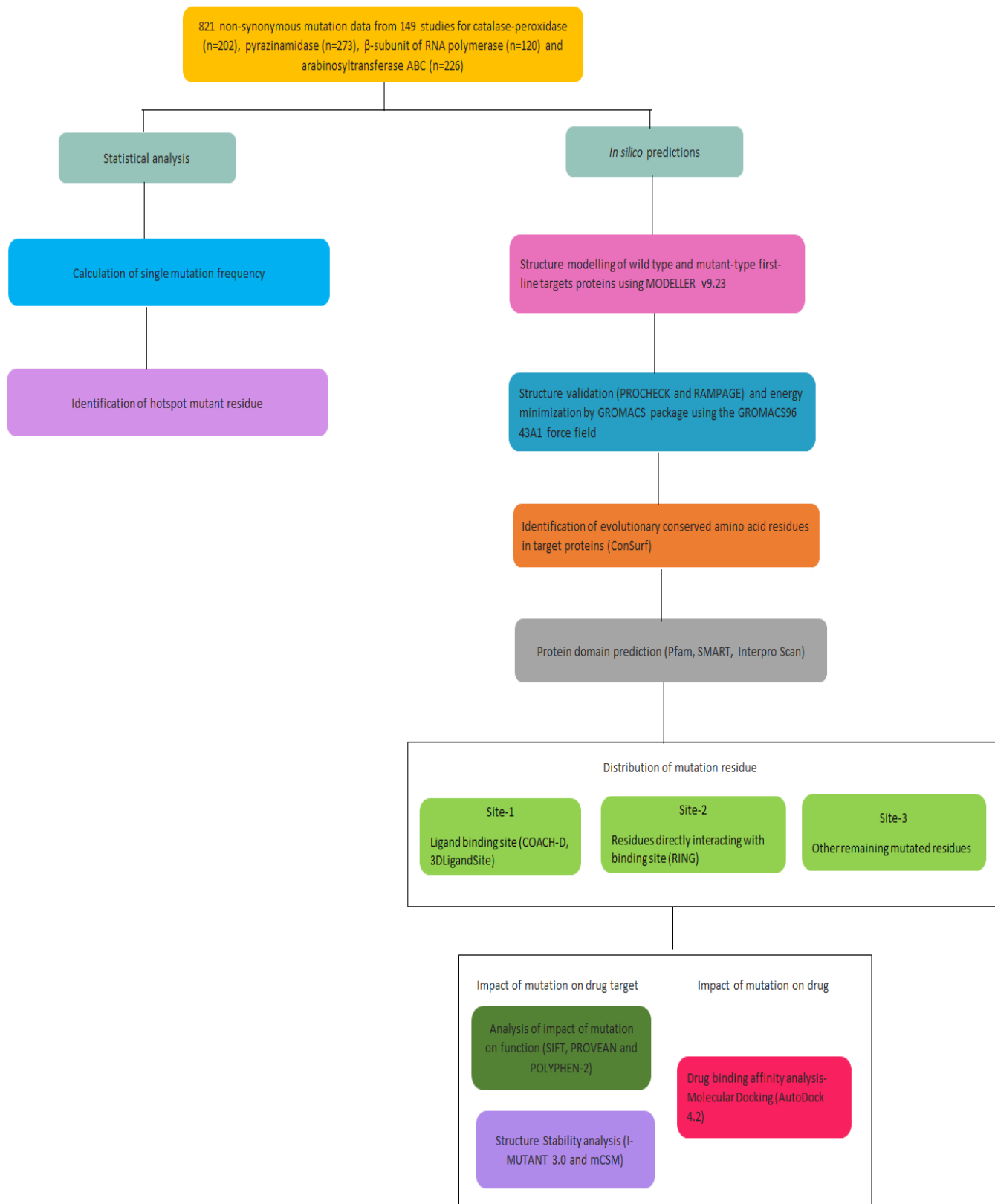
Figure 3. 1: Comprehensive analysis of mutations to predict their impact on first-line TB drug target proteins and first-line drugs (821 non-synonymous mutations were found in this study from 149 publications covering mutations in the last 30 years as discussed in Results section)

3.3   Materials and Methods

Resistance against tuberculosis drugs is increasing at an alarming rate. Thus, there is a need for understanding the potential impact of mutations in first-line drug target proteins as these mutations play a significant role in drug resistance. Mutation cause changes in the protein structure, inhibits drug activity, changes drug binding sites and inhibits drug binding or enables a different resistance level depending on the specific mutation and changes in target protein function, leading to non-activation of a drug. Mutation can either be located in the coding or non-coding region of the genome. Most of the mutations in *Mycobacterium tuberculosis* are in the coding regions of the genome. These mutations can be further divided into synonymous (no change in the amino acid at mutation position) and non-synonymous mutations (substituting one amino acid for another at the mutation position) depending on whether they lead to changes in the corresponding amino acid sequence. Non-synonymous mutations in the coding region have a substantial impact on *Mycobacterium tuberculosis*. This research is focused on creating a catalogue of the first-line drug-resistant non-synonymous mutations in *katG*, *rpoB*, *pncA* and *embCAB* and identifying the most impactful and lethal mutations from them. Study will investigate mutational impact further to unravel how bacteria use the mutations to resist drugs and form their survival strategy.

3.3.1   Data Collection

3.3.1.1 Literature search

A search was performed for identifying research published (relating to tuberculosis drug resistance) in PubMed, Scopus, Google Scholar and online archives of the International Journal of Tuberculosis and Lung Disease. The search was undertaken using the following keywords individually and in exhaustive/extensive combinations by applying the AND operator: 'tuberculosis,' 'incidence,' '*Mycobacterium tuberculosis*,' 'tuberculosis patients,' 'first-line TB drugs,' '*katG*,' '*rpoB*,' '*pncA*,' '*emb*,' 'isoniazid,' 'rifampicin,' 'pyrazinamide,' 'ethambutol,' 'prevalence of drug-resistant tuberculosis,' 'drug resistant first-line TB drugs,' 'drug-resistant tuberculosis,' 'first-line drug resistant tuberculosis,' 'multidrug-resistant tuberculosis,' 'MDR-TB,' 'isoniazid resistance,' 'rifampicin resistance,' 'pyrazinamide resistance' and 'ethambutol resistance.' We did not exclude any publications based on their published date.

3.3.1.2 Study inclusion and exclusion criteria

Publications were selected based on the following criteria:

a) published original data,

b) analysis performed on strains of *Mycobacterium tuberculosis* obtained from a clinical sample,

c) phenotypic drug susceptibility testing (DST) was done,

d) characterizing mutations by sequencing (either DNA or pyrosequencing),

e) information available on individual amino acid mutation.

Publication's exclusion criteria:

a) laboratory *Mycobacterium tuberculosis* strains used in the study,

b) phenotypic DST not performed by the author,

c) if sequencing was not done as a method for determining drug-resistant mutations,

d) if individual amino acid mutation data not included in the study,

e) if the publication contained inconsistent mutation data,

f) publications in different languages other than English.

Duplicate publications and studies that did not meet inclusion criteria were eliminated to ensure the accuracy of the data collected.

3.3.1.3 Quality assessment

Mycobacterium tuberculosis H37Rv (GenBank accession number NC_000962.2) was taken as a reference genome to control errors in the data selected. Every mutated amino acid reported was compared to its reference H37Rv sequence. Mutations inconsistent with the reference amino acids were excluded from the analysis, while those with the correct amino acid reference were included in the study.

3.3.2 Creating an atlas of non-synonymous mutations in first-line target genes

An atlas/catalogue for non-synonymous first-line TB drug resistant mutations could prove a starting point for designing better molecular diagnostic tests and serve as reference data for studying drug resistance in tuberculosis.

3.3.2.1 Data extraction

The following information was extracted and compiled in an excel file using Microsoft Excel software:

- PubMed ID
- Author name
- Sample collection year
- Publication year

- Geographic location of sample collection
- Total number of isolates used in each study
- Total number of resistant strains identified in each study
- Total number of drug target (*katG*, *rpoB*, *pncA* and *embCAB*) mutations in each publication

3.3.2.2 Analysis of the prevalence of first-line drug target mutations by WHO-region

The data collected were further stratified into the six WHO-defined regions (Africa, Americas, Eastern Mediterranean, Europe, South-East Asia and the Western Pacific) based on the sample's origin. Finally, the relative frequency of mutations in first-line TB drug target genes was calculated for each region.

3.3.2.3 Data collation and calculation of single mutation frequency

Data showing resistant non-synonymous mutations in *katG*, *rpoB*, *pncA* and *embCAB* genes were collected from different publications according to the above guidelines. The data collected were filtered based on the inclusion criteria and mutations associated with resistance to INH, RIF, PZA and EMB were grouped according to their target genes (*katG*, *rpoB*, *pncA* and *embCAB*, respectively). This arrangement of data was undertaken to identify the global distribution of drug resistant mutations in first-line TB drug targets. The total number of specific amino acid mutations was calculated by combining the number of incidences of a particular amino acid mutation in all publications. Single mutation frequency was calculated using the following formula (Georghiou et al., 2012)

$$\text{frequency \% of single mutation} = \frac{\text{Total number of specific amino acid mutations in a particular drug target} * 100}{\text{Total number of resistant isolates for a particular drug target}}$$

3.3.2.4 Identification of mutations in the hotspot region and hotspot residue site

The next step was to calculate the frequency of hotspot mutations. The frequency of the different types of mutations (substituting one amino acid for other) at each position in the target gene was calculated. For example, amino acid S present at position 315 of a drug target can be substituted by how many different amino acids? This frequency analysis helps identify hotspot residue sites and provides information on the prevalence of mutations at a specific site or region in a gene.

3.3.3 Structure modelling and validation of wild-type and mutant first-line TB drug target proteins

3.3.3.1 Homology modelling

3D structures of wild-type target proteins, catalase-peroxidase, pyrazinamidase, arabinosyl transferase, and DNA-directed RNA polymerase subunit beta of *Mycobacterium tuberculosis,* were

constructed using MODELLER v9.23 (Eswar et al., 2008; Webb & Sali, 2016). Figure 3.2 describes the steps used in creating a model structure for drug target proteins of *Mycobacterium tuberculosis*.
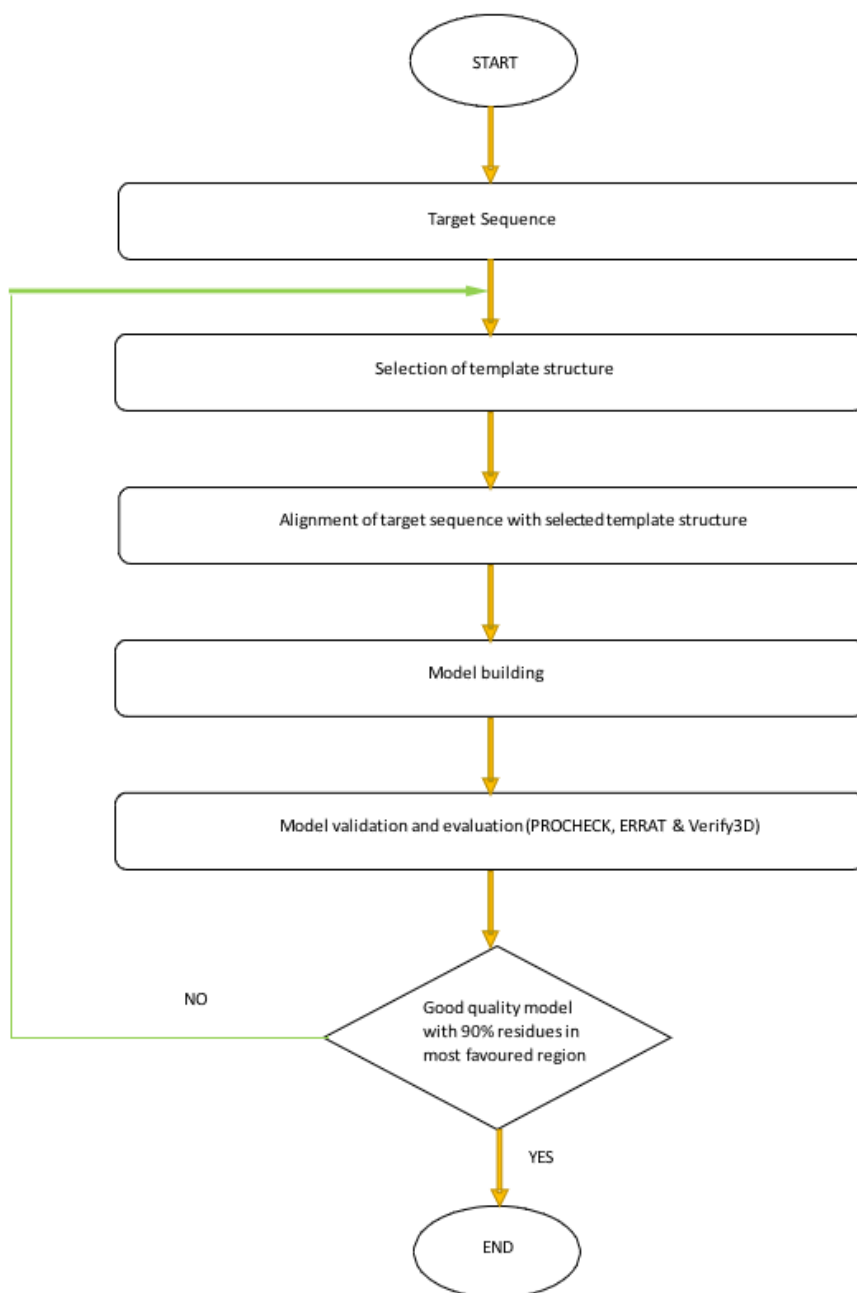
Figure 3. 2: Step-by-step workflow of MODELLER v9.23 used for the generation of a 3D structural model for wild-type and mutant-type protein

(1) **Target sequence:** The target protein sequence retrieved from the UniProt database to construct structural models.

(2) **Template selection:** For selecting a suitable template, PSI-BLAST was performed by taking the protein databank (PDB) database as the source database. A template is a protein structure available in the PDB database and used for constructing structural model of a query protein sequence. The template was chosen with a maximum identity percentage and query coverage.

The identity percentage describes how similar the query sequence is to the template. The higher identity percentage shows a significant alignment between a query sequence and the template. Query coverage is the percentage of the query sequence length included in the alignment with the template. The higher the query coverage score, the better the alignment.

(3) **Target sequence and template structure alignment:** To align the target sequence with the template structure, the "align 2d()" command in MODELLER v9.23 was used. The alignment algorithm used by MODELLER v9.23 differs from other sequence-sequence alignment algorithms, as it uses structural information from a template for constructing the alignment. This was achieved using the gap penalty function to reduce errors during the alignment process.

(4) **Model building:** After aligning the target sequence and template structure, the "automodel" class of MODELLER v9.23 was used for generating a structural model. The modelling module was instructed to produce ten models.

(5) **Selection of model**: The model building results were stored in a log file, which contained the molpdf, DOPE and GA341 scores. The 3D structural models with the lowest DOPE score and highest GA341 score were selected for structural validation.

3.3.3.2 Structure validation

After constructing the 3D structure of the wild-type and mutant protein structures, it is essential to check the structural validity. In our study, we used PROCHECK, ERRAT and Verify 3D for structure validation. PROCHECK is a suite of programs used for checking the stereochemical quality of the protein modelled. It examines the overall structural geometry and analyses residues by residue geometry. It provides detailed information about the chi angles, chirality, planarity of the peptide bonds, disulphide bonds, non-bonded interaction, the main chain hydrogen bonds and the backbone dihedral angles ψ against ϕ by Ramachandran plot (Laskowski et al., 1993). ERRAT is a structure validation algorithm that helps in checking the reliability of the protein model. ERRAT is specifically designed for evaluating the progress of crystallographic model building and refinement. In addition, ERRAT analyses the statistics of the non-bonded interactions between different atom types (Bowie et al., 1991). Verify 3D works by determining the compatibility of the 3D atomic model with its amino acid sequence by assigning a structural class based on location and the environment (alpha-helix, beta-sheet, loops) (Colovos & Yeates, 1993). All three servers help in selecting the best-modelled protein structure. If the model had more than 90% residues in the most favoured region, it was considered the best quality model and used for further analysis (Figure 3.2).

3.3.3.3 Energy minimization

After selecting the model protein for first-line wild-type and mutant proteins, it is essential to perform energy minimization to reduce unfavourable bond lengths, bond angles, distorted geometry or non-bonded interactions. The main goal of energy minimization is to obtain a structure corresponding to its natural geometry. We used the GROMACS 4 (Groningen Machine for Chemical Simulation) package (Hess et al., 2008) for energy minimization of the wild-type and mutant first-line drug target proteins. GROMACS works in a LINUX environment. The GROMACS96 43A1 force field was used in the energy minimization process. Table 3.1 describes the energy minimization steps and commands used to perform each step in GROMACS 4. Lastly, Chimera was used for protein visualization (Pettersen et al., 2004).

Table 3. 1: Energy minimization steps performed in GROMACS 4

Energy minimization steps	Command used in GROMACS 4
(1) Create a gromacs topology file from protein's PDB file	*pdb2gmx*
(2) Define a box around molecule	*editconf*
(3) Add solvent molecules to the box	*genbox*
(4) Prepare file for mdrun	*grompp*
(5) Final step for energy minimization	*mdrun*

3.3.4 Comprehensive computational analysis for identifying the impact of non-synonymous drug resistant mutations on *Mycobacterium tuberculosis*

In this study, we conducted a detailed bioinformatics analysis for understanding mutational changes at the molecular level affecting the function, structural stability, protein folding, physio-chemical properties, protein sequence conservation and drug binding affinity in the target drug protein (Figure 3.1).

3.3.4.1 Identification of conserved amino acid residues

ConSurf is a bioinformatics tool used for the estimation of the conserved amino acid residues in a protein. It performs multiple sequence alignments between the input protein sequences and homologous sequences and then constructs a phylogenetic tree to determine the evolutionary relations between them (Ashkenazy et al., 2016). It grades conservation scores on a scale of 1-9 where 1-3 are variable, 4-6 are average and 7-9 represent highly conserved amino acid residues in a protein (Celniker et al., 2013). Here, we grouped different mycobacterial species to determine the evolutionary relations between them using ConSurf.

3.3.4.2 Determining the occurrence of mutations in a protein domain

A protein domain is a conserved part of a protein that can independently exist, fold and function separately from the rest of the protein structure (Richardson, 1981; Wetlaufer, 1973). Some proteins have a single domain, while others have multiple domains. The presence of mutations in a domain region may impact its biological function or protein folding.

Pfam, Interpro Scan and SMART are freely accessible databases online. These three were used to predict protein domains in the first-line target proteins of *Mycobacterium tuberculosis* (catalase-peroxidase, pyrazinamidase, arabinosyl transferase and DNA-directed RNA polymerase subunit beta). Pfam is a curated database of protein families constructed using the profile hidden Markov model (HMM) (Finn et al., 2014). Interpro Scan classifies a given protein sequence into protein families and then predicts the presence of a functionally important protein domain (Jones et al., 2014). Interpro is based on the signature recognition method and is constructed by combining signature sequences from multiple databases into a single searchable resource (Mitchell et al., 2019). The simple modular architecture research tool (SMART) uses a SMART database to predict and annotate protein domains. The database manually integrates curated HMM for several domains (Letunic & Bork, 2018). After identification of the protein domains, we assessed the presence of our drug-resistant non-synonymous mutations in the protein domains.

3.3.4.3 Identification of essential sites in a protein

In this step, the mutated residues were distributed into three sites: site-1 is the ligand-binding site, site-2 contains the residues directly interacting with the binding site, and site-3 contains other mutated residues.

The structure of a protein is complex and consists of ample surface pockets, cavities and cross channels. These topographic characteristics are the structural basis of a protein and help it perform significant tasks, such as ligand binding, DNA or protein interactions and enzymatic activity. The presence of a mutation in a ligand-binding site interferes with the binding of the first-line drugs to their respective drug targets. A mutation can impact ligand binding site residues and the neighbouring region of a protein's ligand site, affecting the drug, substrate or metal binding. Prediction of the binding cavity helps in the understanding of protein-ligand interactions. Also, the amino acid residues having a direct connection with residues present in ligand-binding sites play a crucial role in the correct functioning and stability of a protein and mutations present in these contact residues can have a destabilizing effect on protein structure.

For identifying the binding sites (site-1) for catalase-peroxidase (*katG*), pyrazinamidase (*pncA*), DNA-directed RNA polymerase subunit beta (*rpoB*) and arabinosyl transferase C, A and B (*embCAB*) proteins

of *Mycobacterium tuberculosis*, COACH-D and the 3D LigandSite web server were used. COACH-D predicts binding for a given protein sequence or structure using five methods (Wu et al., 2018): ConCavity (Capra et al., 2009), TM-SITE (Yang et al., 2013), FINDSITE (Brylinski & Skolnick, 2008), COFACTOR (Roy et al., 2012) and S-SITE (Yang et al., 2013). After submitting the protein structure in COACH-D, the above-mentioned five different methods will predict protein-ligand binding sites. 3D LigandSite uses MAMMOTH (Ortiz et al., 2009) to perform structural alignment to identify similar structures with bound ligands to identify ligand binding sites in the query protein (Wass et al., 2010). The Ring webserver was used to predict the site-2 residues. The webserver produces a residue interaction network (Piovesan et al., 2016) that helps determine intra- and inter-chain contacts among amino acid residues of a protein.

3.3.4.4 Assessment of the functional impact of mutations

The mutations present in a protein can have substantial effects on function by either the loss or gain of function. Some mutations do not have any impact on the role of a protein. However, the position and property of some mutating amino acid residues can have a significant effect on the functioning of a protein. To determine the phenotypic impact of drug resistant mutations of *Mycobacterium tuberculosis*, SIFT, PROVEAN and POLYPHEN-2 computational tools were used in the study to predict deleterious mutations, mutations with intermediate effect and benign mutations.

SIFT (**s**orting **i**ntolerant **f**rom **t**olerant) is a sequence homology-based tool to predict intolerant and tolerant mutations (Ng & Henikoff, 2001). The FASTA sequence of target proteins is submitted to each computational tool. When a query protein sequence is submitted, SIFT performs PSI-BLAST (Altschul & Koonin, 1998) to align all functionally relevant proteins and provides a tolerance index for the mutation based on the conservation score obtained from the sequence alignment. The tolerance index ranges from 0-1, where scores ≥0.05 are "tolerant" and scores ≤ 0.05 are considered "intolerant" or "deleterious" (Kumar et al., 2009). PROVEAN is also a sequence homology-based tool that performs sequence alignment using BLAST (**b**asic **l**ocal **a**lignment **s**earch **t**ool) (Altschul et al., 1990). The alignment hits with more than 75% sequence identity are clustered together to generate a PROVEAN score (Choi et al., 2012). If the PROVEAN score is ≥ -2.5, the mutation has a "neutral" effect, while if the score is ≤ -2.5, the mutation has a "deleterious" or "harmful" effect on a protein (Choi & Chan, 2015). POLYPHEN-2 predicts the impact of a mutation on protein structure and function using specific considerations (structure and evolution). POLYPHEN-2 generates the PSIC (position-specific indecent count) score by performing BLAST against protein structures similar to query protein in the PDB database. The outcome for query protein can be "probably damaging," "possibly damaging," or "benign" mutation (Adzhubei et al., 2013; Ramensky, 2002).

3.3.4.5 Prediction of structural stability change

Non-synonymous amino acid substitution occurring in a protein may have a significant impact on the protein's structure. The mutations can alter structural stability, protein folding, solvent accessibility, and the structural motifs involved in critical molecular mechanisms. Mutations in amino acid residues present on the surface affect the protein binding with its interacting molecules (protein, DNA or ligand). Amino acid substitutions occurring in buried residues have a high probability of impacting protein folding or causing the unfolding of the protein, thereby affecting stability. Thus, the presence of mutations in the buried residues first-line protein targets (catalase-peroxidase, pyrazinamidase, arabinosyl transferase and DNA-directed RNA polymerase subunit beta) of *Mycobacterium tuberculosis* may cause destabilization of the protein.

I-MUTANT 3.0 and mCSM were used in this study for analysing the mutational impact on protein stability. I-MUTANT 3.0 uses a support vector machine (SVM) algorithm to predict changes in the stability of a protein caused by a single non-synonymous mutation. The FASTA sequence of a query protein is submitted along with information on mutation and changes in the amino acid residue. The output is classified into three categories: "large increase" (DDG >0.5kcal/mol), "neutral" (-0.5 ≤ DDG ≤ 0.5 kcal/mol) or "large decrease" (DDG≤ -0.5kcal/mol) (Capriotti et al., 2005) in stability. mCSM is a machine learning tool used to determine the effects of mutations using graph-based structural signatures. The signatures of mCSM were obtained from the Cutoff Scanning Matrix (CSM), a graph-based concept used to study biological systems to represent the distinct pattern of network topology (Pires et al., 2016). The output of mCSM provides information on the change in stability as either "destabilizing" or "stabilizing".

3.3.4.6 Drug binding affinity analysis

The drug binding energy score predicts the strength of interaction between ligand and protein. Mutations occurring in drug target proteins can have a potential impact on the binding affinity of a drug. Sometimes this may not allow the drug to bind to its target protein. Our study performed a docking procedure to determine the impact on drug binding of a mutation in catalase-peroxidase, pyrazinamidase, arabinosyl transferase, and DNA-directed RNA polymerase subunit beta proteins of *Mycobacterium tuberculosis.*

Molecular docking is a computational simulation that involves the interaction between a ligand and a receptor to give a stable conformation (Ferreira et al., 2015; Shoichet et al., 2002). Our study's purpose is to determine the predominant drug binding affinity of wild-type and mutant proteins with their respective first-line drugs. The AutoDock 4.2 docking tool, was used for determining the binding orientation and binding affinity between a drug and its respective drug target (Morris et al., 1998). AutoDock suite consists of two programs: autogrid for pre-calculating the grids and AutoDock program

for performing docking of the ligand to protein with pre-calculated grids (Morris et al., 2009). The following steps were performed for molecular docking using AutoDock 4.2:

(1) **Preparation of coordinate file**: In the docking process, *Mycobacterium tuberculosis* proteins were considered rigid and their binding ligands were deemed flexible (Rauf et al., 2015). The ligands INH, PZA, EMB and RIF for catalase-peroxidase, pyrazinamidase, arabinosyl transferase and DNA-directed RNA polymerase subunit beta proteins were retrieved from the PubChem database (Wang et al., 2009) in the mol2 format and then converted into the PDB format using PyMOL molecular graphics system. The protein and ligand files should be converted into the PDBQT format and then used for docking.

(2) **Generating a grid box**: Autodock4 used a grid-based approach to approximate the energy calculations. The docking area was defined by 100 × 100 × 100 points with a grid spacing of 0.375 Å. The grid box was centred on the macromolecule and covered almost the whole ligand-binding site of the target protein.

(3) **Docking**: Lamarckian genetic algorithms implemented with a 200-docking population size and two million energy evaluations, were used for docking experiments (Morris et al., 1998). Twenty docking poses were generated for each docking process.

(4) **Analysis**: The docking output clusters were analysed for determining the best conforming one. The best output can be selected as that possessing strong binding affinity and lowest binding energy. Furthermore, the binding energy of the mutant proteins was compared with their corresponding wild-type protein to determine the effect of mutation on binding affinity.

3.4   Results

3.4.1   Selection of drug-resistant mutation studies and data extraction

Here we provide the details of the extraction of mutations for this study. In the initial extensive search, a total of 1511 studies were selected. Out of the 1511 studies, 1486 studies were retrieved through electronic searches in PubMed, Scopus, Google Scholar and online archives of the International Journal of Tuberculosis and Lung Disease using keywords described in the Material and Methods section. An additional 25 research papers were identified by looking at the reference lists of the 1486 studies. The removal of 252 duplicate papers left 1259 studies for further review. After completing a thorough search based on the inclusion and exclusion criteria shown in Figure 3.3, 149 studies were selected for TB drug-resistant mutation study.

Identification

1486 studies identified using different keywords related to drug resistant mutations in first-line TB drugs in PubMed, Scopus, Google Scholar and online archives of International Journal of Tuberculosis and Lung Disease.

25 relevant published studies identified in the reference list of the publications searched

252 duplicates studies removed

Screening

1259 publications for further screening after removal of duplicates

807 relevant studies screened for first-line TB drug resistance
(545 publications not related to TB drug resistance excluded after reading title and abstract)

Eligibility

262 full texts assessed for first-line TB drug resistance

113 publications removed based on inclusion and exclusion criteria

- laboratory Mtb strains used in the study (51)
- phenotypic drug-susceptibility testing not performed (15)
- sequencing was not performed (14)
- individual mutations not studied (27)
- publications in different languages other than English (6)

Included

149 studies included for assessing mutations in first-line TB drugs

Figure 3. 3: **The process of selection of studies based on keyword search and eligibility criteria. The initial investigation was done using keywords individually and combination of 'AND,' 'OR' operators: 'tuberculosis,' 'incidence,' *Mycobacterium tuberculosis,'* 'tuberculosis patients,' 'first-line TB drugs,' *'katG,'* *'rpoB,'* *'pncA,'* *'emb,'* 'isoniazid,' 'rifampicin,' 'pyrazinamide,' 'ethambutol,' 'prevalence of drug-resistant tuberculosis,' 'drug resistant first-line TB drugs,' 'drug-resistant tuberculosis,' 'first-line drug resistant tuberculosis,' 'multidrug-resistant tuberculosis,' 'MDR-TB,' 'isoniazid resistance,' 'rifampicin resistance,' 'pyrazinamide resistance' and 'ethambutol resistance'**

The data collected from 149 studies included the PubMed ID of the study, name of the author, publication year, year of sample collection from TB patients, geographic location of the collected sample, first-line TB drug studied, total number of clinical isolates used in each study and the total number of resistant strains identified for the first-line TB drug targets (*katG*, *rpoB*, *pncA* and *embCAB*) in each study. The Appendix Table A.1 provides detailed information on the 149 selected studies. In addition, Table 3.2 shows the number of research studies carried out for each and in combination for first-line TB drugs in the 149 selected studies.

Table 3. 2: The different first-line TB drugs studied in the selected 149 studies. INH-isoniazid, RIF-rifampicin, PZA-pyrazinamide and EMB-ethambutol

First-line TB drugs studied in the selected studies	Number of studies
INH	24
RIF	20
PZA	25
EMB	19
INH and RIF	36
INH and EMB	1
RIF and PZA	1
EMB and PZA	1
INH, RIF and EMB	7
INH, RIF and PZA	2
INH, RIF, PZA and EMB	13

The timeframe of sample collection ranges from 1988 to 2018 (the final year of clinical isolate collection) spanning 30 years. From the 149 studies, 33 did not provide the details of the sample collection year. In terms of publication year, 88 studies were published between 2010 and 2019. From 94,687 isolates collected over 30 years from patients diagnosed with TB, 31,073 isolates (from seven to 2081 *Mycobacterium tuberculosis* isolates per study) were used for studying resistance to different first-line TB drugs. Out of 31,073 isolates, mutation data for *Mycobacterium tuberculosis* isolates showing resistance to the four first-line TB drugs, isoniazid (n= 5703), rifampicin (n=5282), pyrazinamide (n=2608) and ethambutol (n=1771), were available. After collection and cleaning, this 30-year mutation dataset formed the basis for the study of the strategy of *Mycobacterium tuberculosis* for drug resistance.

A total of 12,616 non-synonymous mutations (p) identified in drug-resistant *Mycobacterium tuberculosis* isolates (determined using *Mycobacterium tuberculosis* H37Rv as the reference genome) conferred mutations on first-line TB drug targets in *katG* (p=4589), *pncA* (p=1227), *rpoB* (p=5191) and *embCAB* (p=1609). *katG* and *rpoB,* targets of the two most crucial TB drugs INH and RIF, respectively, showed a high mutation count compared to *pncA* and *embCAB*. INH and RIF are used in both the intensive and continuation phases of the DOTS strategy. The latter two target drugs, PZA and EMB, are used in the two- to four-month intensive phase of TB treatment. The prolonged exposure of INH and RIF, for six to nine months, during TB treatment could be a reason for the higher mutation counts in *katG* and *rpoB*. This suggests that prolonged exposure to TB drugs can allow *Mycobacterium tuberculosis* to evolve potentially stronger drug resistance strategies.

3.4.2   Global distribution of first-line TB drug resistant mutations

Accurate geographic locations were reported in 139 research papers for 51 different countries. Of the 51 countries, 15 belonged to high TB burden countries. Some of the high TB burden countries in the selected studies were from Bangladesh, China, Ethiopia, India, Myanmar, Pakistan, Philippines, Russia, South Africa and Zambia. Geographically, 18 studies were conducted in China, 13 in India, 10 in U.S.A and seven each in South Africa and South Korea.

For understanding the global prevalence of drug resistant mutations in tuberculosis, computation of the number of mutations present for a drug in a specific country was needed. Unfortunately, country-level mutation data were very sparse, leading to the aggregation of mutational data into six WHO regions. The 51 countries covered the six WHO regions: Africa- 6, Americas- 10, Eastern Mediterranean- 5, Europe- 15, South-East Asia- 5, Western Pacific- 10. Figure 3.4 shows first-line TB drug resistant mutations found in six WHO regions and Table 3.3 shows that the occurrence of drug resistant mutations varies among the WHO regions.
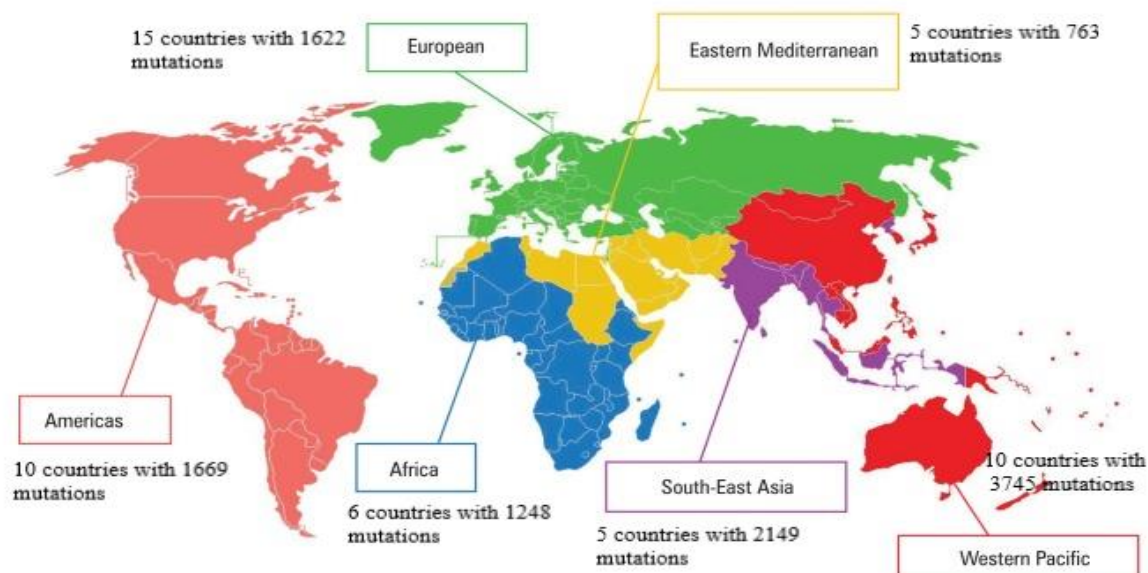
Figure 3. 4: Geographic distribution of first-line TB drug resistant mutations in the six WHO regions

As shown in Figure 3.4, mutations occurred most frequently in the Western Pacific (29.68%), followed by South-East Asia (17.03%), the Americas (13.23%), Europe (12.86%), Africa (9.89%) and Eastern Mediterranean (6.05%). Mutations of unknown origin accounted for 11.26% of the total. Mutations in all four drug targets (*katG, rpoB, pncA* and *emb*) were found most frequently in the Western Pacific region and least frequently in Africa (*katG* and *emb*) and the Eastern Mediterranean (*rpoB* and *pncA*) (Table 3.3). At the WHO regional level, the mutation frequencies of *katG* and *rpoB* were considerably higher than that of *pncA* and *embCAB* in all six WHO regions. The differences in mutation count could be due to several reasons: weather conditions, the health of the patient, prolonged exposure to drugs, and poor performance of some sequence-based diagnostic testing. Inconsistencies in the identification of drug resistance could lead to incorrect TB treatment plans posing a threat to drug-resistant TB patients' lives. Thus, it is essential to develop better diagnostic methods and perform whole genome sequencing for predicting drug resistance in first-line TB drugs.

Table 3. 3: Number of drug resistant mutations in first-line drug targets in the six WHO regions. Fifty-one countries were divided into six WHO regions (from 139 publications) and unknown regions (10 studies did not specify exact location)

Drugs/ WHO regions	Africa	Americas	Eastern Mediterranean	Europe	South-East Asia	Western Pacific	Unknown
katG	308	919	349	436	973	1075	529
rpoB	849	490	317	676	954	1455	450
pncA	71	146	2	244	69	332	363
embCAB	20	114	95	266	153	883	78

3.4.3   Single mutation frequency of the first-line TB drug targets

A total of 821 non-synonymous mutations (m) occurring in the coding region of *katG* (m=202), *rpoB* (m=120), *pncA* (m=273) and *embCAB* (m=226) were identified. The single mutation frequency for each mutation identified was calculated using the formula described in the Materials and Methods, section 3.3.2.3, to determine the prevalence of a mutation at a specific site. An atlas of non-synonymous mutations identified in the first-line TB drug targets of *Mycobacterium tuberculosis* was created describing their amino acid substitution, phenotypically resistant clinical isolates sequenced for a specific target or position, mutation count and single amino acid mutation frequencies (Appendix B). The previous section revealed that more mutations in *katG* and *rpoB*, compared to *pncA* and *embCAB*, suggest that prolonged exposure to a drug allows *Mycobacterium tuberculosis* to experiment with more mutation options. After calculating the single mutation frequency, it was observed that the mutation count varies significantly in different first-line TB targets and within each target. This suggests that *Mycobacterium tuberculosis* either purposefully or through random trial and error has found strategies for mutating the drug target by trying diversity of mutations.

(i) *katG* mutations associated with INH resistance

Isoniazid is used for inhibiting mycolic acid biosynthesis, thus weakening the *Mycobacterium tuberculosis* cell wall. This prodrug is activated by *katG.* A total of 202 point mutations were found at 140 different positions in the coding region of *katG.* The range in mutation counts varied from 1 to 3690 (Figure 3.5). The most frequent mutations were observed at the *katG* S315 position, with 3690 mutations in 5667 INH resistant isolates. *katG* R463 was the second most commonly mutated codon with 333 mutations in 2400 resistant isolates. Approximately 24 different positions in *katG* had two mutations and 69 had only one mutation.

The calculation of single mutation frequency revealed that *katG* S315T was the most prevalent amino acid substitution with a single mutation frequency of 60.57%, followed by R463L with 13.75% and S315N with 2.95%. The crystal structure of *katG* with 1SJ2 PDB entry (Bertrand et al., 2004) shows that S315 is the INH binding site. Therefore, a mutation at the INH binding site might bring conformational changes in *katG*, preventing activation of INH. The mutation pattern of *katG* suggests that *Mycobacterium tuberculosis* uses a strategic plan to safeguard itself against INH.
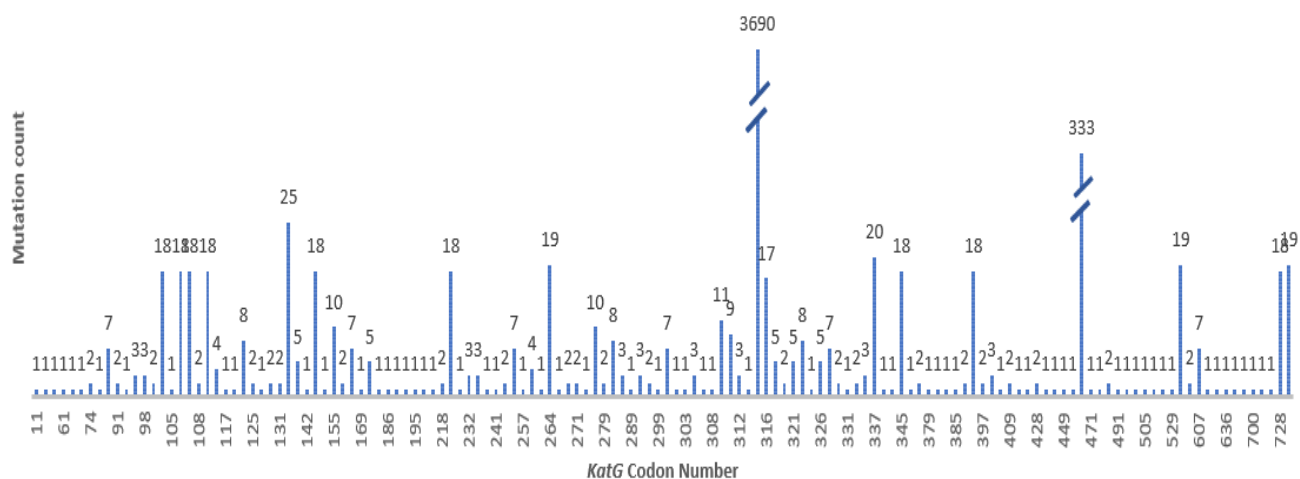
Figure 3. 5: Mutation count of *katG* observed in INH resistant *Mycobacterium tuberculosis* isolates

(ii) *rpoB* mutations associated with RIF resistance

RIF is used to attack the genetic machinery of *Mycobacterium tuberculosis*. It targets the β-subunit of RNA polymerase encoded by *rpoB,* thus, inhibiting the transcription process (Zhang & Yew, 2009). Generally, the *E. coli* numbering system for *rpoB* is used as a standard reference for identifying mutations associated with RIF resistance. This numbering system has sometimes led to misinterpretation of drug resistant mutations. In this study, we used the *Mycobacterium tuberculosis rpoB* numbering system, as described by Andre et al. (2017). *rpoB* consists of an 81 base-pair rifampicin-resistance determining region (RRDR), from position 507 to 533, in *E. coli* (Taniguchi et al., 1996) and, from codon 426 to 452, in *Mycobacterium tuberculosis* (Andre et al., 2017). The collected mutation data for *rpoB* found that 42 distinct *rpoB* codons had 120 non-synonymous amino acid mutations. The range of mutation counts for *rpoB* varied from 1 to 3054. The most prevalent mutations were identified at codon number S450 with 3054 mutations accompanied by *rpoB* H445 with 1019 mutations (Figure 3.6). Eight codons had only a single mutation. It was found that most of the RIF drug resistant mutations were present inside RRDR. Forty-nine mutations were present outside RRDR at 17 distinct codons: 170, 413, 424, 453, 454, 455, 457, 460, 480, 481, 482, 483, 487, 488, 491, 493 and 507.

The most commonly occurring *rpoB* mutations at positions *rpoB* S450L, H445Y, D435V and H445D had single mutation frequencies of 56.7%, 8.01%, 6.5% and 4.53%, respectively. The high density of mutations occurring in RRDR seem to hamper the binding of a drug to *rpoB* leading to RIF resistance. Unfortunately, the crystal structure for *Mycobacterium tuberculosis rpoB* is not available.
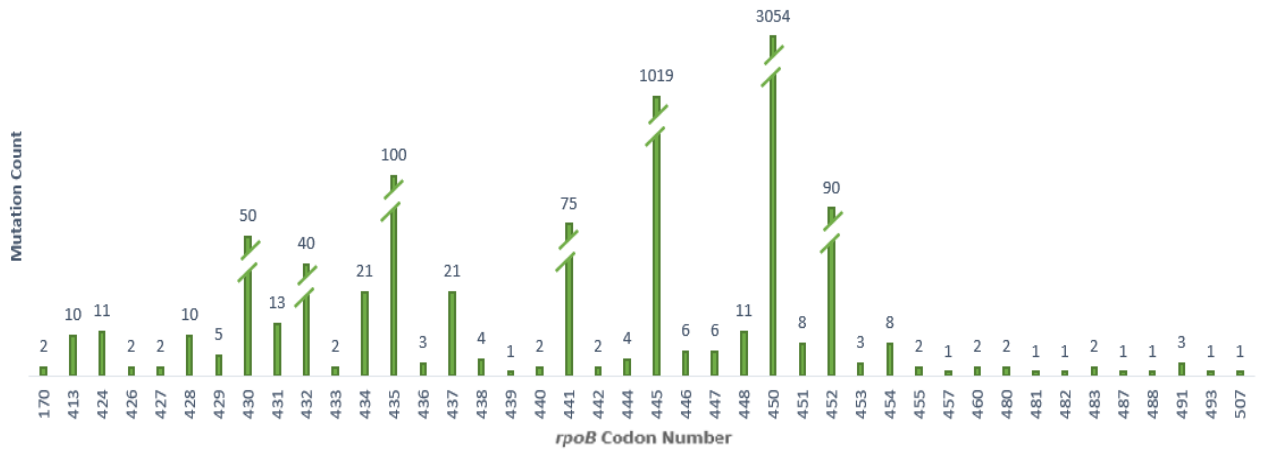
Figure 3. 6: Mutation count of *rpoB* observed in RIF resistant *Mycobacterium tuberculosis* isolates

(iii) *pncA* mutations associated with PNZ resistance

PZA, a prodrug, is activated by pyrazinamidase encoded by *pncA.* It weakens the cell membrane and inhibits transportation across the cell membrane (Zhang & Yew, 2009). For *pncA*, 273 different types of non-synonymous mutations were observed at 111 distinct codons. Unlike other first-line drug targets, mutations in *pncA* are scattered throughout the gene. The mutation count for *pncA* ranged from 1 to 62. The most pervasive mutation was observed at *pncA* W68 with 62 mutations, followed by Q10 with 54 mutations and codon H57 with 52 mutations (Figure 3.7). Scorpio et al. (1997) described three catalytic regions in *pncA*: I5-D12, P69-L85 and G132-T142. We found that the G132-T142 region had more mutations compared to I5-D12 and P69-L85.

In contrast to the mutation count at each amino acid position, the single mutation frequency was observed at Q10P with 1.46%, accompanied by W68R and Y103D with 1.03% single mutation frequencies. The crystal structure of *pncA* is available with 3PL1 PDB entry (Petrella et al., 2011). While studying the crystal structure, we found that the ligand-binding region (D49, H51, H57 and H71) was not frequently mutated except for H57. This indicates that that *pncA* has different resistance mechanisms, which needs investigation.
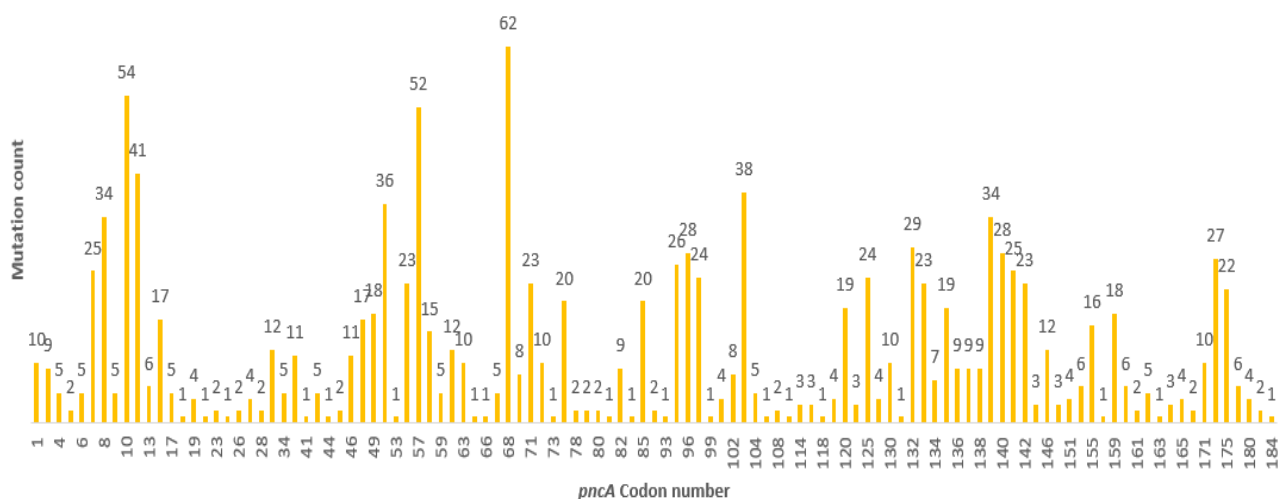
Figure 3. 7: Mutation count of *pncA* observed in PZA resistant *Mycobacterium tuberculosis* isolates

(iv) *emb* mutations associated with EMB resistance

Drug resistant mutations occurring in the *embCAB* operon, encoding for arabinosyltransferase C, A and B, confer EMB resistance. EMB targets arabinosyltransferase for inhibiting arabinogalactan synthesis, thus weakening the cell wall (Bhat et al., 2018). In our study, we observed 168 different point mutations in *embB* at 101 distinct codons. The most frequent mutations in *embB* occurred at the M306 codon with 900 mutations, followed by codon G406 with 164 mutations and codon 497 with 115 substitutions (Figure 3.8(ii)). The highest single mutation frequency was observed at *embB* M306V with 29.18%, followed by M306I and G406A with 18.01% and 4.71%, respectively. For *embA* and *embC*, no significant mutation data were available. From the available data, we identified that there were only two mutations present at the *embA* 4 and 913 codons (Figure 3.8(i)) whereas, *embC* was commonly mutated at codon numbers 270 with 31 mutations and at position 981 with 30 mutations (Fig 3.8(iii)). However, the crystal structure for *embA* and *embB* is not available. 3PTY PDB entry is the known crystal structure for the C-terminal domain of *embC* (Alderwick et al., 2011). Thus, understanding the EMB resistance mechanism in *embCAB* is a challenge due to the limited mutation data and lack of crystal structure.
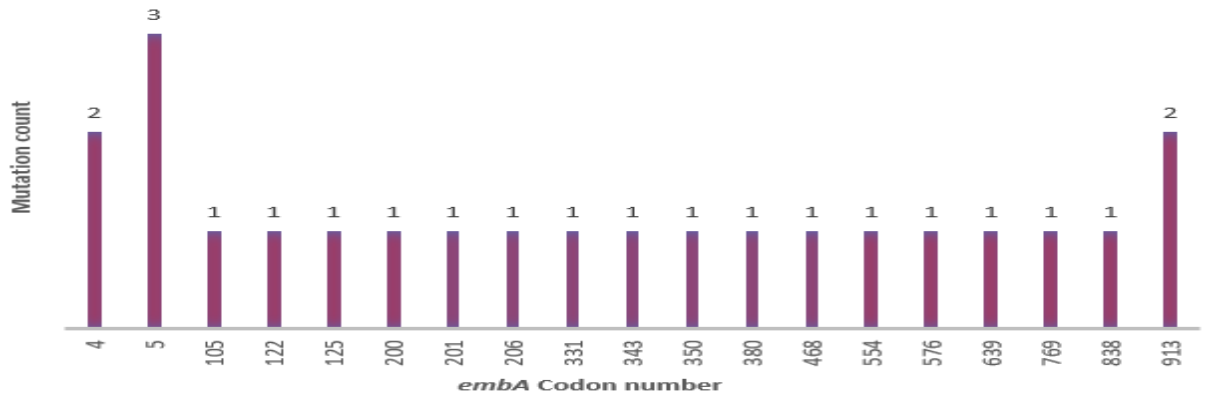
(i)
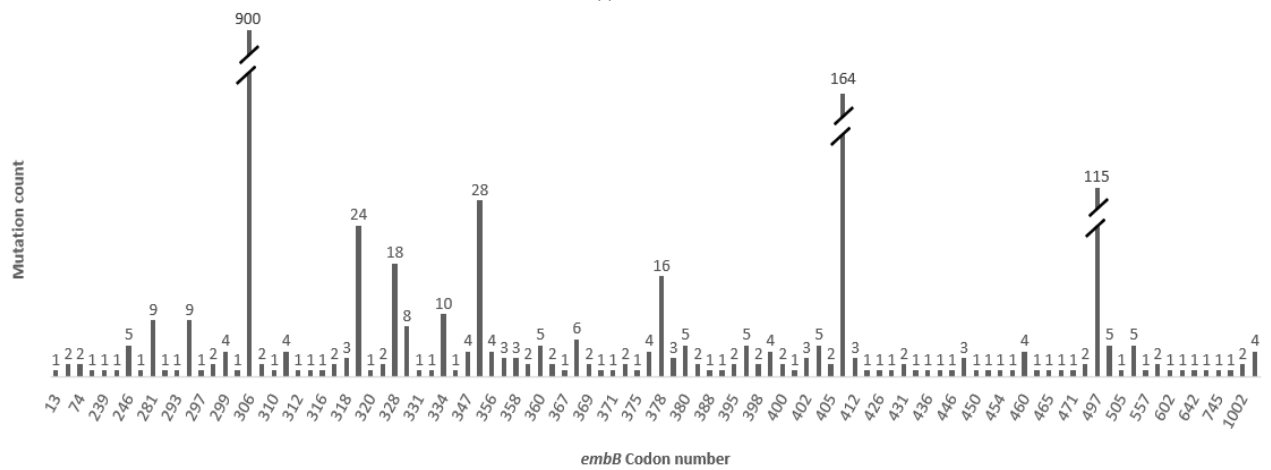
(ii)
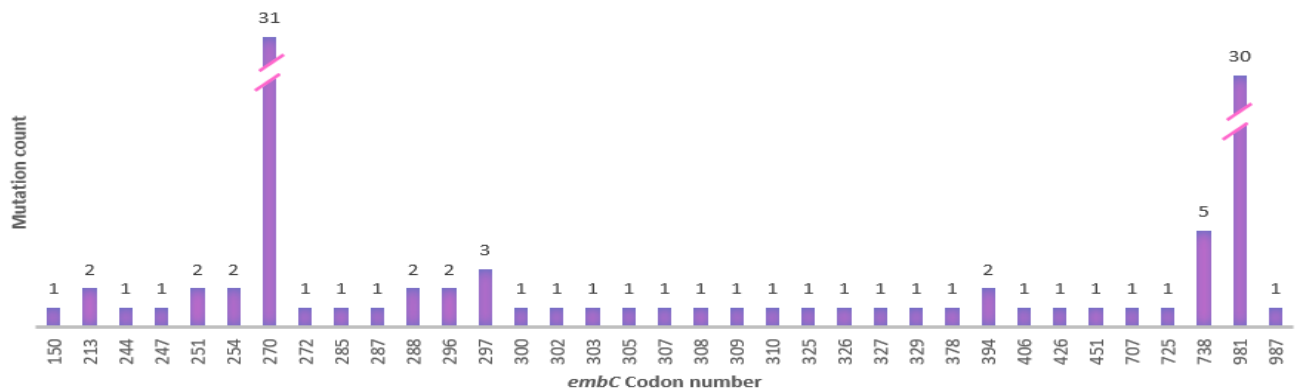
(iii)

Figure 3. 8: Mutation count for *emb* associated with EMB resistance: (i) *embA*, (ii) *embB,* and (iii) *embC* observed mutations in EMB resistant tuberculosis isolates

When we looked deeply into mutation patterns, spread, and frequency, we found that the most commonly occurring mutations were highly prevalent in all six WHO regions. However, their frequency varied in each region. For example, *katG* S315T and *rpoB* S450L were the most highly occurring mutations in all WHO regions.

3.4.4 Hotspot mutation site in the first-line TB drug targets

The number of amino acid substitutions occurring at each mutated position of the first-line TB drug targets was also evaluated (Appendix C). The position in the target gene with three or more substitutions was considered a hotspot residue site in our study. *katG* had 12 sites with multiple amino acid substitutions. S315 was found to have the maximum number of eight substitutions (S→ A, D, G, I, L, N, R, T). In *rpoB,* 14 sites with distinct replacements and H445, in particular, with 13 variants (H→A, C, D, E, G, L, N, P, Q, R, S, T, Y) were identified. A high degree of diversity with 44 hotspot sites was observed for *pncA* with H71 having eight distinct amino substitutions (H→D, E, N, P, Q, R, T, Y) followed by D8, H51, H57 and W68 with six multiple mutations. *embA* and *embC* had no hotspot sites, whereas *embB* showed 18 hotspot sites, with G406 having seven variants (G→P, S, C, K, R, D, A). The relationship between mutation frequency and hotspot site revealed that the frequently mutated position in the first-line TB drug target also had many variants.

The results above (single mutation frequency and hotspot residue site) showed that *Mycobacterium tuberculosis* might have a more focused approach for generating resistance against INH and RIF. The prolonged exposure to these two drugs could be contributing to the evolution of *Mycobacterium tuberculosis.* In INH resistance, the widespread mutation, *katG* S315, in the drug-binding region, is the most commonly used strategy for generating resistance against INH. However, the strategy used for evading the other three drugs is not clearly understood. Therefore, it is essential to determine the location of frequent mutations in drug targets and the impact of mutation in terms of its presence in conserved region, functional change, structural stability change, location as well as changes in drug binding to understand the survival mechanism of *Mycobacterium tuberculosis* against first-line TB drugs. For understanding the strategy used by *Mycobacterium tuberculosis*, we performed a comprehensive computational analysis on the 821 non-synonymous mutations.

3.4.5 Structure modelling, validation and energy minimization of first-line TB drug targets

The structural models for wild-type (WT) catalase-peroxidase, β subunit of RNA polymerase, pyrazinamidase, arabinosyl transferase A, B and C encoded by *katG*, *rpoB*, *pncA, embA, embB* and *embC,* respectively and mutant protein targets were generated using MODELLER v9.23.

(1) **Retrieval of target sequence:** First, the protein sequence of catalase-peroxidase (P9WIE5), pyrazinamidase (Q50575), β-subunit of RNA polymerase (A0A0K0PZB9), arabinosyltransferase A (P9WNL9), arabinosyltransferase B (P9WNL7) and arabinosyltransferase C (P9WNL5) were retrieved from the UniProt database. It was necessary to convert protein sequence into the PIR file format, so that it is easily read and processed by MODELLER v9.23. The converted

sequence was saved as file= "target.ali"; for example, "catp.ali" used for generating model for mutant catalase-peroxidase S315T.

(2) **Selection of template structure:** PSI-BLAST was performed by taking the PDB database as the source database to select template structure. The template was chosen with the maximum identity percentage and query coverage. The chosen template for the wild-type protein was taken as a template for modelling their respective mutant proteins. For example, 1SJ2 was selected as a template for modelling mutant catalase-peroxidase S315T, 3PL1 for pyrazinamidase Q10P and 5UH5 for the β-subunit of RNA polymerase H445R. Table 3.4 shows the result of PSI-BLAST for wild-type proteins.

Table 3. 4: Results of template structure search from PSI-BLAST

Protein	Protein sequence accession number	Amino acid length	Selected template PDB ID	Identity (%)	PDB Chain	Query cover (%)
Catalase-peroxidase	P9WIE5	740	1SJ2	99.86	A	100
Pyrazinamidase	Q50575	186	3PL1	100	A	100
β-subunit of RNA polymerase	A0A0K0PZB9	1172	5UH5	99.83	C	100
Arabinosyltransferase A	P9WNL9	1094	3PTY	40.4	A	35
Arabinosyltransferase B	P9WNL7	1098	3PTY	42.97	A	33
Arabinosyltransferase C	P9WNL5	1094	3PTY	100	A	34

Table 3.4 shows the details of the templates selected for homolog modelling with a sequence identity of the query with selected template and query coverage for structural model generation. The selected templates for catalase-peroxidase, pyrazinamidase and β-subunit of RNA polymerase shows the 100% query coverage with more than 99% sequence identity. 3PTY PDB entry is the known crystal structure for the C-terminal domain of embC. 3PTY was selected as a template for constructing the structural models for arabinosyltransferase A, B and C. Due to the unavailability of structural information for arabinosyltransferase A, B and C, the query coverage and sequence identity were low.

(3) **Alignment of a target sequence and template structure:** The alignment of the target sequence with template structure was undertaken using the "align 2d()" command in MODELLER v9.23. A python script was generated and saved as "target-template.py". Two alignments files were generated after target-template alignment: "target-template.pap" and "target-template.ali".

For example, Figure 3.9 illustrates "catp-1sj2.pap" generated after alignment between target sequence and template (1sj2) for wild-type catalase-peroxidase (P9WIE5). File "catp-1sj2.ali" was used for subsequent model generation.

```
_aln.pos        10        20        30        40        50        60
1sj2A   ----------------------GHMKYPVEGGGNQDWWPNRLNLKVLHQNPAVADPMGAAFDYAAEV
catp    MPEQHPPITETTTGAASNGCPVVGHMKYPVEGGGNQDWWPNRLNLKVLHQNPAVADPMGAAFDYAAEV
_consrvd                      ********************************************

_aln.p  70        80        90        100       110       120       130
1sj2A   ATIDVDALTRDIEEVMTTSQPWWPADYGHYGPLFIRMAWHAAGTYRIHDGRGGAGGGMQRFAPLNSWP
catp    ATIDVDALTRDIEEVMTTSQPWWPADYGHYGPLFIRMAWHAAGTYRIHDGRGGAGGGMQRFAPLNSWP
_consrvd ******************************************************************

_aln.pos 140       150       160       170       180       190       200
1sj2A   DNASLDKARRLLWPVKKKYGKKLSWADLIVFAGNCALESMGFKTFGFGFGRVDQWEPDEVYWGKEATW
catp    DNASLDKARRLLWPVKKKYGKKLSWADLIVFAGNCALESMGFKTFGFGFGRVDQWEPDEVYWGKEATW
_consrvd ******************************************************************

_aln.pos 210       220       230       240       250       260       270
1sj2A   LGDERYSGKRDLENPLAAVQMGLIYVNPEGPNGNPDPMAAAVDIRETFRRMAMNDVETAALIVGGHTF
catp    LGDERYSGKRDLENPLAAVQMGLIYVNPEGPNGNPDPMAAAVDIRETFRRMAMNDVETAALIVGGHTF
_consrvd ******************************************************************

_aln.pos    280       290       300       310       320       330       340
1sj2A   GKTHGAGPADLVGPEPEAAPLEQMGLGWKSSYGTGTGKDAITSGIEVVWTNTPTKWDNSFLEILYGYE
catp    GKTHGAGPADLVGPEPEAAPLEQMGLGWKSSYGTGTGKDAITSGIEVVWTNTPTKWDNSFLEILYGYE
_consrvd ******************************************************************

_aln.pos    350       360       370       380       390       400
1sj2A   WELTKSPAGAWQYTAKDGAGAGTIPDPFGGPGRSPTMLATDLSLRVDPIYERITRRWLEHPEELADEF
catp    WELTKSPAGAWQYTAKDGAGAGTIPDPFGGPGRSPTMLATDLSLRVDPIYERITRRWLEHPEELADEF
_consrvd ******************************************************************

_aln.p  410       420       430       440       450       460       470
1sj2A   AKAWYKLIHRDMGPVARYLGPLVPKQTLLWQDPVPAVSHDLVGEAEIASLKSQIRASGLTVSQLVSTA
catp    AKAWYKLIHRDMGPVARYLGPLVPKQTLLWQDPVPAVSHDLVGEAEIASLKSQILASGLTVSQLVSTA
_consrvd *********************************************** ******************

_aln.pos 480       490       500       510       520       530       540
1sj2A   WAAASSFRGSDKRGGANGGRIRLQPQVGWEVNDPDGDLRKVIRTLEEIQESFNSAAPGNIKVSFADLV
catp    WAAASSFRGSDKRGGANGGRIRLQPQVGWEVNDPDGDLRKVIRTLEEIQESFNSAAPGNIKVSFADLV
_consrvd ******************************************************************

_aln.pos 550       560       570       580       590       600       610
1sj2A   VLGGCAAIEKAAKAAGHNITVPFTPGRTDASQEQTDVESFAVLEPKADGFRNYLGKGNPLPAEYMLLD
catp    VLGGCAAIEKAAKAAGHNITVPFTPGRTDASQEQTDVESFAVLEPKADGFRNYLGKGNPLPAEYMLLD
_consrvd ******************************************************************

_aln.pos    620       630       640       650       660       670       680
1sj2A   KANLLTLSAPEMTVLVGGLRVLGANYKRLPLGVFTEASESLTNDFFVNLLDMGITWEPSPADDGTYQG
catp    KANLLTLSAPEMTVLVGGLRVLGANYKRLPLGVFTEASESLTNDFFVNLLDMGITWEPSPADDGTYQG
_consrvd ******************************************************************

_aln.pos    690       700       710       720       730       740
1sj2A   KDGSGKVKWTGSRVDLVFGSNSELRALVEVYGADDAQPKFVQDFVAAWDKVMNLDRFDVR
catp    KDGSGKVKWTGSRVDLVFGSNSELRALVEVYGADDAQPKFVQDFVAAWDKVMNLDRFDVR
_consrvd ***********************************************************
```

Figure 3. 9: Target-template alignment. The figure shows the "catp-1sj2.pap" file generated after alignment between the target sequence (wild-type catalase-peroxidase (P9WIE5)) and the template (1sj2) using align 2d() command in MODELLER v9.23; "*" represents conserved regions

(4) **Building structural model:** After aligning the target sequence and template structure, "target-template.ali" was generated to generate a model using the "automodel" class of MODELLER v9.23. A python script was used to create, for example, ten similar structural models for wild-type catalase-peroxidase, pyrazinamidase, β-subunit of RNA polymerase,

arabinosyltransferase A, arabinosyltransferase B and arabinosyltransferase C. The result of model building was stored in the "catp_model.log" file, which contained the molecular PDF (molpdf), **d**iscrete **o**ptimized **p**rotein **e**nergy (DOPE) (Shen & Sali, 2006) and GA341 scores (Melo et al., 2009). Figure 3.10 presents the summary of the generated models for wild-type catalase-peroxidase (P9WIE5).

```
>> Summary of successfully produced models:
Filename                    molpdf    DOPE score    GA341 score
------------------------------------------------------------------
catp.B99990001.pdb          4078.44189   -81857.86719     1.00000
catp.B99990002.pdb          4154.82031   -81971.18750     1.00000
catp.B99990003.pdb          3910.85693   -81636.92188     1.00000
catp.B99990004.pdb          3975.54688   -81684.92188     1.00000
catp.B99990005.pdb          3852.26440   -82042.59375     1.00000
catp.B99990006.pdb          3823.98389   -82128.24219     1.00000
catp.B99990007.pdb          4113.00342   -81814.24219     1.00000
catp.B99990008.pdb          3781.69556   -82117.51563     1.00000
catp.B99990009.pdb          4030.13452   -81648.96875     1.00000
catp.B99990010.pdb          4053.96069   -81698.94531     1.00000
```

Figure 3. 10: Summary of the structural models generated for wild-type catalase-peroxidase (P9WIE5). The figure shows the first ten models generated using MODELLER v9.23 along with molpdf, DOPE and GA341 scores.

Molpdf represents the standard homology modelling score function. DOPE score assesses the quality of the atoms in the constructed structural model using the DOPE method (Shen & Sali, 2006). The 3D structure models with the lowest DOPE score are considered good models. GA341 score is used for assessing the quality of the structural model. The structural model with GA341 scores of more than 0.5 is considered a good model (Melo et al., 2009).

(5) **Selection of model**: The results for model building were stored in a log file, which contained molpdf, and the DOPE and GA341 scores. The 3D structure models with the lowest DOPE score and highest GA341scores were selected for structural validation. For example, according to the summary result displayed in Figure 3.10, catp.B9999006.pdb had the highest GA341 score of 1.00000 and the lowest DOPE score of -82128.24219. Thus, "catp.B9999006.pdb" was chosen as a 3D structure model for wild-type catalase-peroxidase (P9WIE5).

Following this, models were generated for the mutant-type proteins. For the 821 mutations, we were able to construct a total of 603 mutant protein structural models. Due to the unavailability of a suitable template for arabinosyltransferase (*emb*), only 12 mutant protein models were generated. (Python scripts used for structural modelling are in the Appendix, section H1 and H2).

The model selected was further evaluated for stereochemical quality using PROCHECK, ERRAT and Verify 3D. The chosen model had more than 90% of residues in the most favoured region, which illustrates the importance of selecting the best quality model. Energy minimization of the protein models was performed by the GROMACS package using the GROMACS96 43A1 force field. Similarly, six wild-type target protein structures were constructed. Figure 3.11 shows the final structural models selected for the main wild-type first-line TB drug targets.

(i)
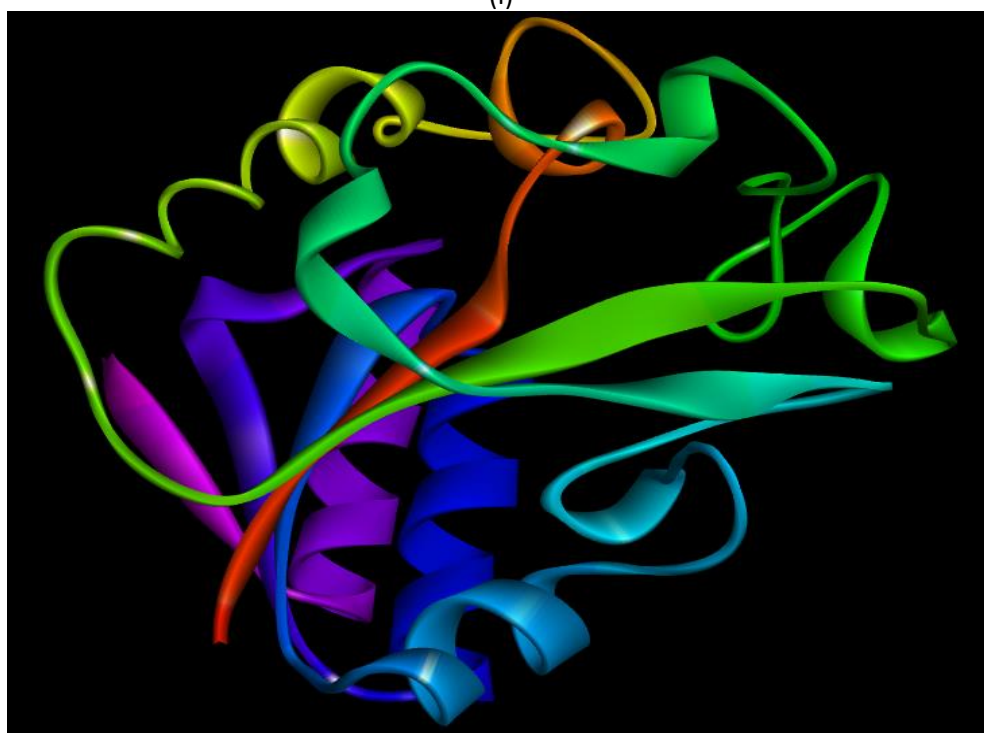
(ii)

(iii)
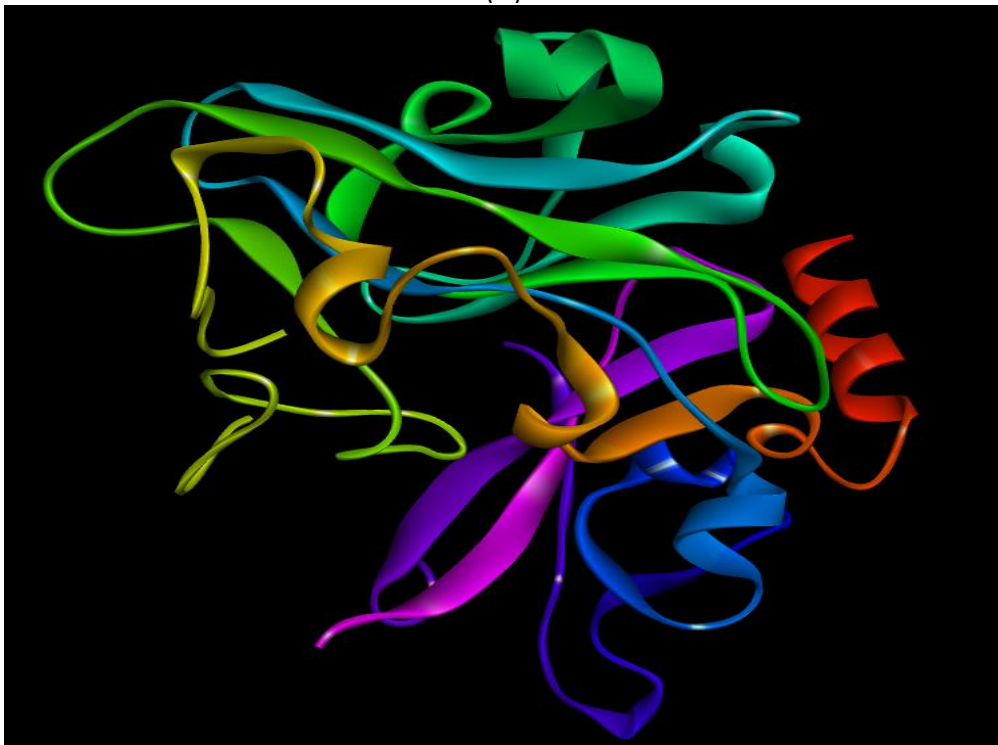
(iv)

Figure 3. 11: 3D structural models constructed for the wild-type first-line drug target proteins: (i) catalase-peroxidase (*katG*), (ii) pyrazinamidase (*pncA*), (iii) β-subunit of RNA polymerase (*rpoB*), and (iv) arabinosyltransferase B (*embB*). The structures are coloured in rainbow colour from violet to red (starting from N-terminus to C-terminus)

3.4.6 Comprehensive computational analysis of the drug resistant strategy of first-line TB drug targets

A mutation in the drug target can change its biological function and structural stability. For a mutation to be drug resistant, it must cause a minimum loss to the infectious agent. Thus, it is essential to identify the impact of first-line TB drug resistant mutations on *Mycobacterium tuberculosis*. Here, we used several bioinformatics tools to determine the effect of the 821 drug resistant mutations on the properties of the target proteins to assess how these mutations impact the fitness of *Mycobacterium tuberculosis*. Figure 3.1 presents the workflow depicting the approach used to systematically understand drug resistance and the number of different bioinformatics tools used for each step (in bold letters). In particular, we combined the results of the impact of mutations on *Mycobacterium tuberculosis* fitness and the effect of mutations on drug binding/activation, and then ranked the 821 mutations into neutral, mild, moderate and lethal categories to uncover the *Mycobacterium tuberculosis* drug resistance strategy. This understanding will help unravel the drug resistance mechanisms of *Mycobacterium tuberculosis* and develop effective strategies to prevent or eliminate TB at its root.

3.4.6.1 Evolutionary conservation analysis

Evolutionary information about a protein sequence reveals the sites that can be more prone to mutation and what impact a mutations in conserved or variable (non-conserved) position can have on a protein. From the perspective of evolution, mutations in less conserved regions will have less impact on structure and function than highly conserved areas. Mutations in highly conserved areas can significantly affect the target proteins' structure and function, leading to improper drug binding contributing to drug resistance. However, these can also impact considerably the fitness of *Mycobacterium tuberculosis*. Therefore, we first investigated if *Mycobacterium tuberculosis* in fact made use of mutations in the conserved regions. For this, we analysed the spread of mutations in the target in terms of sequence conservation. We used ConSurf for determining sequence conservation as ConSurf identifies functional regions based on evolutionary relationships among homologues. ConSurf provides the distribution of mutations in all four drug targets in three regions: highly conserved (HC), less conserved (C) and highly variable (V) (Figure 3.12). Out of the 821 mutations, 66.5% (546) were in HC, 14.13% (117) were in C and 19.37% (159) were in V, indicating that approximately 80% of mutations were in conserved regions (Appendix D summarizes the prediction of sequence conservation). This indicates that drug resistance strategy could come at a fairly significant cost to bacterial fitness depending on whether frequent mutations are in the conserved regions.
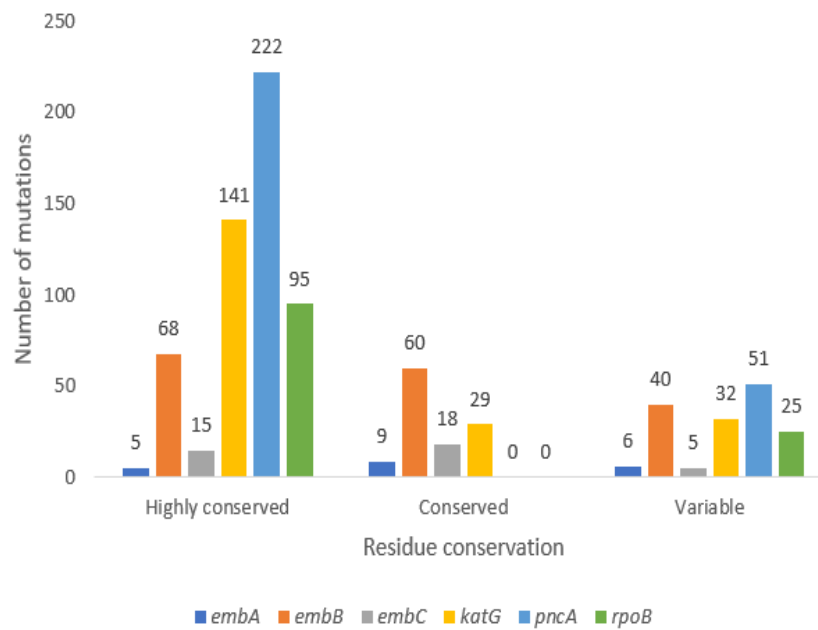
Figure 3. 12: Categorizing first-line TB drug resistant mutations into three categories: highly conserved, conserved and variable

It was interesting to observe that the mutations occurring more frequently were not present in the highly conserved and conserved regions of the first-line drug targets but were in the variable region. The most frequent mutation in catalase-peroxidase (*katG*), S315, and in β-subunit of RNA polymerase (*rpoB*), S450 and arabinosyltransferase B (*embB*), M306, were found in a highly variable region. However, the most frequent mutation in pyrazinamide (*pncA)*, Q10, was present in the highly conserved region. As for the less frequent mutations, we found them occurring equally in the conserved and variable region with no apparent pattern of occurrence across the regions.

Highly conserved residues are supposed to be involved in essential protein interactions crucial for maintaining the function of *Mycobacterium tuberculosis*. Surprisingly, although the mutation count is higher in the conserved regions (Figure 3.12), most of these were the least prevalent mutations. This suggests that *Mycobacterium tuberculosis* experiments with or employs different strategies not to jeopardise the normal functioning and structural stability needed for the TB bacteria survival. Thus, TB bacteria follow an evolutionary strategy with the slightest disturbance to highly conserved regions in attaining drug resistance.

We also identified domain regions in the first-line target proteins using Pfam, SMART and Interpro Scan. (Appendix D describes the protein domain region). Domains are considered as evolutionarily conserved regions in a protein. Amino acid residues present in the domain are responsible for performing essential biological functions. Therefore, mutations in them could lead to more remarkable

functional changes. More than 95% of mutations were in the domain region of their respective proteins. This indicates that *Mycobacterium tuberculosis* experiments with mutations in conserved regions but does so very sparingly, still favouring variable regions to mutate in developing its drug resistance strategy.

3.4.6.2 Distribution of drug resistant mutation into three sites: 1, 2 and 3

We then looked at the position of mutations from another angle; specifically, in terms of drug binding. For this, we categorised mutation positions in the target proteins into three sites: site-1, site-2 and site-3. Site-1 was defined as mutations present in the ligand-binding region; site-2 contains mutant residues directly interacting with site-1, and site-3 had mutations occurring elsewhere. Table 3.5 shows the mutations present at the ligand-binding sites (site-1 mutations) for first-line wild-type protein targets identified using COACH-D and 3D LigandSite. Out of 821 non-synonymous substitutions, 77 mutations (n) were present at 29 different positions (p) in the binding region of the four targets: catalase-peroxidase (n=31, p=17, encoded by *katG* and target of INH), pyrazinamidase (n=36, p=7, encoded by *pncA* and target of PZA), β subunit of RNA polymerase (n=10, p=5, encoded by *rpoB* and target of RIF).

Table 3. 5: Mutations present in site-1 identified using COACH and 3D LigandSite

Protein	Ligand binding site with mutations
Catalase-peroxidase	101 LEU, 104 ARG, 107 TRP, 229 TYR, 230 VAL, 232 PRO, 248 ILE, 269 GLY, 270 HIS, 274 LYS, 275 THR, 314 THR, 315 SER, 317 ILE, 321 TRP, 380 THR, 408 PHE
Pyrazinamidase	8 ASP, 13 PHE, 49 ASP, 51 HIS, 57 HIS, 71 HIS, 102 ALA
DNA dependent RNA polymerase subunit beta	432 GLN, 448 ARG,483 PRO, 487 ASN, 491 ILE

S315 in catalase-peroxidase had the highest and H57 in pyrazinamidase had the third-highest mutation count. For arabinosyltransferase A, B and C (encoded by *embA, B, C* and target of EMB), no mutations were found in site-1.

Table 3.6 shows the mutant residues present in site-2 found from the RING webserver. We found 53 mutations (m) present at 25 different positions (r) in the site-2 of catalase-peroxidase, pyrazinamidase and β subunit of RNA polymerase. Catalase-peroxidase had 22 mutations in site-2 (r=14, m=22) and pyrazinamidase had 30 mutations (r=10, m=30). The β subunit of RNA polymerase had only one mutation present at site-2. W68 in pyrazinamidase had the highest and H445 in β subunit of RNA polymerase had the second-highest mutation count. For arabinosyltransferase A, B and C, there were

no site-2 mutations; all its mutations were present in site-3 with no mutations in site-1 and site-2 (Appendix D provides information on the distributed sites for each first-line TB drug target.)

Table 3. 6: Mutations present in site-2 identified by RING webserver

Protein	Mutated residues interacting with site-1
Catalase-peroxidase	105 MET, 108 HIS, 172 ALA, 176 MET, 234 GLY, 251 THR, 257 MET, 262 THR, 316 GLY, 318 GLU, 328 TRP, 350 ALA, 384 LEU, 415 LEU
Pyrazinamidase	19 LEU, 21 VAL, 47 THR, 54 PRO, 58 PHE, 68 TRP, 72 CYS, 96 LYS, 97 GLY, 133 ILE
DNA dependent RNA polymerase subunit beta	445 HIS

As shown in the above tables, of all the first-line TB targets, only S315 of catalase-peroxidase showed that the presence of a mutation in the ligand-binding region could be the potential source for INH resistance. In contrast, most commonly occurring mutations in the β-subunit of RNA polymerase, pyrazinamidase and arabinosyltransferase were present in sites-2 and 3, indicating that *Mycobacterium tuberculosis* may follow a different, potentially less drastic strategy with these three targets.

3.4.6.3 Determination on the impact of drug resistant mutations on biological function

We used three predictive tools, SIFT, PolyPhen-2 and PROVEAN, to determine the impact of mutations on the function of first-line drug targets. Figure 3.13 shows the deleterious mutations predicted by these computational tools. Deleterious mutations are those that can significantly alter the biological function of the target. The differences found in the predictions shown in Figure 3.13 can be due to the different algorithms used by various tools. SIFT and PROVEAN use sequence information for predicting functional changes, whereas PolyPhen-2 is not solely dependent on sequence homology, but also uses the structural information of a protein for determining the impact on function. SIFT and PROVEAN directly predict deleterious mutations. In the case of PolyPhen-2 prediction, we considered 'probably damaging' and 'possibly damaging' mutations as 'deleterious.' Out of 821 amino acid substitutions, 507 were predicted to be 'deleterious' (D) (Figure 3.13) and 79 as 'neutral' (N) by all three programs (Appendix E describes functional change for each mutation in drug targets as predicted by each tool).
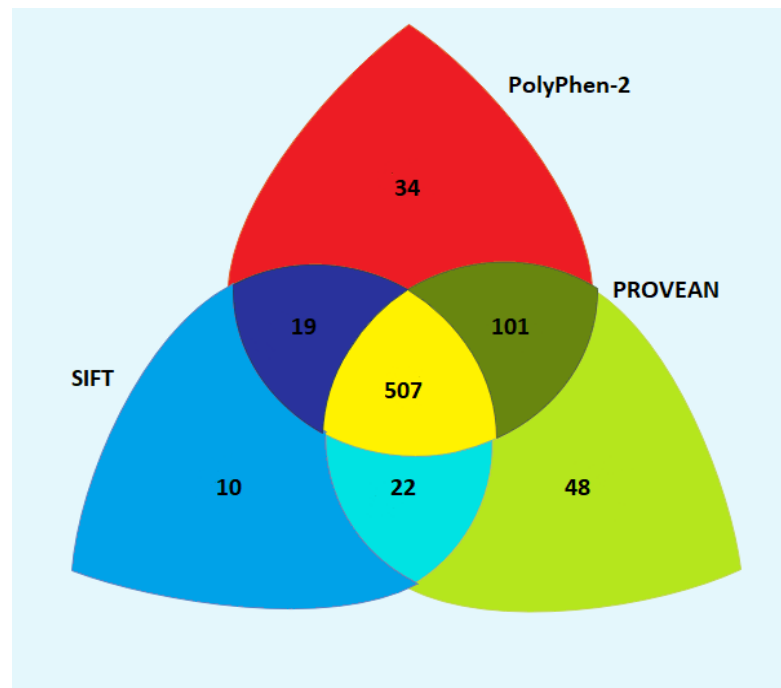
Figure 3. 13: Deleterious drug resistant mutations predicted by SIFT, PROVEAN and PolyPhen-2

Catalase-peroxidase had 148 deleterious mutations, pyrazinamidase had 158 mutations, the β-subunit of RNA polymerase had 106 mutations and arabinosyltransferase had 95 deleterious mutations, as predicted by all three tools. A larger number of deleterious mutations were present in pyrazinamidase and catalase-peroxidase than in the other two first-line targets. The changes in biological function are not beneficial to *Mycobacterium tuberculosis*; therefore, there must be some form of compensation for the lost function.

In terms of location, 67 deleterious mutations were found in site-1 and were distributed among the four targets as: catalase-peroxidase (encoded by *katG*) (n=25), pyrazinamidase (*pncA*) (n=34) and β-subunit of RNA polymerase (*rpoB*) (n=8). Only one mutation in site-1 of catalase-peroxidase I317L was found to be neutral indicating that catalase-peroxidase undergoes a significant change in the binding region which may have contributed to the deleterious effect on its biological function. All 95 deleterious mutations of arabinosyltransferase (*emb*) were in site-3. Functional changes due to mutation in site-1 could be directly involved in altering the ability of the target to bind with the first-line TB drug. Change in function occurring in target protein because of site-2 and site-3 variations could be indirectly involved in reducing the enzymatic activity of the target, for example, non-activation of pro-drugs such as INH and PZA.

The presence of a large number of deleterious mutations shows that the target proteins undergo significant biological changes to acquire drug resistance. The deleterious mutations were less frequently occurring. In contrast, frequently occurring mutations were found to have a mild to

moderate impact on the biological function of first-line TB drug targets indicating a favourable drug resistant strategy. These predictions enable researchers to give preference to and work on mutations based on their biological significance.

3.4.6.4 Predicting the structure stability changes in first-line targets due to mutations

Stability is crucial for correct protein folding and maintaining the biological function of proteins as it provides structural integrity to carry out their roles. For determining changes in structural stability, we used I-MUTANT 3.0 and mCSM. mCSM predicted that some of the mutations (16%) drastically change stability (highly destabilising, HDT); an example is the W68G position found mutated at site-2 of pyrazinamidase that had a highly destabilizing effect. Figure 3.14(i) shows that out of the 821 mutations, most (96%) destabilise the target, with only marginal differences in their free energy values as predicted by mCSM. As for I-MUTANT 3.0 predictions, out of the 821 mutations, 678 (82.6%) mutations 'decrease' and only 143 (17.4%) 'increase' structural stability (Figure 3.14(ii)) (Appendix E describes the structural stability changes for each mutation in drug target predicted by each tool).

In particular, the most frequent mutation site, S315 of catalase-peroxidase, caused a minor change in structural stability in the target structure, which could be a reason for its prevalence in drug resistance. Similarly, frequent mutations in the β-subunit of RNA polymerase cause little destabilization indicating that it also tries to maintain its stability while reducing drug binding. Mutations in *pncA* and *emb* had the most destabilising effect on stability. Overall, mutations predominantly decrease the stability of the target proteins with only a marginal increase in stability. Further, the most frequently occurring mutations in first-line drug targets cause moderate changes in protein stability (without highly destabilising it) in resisting drugs but they cause greater harm to biological function. Considering the effects of mutations on target protein function and stability, this appears to be the strategy of *Mycobacterium tuberculosis* for drug resistance: sacrificing function for stability. In the later section of this chapter, we explored how changes in biological function and structural stability specifically affect *Mycobacterium tuberculosis* survival.

(i)   (ii)
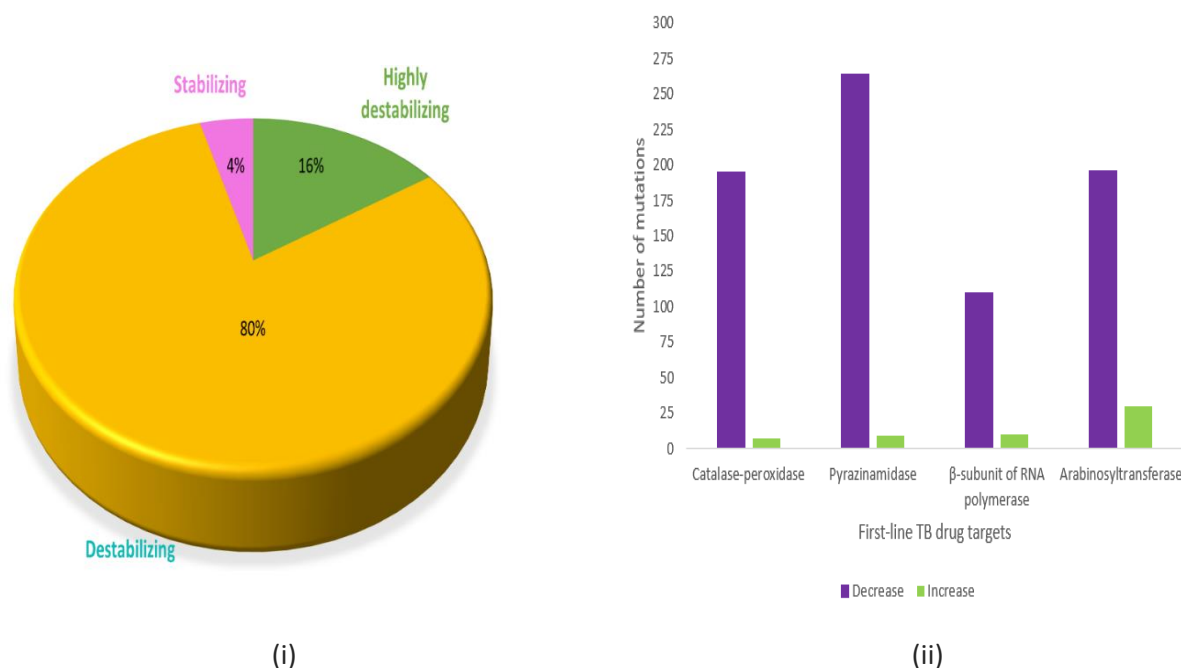
Figure 3. 14: First-line mutant target protein stability changes predicted by I-MUTANT 3.0 and mCSM: (i) Distribution of stability changes into 'highly destabilizing,' 'destabilizing' and 'stabilizing' identified for 821 mutations in first-line targets as predicted by m-CSM, and (ii) 'Decrease' and 'Increase' in stability in first-line mutant targets as predicted by I-MUTANT 3.0

3.4.6.5 Analysing the impact of drug binding affinity

Our analysis observed that each mutation directly or indirectly affects drug binding with its target. To determine the pronounced effect of mutations on the conformation of the first-line TB mutant proteins, drug binding energies were calculated between ligand (first-line drugs) and receptor (first-line drug targets) using AutoDock. Figure 3.15 shows the binding energy calculated for catalase-peroxidase (*katG*), pyrazinamidase (*pncA*) and the β-subunit of RNA polymerase (*rpoB*) mutant first-line TB targets. Only fifteen binding energies were calculated for arabinosyltransferase, 3 wild-type and 12 mutant-type, due to the unavailability of a suitable template for arabinosyltransferase (*emb*). It was observed that wild-type proteins' binding energies were more positive than the mutant proteins in contributing to drug resistance. Some less frequent mutations had more positive binding energies than wild type proteins.

For rifampicin, binding energy with the WT β-subunit of RNA polymerase was -8.23 kcal/mol and with the most frequent mutant H445R (site-2) was -7.78 kcal/mol. The docking score calculated for the WT catalase-peroxidase (*katG*) complex with INH was -5.53 kcal/mol. In contrast, binding energies with the identified most frequent mutants S315T, S315N and S315R of site-1 were -4.99 kcal/mol, -5.15 kcal/mol and -5.2 kcal/mol, respectively. WT pyrazinamidase had a docking score of -4.73kcal/mol and

the most frequent mutants Q10P (site-3) and W68R (site-2) had -3.9 had -4.29 kcal/mol, respectively. For arabinosyltransferase C, the docking score for WT was -4.03 kcal/mol and the most frequent mutant V981L (site-3) had binding energy of -3.78 kcal/mol. This weakened affinity indicates that these mutations cause changes that reduce drug binding and enzymatic activity of the drug targets.

The overall view of drug binding was that mutations in site-1 showed a more weakened drug binding affinity than mutations in site-2 and site-3. The changes in binding strength/affinity due to mutations in site-2 and site-3 could be due to some underlying mechanisms that directly (site-2) or indirectly (site-3) affect the orientation of the drug binding region through different mechanisms. We have already established that site- 1 also experiences deleterious changes in function and moderate changes in stability. Any change in function and stability comes with a fitness cost to *Mycobacterium tuberculosis*. As drug resistance appears to come with a fitness cost to *Mycobacterium tuberculosis*, it is essential to investigate the fitness cost of underlying changes due to mutations.
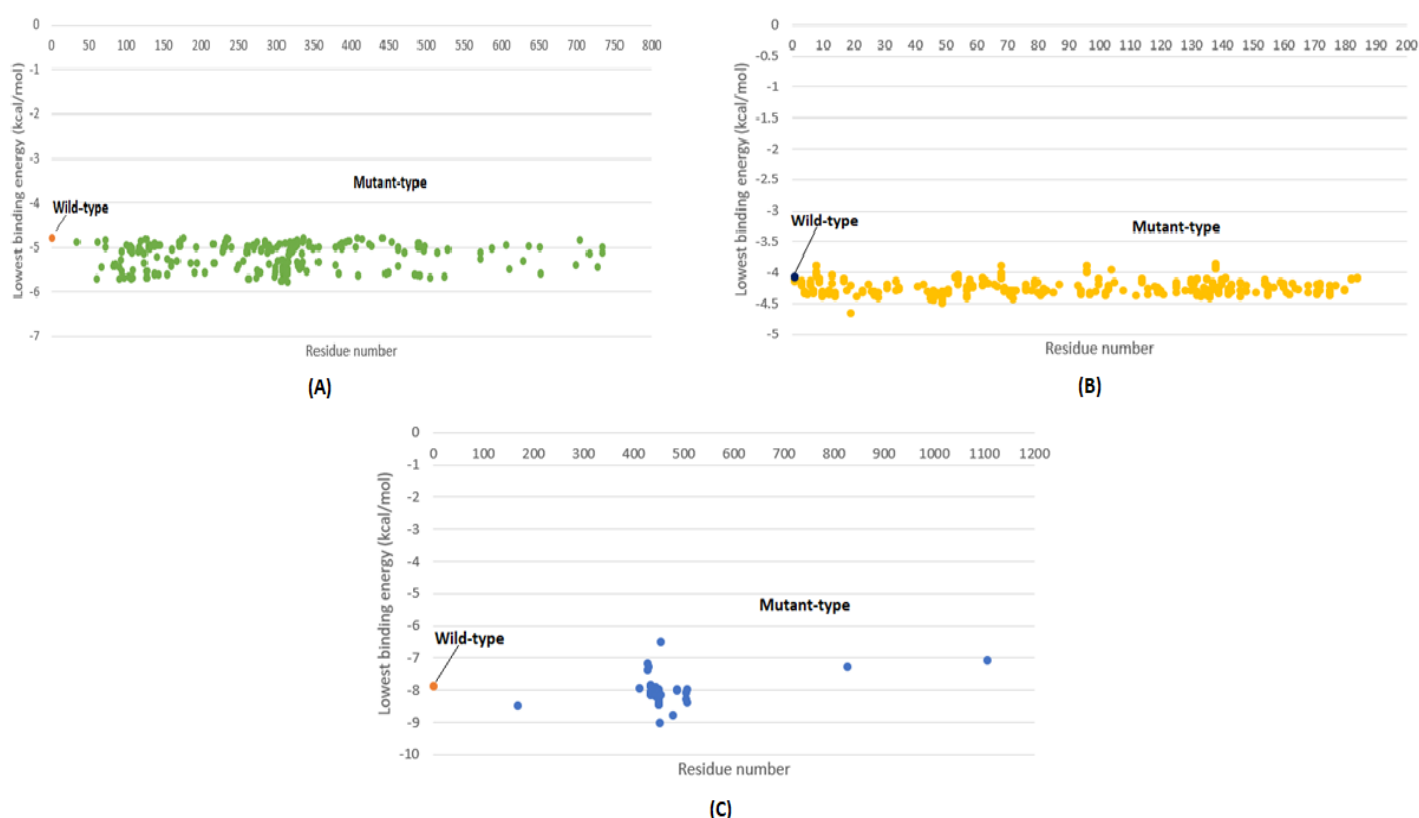
Figure 3. 15: Drug binding affinity of wild-type and mutant drug target protein: (A) catalase-peroxidase (*katG*), (B) pyrazinamidase (*pncA*), and (C) β-subunit of RNA polymerase (*rpoB*). The X-axis represents the residue number (position of the mutation in target protein) and Y-axis represents the lowest binding energy (kcal/mol) identified in each docking

3.4.6.6 Categorisation of mutation into lethal, moderate, mild and neutral

As hypothesised earlier, *Mycobacterium tuberculosis* may favour mutations harmful to the drug but benign to itself. However, in reality, *Mycobacterium tuberculosis* fitness may be compromised to a small or large extent while achieving this balance. Furthermore, a realistic assessment of the fitness cost is challenging because, as shown earlier, the various bioinformatics tools used for mutation analysis have varying degrees of overlap due to the different approaches they use, such as sequence-based or structure-based. Therefore, a deeper investigation into these results is needed to validate them and properly assess the fitness cost.

A concordance analysis was carried out to integrate results from the previous mutation analysis for proper validation and categorisation of mutations into lethal, moderate, mild and neutral. The mutations that were deleterious and present in the highly conserved region of drug target with a significant reduction in structural stability and binding affinity, were categorised as 'lethal.' As the level of these decreases, the ranking accordingly goes down to 'moderate' and 'mild.' Mutations with no harmful impact on *Mycobacterium tuberculosis* were defined as 'neutral.' Out of 821 variations, 340 were identified as 'lethal,' 284 as 'moderate,' 185 as 'mild' and 12 as 'neutral.'

Table 3.7 presents the distribution of first-line TB target mutations in all four categories. Appendix F contains detailed information of all four types. The number of lethal mutations was higher, but when we looked at their prevalence, we found that most lethal mutations were the least prevalent across the world. In contrast, mutations with a mild impact on *Mycobacterium tuberculosis* were highly persistent. For example, catalase-peroxidase and the β-subunit of RNA polymerase had more lethal mutations than moderate and mild mutations. Still, their most prevalent mutations were found to be in the mild mutation category. Due to the unavailability of complete structural information for arabinosyltransferase B (*embB*), we could not find any mutations in site-1 and site-2. Still, we found that the most frequently occurring mutations at position M306 present at site-3 had a moderate impact on *Mycobacterium tuberculosis.* In contrast, lethal mutations of pyrazinamidase (*pncA*) were highly prevalent.

Catalase-peroxidase (*katG*) and arabinosyltransferase are essential for cell wall synthesis, and the β-subunit of RNA polymerase (*rpoB*) is required for the transcription process (Almeida Da Silva & Palomino, 2011; Lange et al., 2018). The results found for these three targets indicate that *Mycobacterium tuberculosis* prefers mild to moderate mutations to achieve resistance for INH and RIF and most likely for EMB resistance as well while still being viable. This makes sense as *Mycobacterium tuberculosis* needs to protect the cell wall and protein production machinery to safeguard its survival.

Some studies suggest pyrazinamidase (*pncA*) is not considered essential for *Mycobacterium tuberculosis* growth and development (Baddam et al., 2018). Hence, *Mycobacterium tuberculosis* can afford to resist the drug more strongly with lethal mutations in pyrazinamidase without compromising its viability.

Table 3. 7: Number of lethal, moderate, mild and neutral mutations in first-line TB drug targets identified from comprehensive bioinformatics analysis

Lethal mutations (340)					
	Catalase-peroxidase (105)	Pyrazinamidase (115)	β-subunit of RNA polymerase (58)	Arabinosyltransferase A (5)	Arabinosyltransferase B (43)	Arabinosyltransferase C (14)
site-1	14	19	6	0	0	0
site-2	19	27	7	0	0	0
site-3	72	69	45	5	43	14
Moderate mutations (284)					
	Catalase-peroxidase (53)	Pyrazinamidase (126)	β-subunit of RNA polymerase (51)	Arabinosyltransferase A (9)	Arabinosyltransferase B (40)	Arabinosyltransferase C (5)
site-1	6	14	3	0	0	0
site-2	3	3	6	0	0	0
site-3	44	109	42	9	40	5
Mild mutations (185)					
	Catalase-peroxidase (40)	Pyrazinamidase (31)	β-subunit of RNA polymerase (11)	Arabinosyltransferase A (6)	Arabinosyltransferase B (78)	Arabinosyltransferase C (19)
site-1	11	3	1	0	0	0
site-2	0	0	0	0	0	0
site-3	29	28	10	6	78	19
Neutral mutations (12)					
	Catalase-peroxidase (4)	Pyrazinamidase (1)	β-subunit of RNA polymerase	Arabinosyltransferase A	Arabinosyltransferase B (7)	Arabinosyltransferase C
site-1	0	0	0	0	0	0
site-2	0	0	0	0	0	0
site-3	4	1	0	0	7	0

3.5 Chapter Summary

Drug-resistant TB is a severe threat globally. Understanding drug-resistant mutations for their prevalence, global spread, drug resistance mechanisms and impacts on the drug and *Mycobacterium tuberculosis* is a challenge. This study aims to better understand the prevalence, regional differences, mutation diversity with different hotspot positions, and impact of drug-resistant mutations on *Mycobacterium tuberculosis* and first-line TB drugs in terms of functional change, stability change and sequence conservation of the target and changes in drug binding to its target protein. For this, we collected mutational data from 31,073 *Mycobacterium tuberculosis* isolates published in 149 studies over the last 30 years.

From our 149 studies, we found 821 non-synonymous drug resistant mutations in first-line drug targets (*katG*-202, *pncA*-273, *rpoB*-120 and *embCAB*-226). For these, we calculated single mutation frequencies for each substitution for a better understanding of the prevalence and diversity of mutations in first-line TB drug targets. The pattern of mutations for the first-line TB drugs appeared to be different for each drug target with respect to the frequency of mutations, their positions, the overall spread of mutations in the target and hotspot positions (various amino acid substitutions at a single position).

In *katG,* we found that two positions were frequently mutated, with S315 (80.4%) being the most commonly mutated position; S315T was highly prevalent (60.58%) followed by R463 (7.25%) among the INH resistant isolates. For *rpoB,* most of the mutations were present in the rifampicin-resistance determining region (RRDR), with S450 (58.83%) being the most frequently mutated, along with H445 (19.63%), the second most common, occurring in RIF resistant isolates. M306 and G406 were two commonly mutating positions of *embB* showing EMB resistance. For *embC* and *embA*, studies did not provide significant mutation data. Unlike the other first-line targets, mutations in *pncA* were highly scattered, with W68, Q10 and H57 being the most frequent among all the other mutations (as shown in Figure 3.7).

Mutations in all four targets having the highest single mutation frequency were *katG* S315T (60.58%), *rpoB* S450L (56.70%), *pncA* Q10P (1.46%) and *embB* M306V (29.18%). *katG* and *rpoB* have one very frequent mutation and both have a large number of rare mutations scattered across the targets. These drugs are in both phases of treatment, and prolonged exposure to them seems to have enabled *Mycobacterium tuberculosis* to acquire many mutations. *embCAB* has a large number of low-frequency mutations spread across the target. *pncA* has a very low frequency (less than 2%) of mutations scattered across the region. These two drugs are only used in the first phase of treatment, and *Mycobacterium tuberculosis* might not produce highly frequent mutations in these two in the shorter

period of exposure to them. We also found that the frequent mutations had undergone frequent amino acid substitutions. When we looked closely at these sites, we found that most of these mutations were in non-conserved sites in the target protein except for *pncA* W68 and Q10 and *rpoB* H445. This means that *Mycobacterium tuberculosis* usually escapes from drug actions by mutating more variable regions compared to conserved.

Looking at the geographical occurrence of the mutations, we found a higher mutational burden of *katG* and *rpoB* compared to *embCAB* and *pncA*. Higher mutational frequency can be due to prolonged exposure (over six to twelve months) to INH and RIF during TB treatment. When we looked deeply into mutation patterns, spread, and frequency, we found that the most commonly occurring mutations were highly prevalent in all six WHO regions. However, their frequency varied in each region. For example, *katG* S315T and *rpoB* S450L were the most highly occurring mutations in all WHO regions. However, the total number of mutations was high in the Western Pacific and South-East Asia compared to the other four regions. The environmental and economic conditions and poor TB diagnosis might be the reason behind the differences in frequency of mutation count in different geographic areas (WHO Global Tuberculosis Report, 2020). Thus, *Mycobacterium tuberculosis* has followed a generic pattern across the globe over the last 30-35 years to develop resistance to drugs.

Drug resistant mutations are known to impair the growth and development of *Mycobacterium tuberculosis*. Information on the degree of impact on fitness is crucial for understanding the nature of mutations in different drug targets and for assessing how bacteria attain drug resistance. Mutations usually alter the drug targets' structural stability and biological function. Analysing the extensive data (821 drug resistant mutations) by wet-laboratory experiments would be difficult, time-consuming, and expensive. Thus, we performed a comprehensive analysis for predicting each mutation's impact on its respective protein using several bioinformatics tools relatively quickly with reliable accuracy to scrutinize the changes in sequence conservation, function, stability, and drug binding affinity in the mutant proteins.

Drug resistance is about *Mycobacterium tuberculosis* manipulating this arsenal of mutational capabilities to find a strategy that minimises its drug binding affinity with a minimum fitness cost. In the current study, we used ConSurf, SIFT, PROVEAN, PolyPhen-2, I-MUTANT 3.0 and mCSM (see Method section) that use information on sequence conservation and physiochemical or structural properties for making predictions on whether *Mycobacterium tuberculosis* employs mutations in conserved regions and predict functional and stability changes resulting from mutations. We used AutoDock for calculating drug binding affinity. Use of a number of prediction tools together reduces the chance of errors and provides accurate results.

To our knowledge, this is the first study using systematic sorting and comprehensive *in silico* analysis of 821 non-synonymous mutations in first-line drug targets, based on five crucial factors- sequence conservation, distribution of mutations into three sites, function, structure stability and drug binding, to probe into drug resistance mechanisms and *Mycobacterium tuberculosis* strategies for survival. Previous studies have focused on few mutations in studies of drug resistance. From our analysis, we found a much larger number of mutations and that: (i) Out of 821 mutations, 66.5% were present in the highly conserved sites, but when we looked at their frequency of occurrence, we found that they were infrequent (<9% in frequency). In contrast, frequent mutations (> 60% in frequency) were at variable (non-conserved) sites, (ii) in *katG* S315T was the only frequent mutation in site-1, whereas, in *rpoB, pncA* and *embCAB,* the most commonly occurring mutations were in site-2 and -3, (iii) 507 mutations out of 821 were found to be affecting significant functional changes leading to these mutants acquiring drug resistance, (iv) Around 80% of 821 mutations had a destabilizing impact with a small proportion of mutations improving structural stability, and (v) More than 85% of 821 mutations showed reduced drug binding affinity.

Based on the above findings incorporating the five important factors mentioned and the level of harm caused by *Mycobacterium tuberculosis* drug resistant mutations, we categorized mutations into four ranks - lethal, moderate, mild and neutral. Out of 821 non-synonymous mutations, we identified 340 'lethal,' 284 'moderate,' 185 'mild' and 12 'neutral' mutations. Out of the four first-line TB targets, three targets were required for survival of *Mycobacterium tuberculosis,* i.e., *katG* (catalase-peroxidase), *rpoB* (β-subunit of RNA polymerase) and *embCAB* (arabinosyltransferase CAB). While ranking drug the resistant mutations, we observed that the highly prevalent mutations in these three drug targets had mild to moderate impacts on drug binding with reduced drug binding energies, changes in enzymatic activity and low steric hindrance caused by structural changes; for example, the S315T mutation of *katG*, the S450L mutation of *rpoB* and the M306V mutation of *embB* were found to be highly frequent across the globe but had only a mild impact on their respective proteins. The number of lethal mutations was high for *katG* and *rpoB*, but these mutations were less prevalent.

This shows that *Mycobacterium tuberculosis* prefers mild mutations in targets essential for its survival in developing resistance to first-line TB drugs. On the other hand, *pncA* (pyrazinamidase), which is not essential for the survival of *Mycobacterium tuberculosis*, was found to have frequent lethal mutations. There can be several reasons for the occurrence of lethal mutations - the health of a patient, living conditions, severe antibiotic pressure and errors in DNA replication (McGrath et al., 2014).

Our comprehensive analysis for determining the prevalence and ranking of mutations based on its effect on *Mycobacterium tuberculosis* has improved our knowledge of the survival strategies used by this bacterium to maintain its fitness. Our study found that *Mycobacterium tuberculosis* follows some evolutionary pathways to balance the harmful impact of drug-resistant mutations on itself against drug resistance. Some literature studies have provided information regarding compensatory mutations occurring in *Mycobacterium tuberculosis* to compensate for fitness cost; for example, the fitness cost of RIF-resistant mutation, S450L, in the *rpoB* gene was overcome by a compensatory mutation in the *rpoA* and *rpoC* genes (Comas et al., 2012). The mutation leading to the overexpression of the *ahpC* gene compensates for S315T *katG* gene mutation (Sherman et al., 1996). We found that most frequent mutations had followed a generic pattern over the last three decades, suggesting that *Mycobacterium tuberculosis* will also follow a generic pattern across the globe in the future.

Drug resistance mechanisms and survival strategy of *Mycobacterium tuberculosis* found in this study can greatly contribute to eradicating TB globally. It can help develop effective and personalised treatment plans, develop new drugs and repurpose existing drugs for the frequent mutations worldwide. Our methods can be used to train new computational models for predicting positions in proteins with a higher tendency for acquiring new mutations and their consequences. Better insights into drug resistance mechanisms will aid in developing novel diagnostic tools that can help in the early diagnosis of drug resistance TB, reducing the transmission rate and planning proper effective treatment for the patients. When designing new drugs, our method can help predict the impact of mutations on their respective drug targets for developing better drugs for TB treatment. The ranking of mutations into four different categories can assist in developing inhibitors for a specific mutation or group of mutations and help develop personalized treatment plans for TB patients. In summary, this study provides an in-depth understanding of the impact of mutations at the evolutionary, functional and structural levels. The method developed in this study can also help in studying future mutations and there is the scope for introducing new steps in the method for improvement. Further, our method can be used to study the nature and impact of mutations in other infectious diseases.

Chapter 4

Developing and testing a conceptual and computational framework towards an effective human TB vaccine

Chapter-4 of aims to accomplish objective-2 of our study. In chapter-2, we discussed the host's immune response against TB bacteria and the survival strategies of the *Mycobacterium tuberculosis* against immune response. We also discussed the available TB vaccine, its lack of efficacy and the challenges in vaccine development. In this chapter, we developed a framework based on a new concept introduced in this research for TB vaccine development after gaining a deeper understanding from chapter-2. We tested the framework using bioinformatics tools to provide a potential vaccine solution to reduce the burden of TB worldwide. An overview of the chapter is provided in section 4.1. In section 4.2, the gaps in TB vaccine research are discussed. The formulation of the framework for TB vaccine development is discussed in section 4.3. The step-by-step process of the computational framework developed, and the bioinformatics tools used in each step, are explained in section 4.4. The results (effective vaccine candidates) are discussed in section 4.5. Finally, a summary of the chapter is presented in the last section (Section 4.6).

4.1   Overview

Vaccines are cost-effective pharmaceutical products that play an essential role in the elimination and eradication of infectious diseases. Despite the massive success of vaccines, new vaccines are still needed for re-emerging and drug-resistant pathogens. The objective of the present study is to identify potential vaccine candidates against *Mycobacterium tuberculosis*. The World Health Organization considers TB a global threat with significant mortality and morbidity rates. BCG, the only licensed vaccine available, and prepared from a live-attenuated strain of *Mycobacterium bovis,* has shown protection for babies and young children. The inefficiency of BCG in neither reducing the prevalence of the disease nor protecting adults is so far not understood. Some important factors might include *Mycobacterium bovis* being less virulent and not a primary causative agent of TB, diversity in TB strains and over-attenuation of the presently used BCG strain. For complete elimination and eradication of TB globally, an effective and powerful vaccine is needed for curtailing the dissemination and transmission of tuberculosis.

The conventional vaccine development and production process is very costly, and it takes many years to develop effective vaccines. Advances in genomics, transcriptomics, and proteomics have reduced the time and cost of vaccine development by focusing mainly on selecting novel antigens or epitopes.

One of the important advancements in vaccine development is computational vaccinology that uses different bioinformatics approaches, such as comparative genome analysis, reverse vaccinology, immunoinformatics and structural vaccinology for identifying potential vaccine candidates. Reverse vaccinology identifies the outer-membrane antigenic proteins with unique characteristics, such as signal peptides, membrane-spanning regions, lipoprotein signatures, adhesion probability, motif attachment from the genome or proteome a single pathogen or pathogens. Immunoinformatics performs computational analysis of antigenic proteins for the prediction of T-cell and B-cell epitopes. Epitopes comprise a short stretch of amino acids that are recognised by B-cell and T-cell receptors and specific antibodies. The vaccine targets identified are expected to evoke a specific immune response to provide broad immune protection against a large number of *Mycobacterium tuberculosis* strains. Our study used computational vaccinology on a large spread of strains to develop an epitope-based vaccine that is expected to be effective and safer against tuberculosis.

This study is designed to address the challenges of conventional vaccine development that include: expensive, time-consuming and arduous experimental testing; safety concerns while culturing the pathogen in a laboratory; identification of surface exposed, secreted, adhesin proteins; identification of conserved epitopes in highly variable or drug-resistant *Mycobacterium tuberculosis*; identifying immunodominant epitopes for inducing a potent humoral and cell-mediated immune response; elimination of cross-reactive epitopes; and immunogenicity assessment of the selected targets. Unlike traditional methods, computational vaccinology saves time and costs by minimising repeated laboratory testing.

4.2 Gaps in TB vaccine research

BCG, a hundred-year-old vaccine, has failed to provide complete protection against *Mycobacterium tuberculosis*. The formulation of the live-attenuated vaccine, BCG, involves the cultivation and attenuation (reducing the virulence) of *Mycobacterium bovis* using various *in vitro* techniques in the laboratory. The protection by BCG is highly variable in adults and the prevention of chronic infection is not high. The reasons for such high variability of BCG vaccine include differences in clinical assays, genetic variability in a sample population, different levels of protection against the clinical forms of tuberculosis, malnutrition and variability in *Mycobacterium tuberculosis* strains (Barreto et al., 2006; Doherty & Andersen, 2005).

Currently, there are 14 vaccines in different stages of clinical trials. Out of the 14, seven are protein and viral vectored vaccines. The remaining seven vaccines are live-attenuated and heat-killed vaccines. The main limitation of live-attenuated and inactivated or heat-killed vaccines is the maintenance of purity and safety while culturing the microorganism in the laboratory. A major drawback of live

attenuated vaccines is the reversion of the pathogen into its original pathogenic form (Versteeg et al., 2019). The subunit TB vaccine developed using the conventional approach contains one or more antigenic TB proteins for initiating the immune response in the host. The traditional method requires cultivating pathogens in the laboratory and then performing different biochemical, microbiological, immunological tests to detect the vaccine candidates. This approach is laborious, requires expensive experimental testing and sometimes fails to reveal suitable antigens. Rodo et al., (2019) used clinical data for six subunit TB vaccines (MVA85A, ID93+GLA-SE, AERAS-402, H56:IC31, M72/AS01E and H1:IC31) and found that only M72/AS01E provided protective immunity against TB. This study suggested the need for a new and improved vaccine with a more specific immune response. The vaccines in the clinical trial mainly depend on inducing an immune response using either weak mycobacterial strains or using two to three antigens to produce recombinant subunits and viral vectored vaccines. These vaccines do not guarantee to yield a broad coverage immune response. The drug-resistant issue is also not adequately addressed by any of the fourteen vaccines in the clinical trial.

Current vaccine development approaches focus on enhancing cell-mediated immunity by using one or more antigens and does not consider global coverage. Some computational studies have also been performed to address the vaccine problems for TB disease (Hossain et al., 2017; Monterrubio-López et al., 2015; Ortega-Tirado et al., 2020). Monterrubio-López et al. (2015) applied the reverse vaccinology approach to the proteome of *Mycobacterium tuberculosis* H37Rv (Figure 4.1). They identified six antigenic proteins as potential vaccine candidates PBP1, PPE65, EsxL, PE26, PE PGRS49, and Erp (Monterrubio-López et al., 2015). Hossain et al. (2017) identified T-cell epitopes present on the extracellular protein 85B of *Mycobacterium* spp. The PE_PGRS33 protein was used to identify T-cell epitopes by Ortega-Tirado et al. (2020). Both these studies used only one protein for recognising promiscuous T-cell epitopes. The major drawback of the above-described studies was that they did not provide a solution to pathogen polymorphism, allergenicity and genetic variability among the hosts.

A well-designed epitope-based TB vaccine based on an understanding of the tuberculosis disease and the above concepts can become a powerful therapeutic tool against *Mycobacterium tuberculosis.* Developing a TB vaccine using the conventional vaccine development process would be time-consuming and expensive. The purification and attenuation of TB vaccine products in the laboratory is arduous and leakage of *Mycobacterium tuberculosis* is always a risk in the laboratory. The conventional vaccine development works only with one or few strains yielding limited coverage. Therefore, a vaccine developed based on many strains with a broad coverage and based on a deep understanding of host-pathogen interaction can greatly contribute to eliminating TB globally or drastically minimise its prevalence.

Our study aims to provide a potential solution to reduce the burden of tuberculosis disease globally. With this aim, we developed a conceptual and computational framework for designing a universal vaccine against *Mycobacterium tuberculosis.* We performed a holistic study to gain deeper insights into the pathogenesis of tuberculosis and the challenges in TB vaccine development and developed a method using computational vaccinology to identify potential vaccine candidates for TB. As we know, the current explosion in bioinformatics has revolutionized the field of vaccine development. Using a myriad of bioinformatics tools, the computational vaccinology approach facilitates identification of potential vaccine targets without culturing *Mycobacterium tuberculosis* in the laboratory. Also, information about the proteome of *Mycobacterium tuberculosis* can help identify novel vaccine candidates that are important for developing effective TB vaccines.
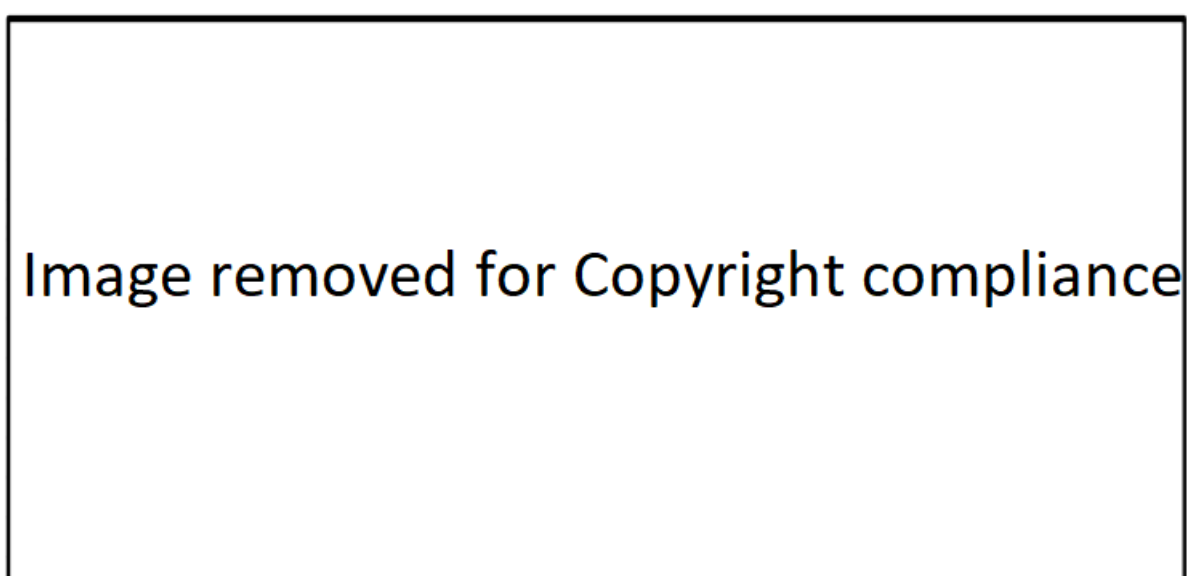
Figure 4. 1: Reverse vaccinology applied to *Mycobacterium tuberculosis* proteome by Monterrubio-López et al., 2015

4.3 Conceptual and holistic framework designed for identifying TB vaccine candidates

Epitope-based vaccines are more potent when controlled properly, and they induce a specific immune response to a broad range of immunodominant epitopes, target multiple-conserved epitopes and disrupt tolerance. An epitope-based vaccine development helps select vaccine candidates that are not harmful when following the safety measures of the laboratory. In this study, we proposed a conceptual framework for developing an epitope-based TB vaccine providing broad spectrum of protection against many *Mycobacterium tuberculosis* strains and drug resistance. We addressed most of the challenges in conventional TB vaccine development using computational vaccinology. Thus, we developed a framework for designing a TB vaccine aimed to evoke a strong humoral and cell-mediated immune

response in humans. By bridging the gaps in vaccine development, a powerful, effective and universal TB vaccine can be developed that:

- brings down the incidence rate of infection
- eradicates latent infection
- quells the spread of drug-resistance pathogens
- prevents and protects from the overuse of antibiotics
- elicits a specific and swift immune response, and
- averts autoimmunity or hypersensitive reactions.

4.3.1 Computational vaccinology

The use of bioinformatics tools in biotechnology, immunology and vaccinology has developed an innovative field known as computational vaccinology (Vivona et al., 2008). Computational vaccinology performs *in-silico* analyses to identify novel vaccine candidates within a pathogen's genome (He et al., 2010). The aims of computational vaccinology include minimising the number of targets, identifying protective antigens that can be safely expressed in the laboratory, and reducing the number of experimental tests and the costs required to validate vaccine targets. Computational vaccinology uses reverse vaccinology and immunoinformatics to identify potential vaccine candidates, aiming to develop an effective vaccine (Figure 4.2). The main advantage of computational vaccinology is that there is the scope for improving the prediction of vaccine candidates by performing several additional steps.
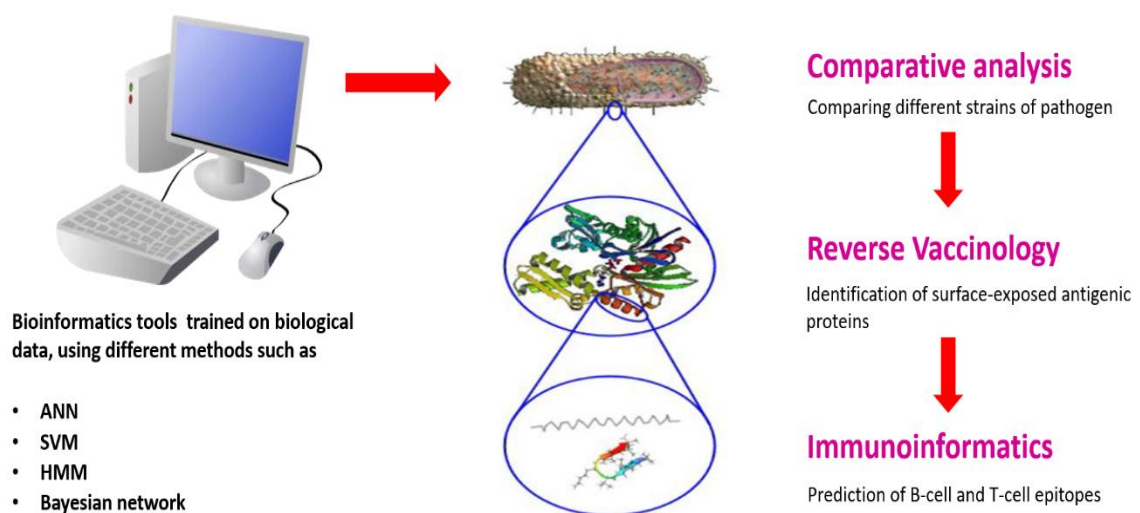
Figure 4. 2: Generic computational vaccinology approach to identify vaccine candidates

4.3.1.1 Reverse vaccinology

The availability of the genome sequence of disease-causing pathogens facilitates the prediction of antigenic proteins for vaccine development, without the need to culture them in the laboratory. As stated in Chapter-1, *in-silico* analysis of the entire genome of a pathogen to identify potential vaccine targets is known as 'reverse vaccinology' (RV) (Seib et al., 2012). Reverse vaccinology emerged as a novel approach that accelerates the vaccine design process and has revolutionised both immunology and biotechnology. The critical perspective of RV includes identifying vaccine targets that are either surface-exposed or secreted into the extracellular surrounding of a pathogen (Donati & Rappuoli, 2013). The unique features of the RV approach have attracted the attention of researchers developing vaccines against pathogens where the conventional method has failed to produce an effective vaccine (Grandi, 2001). The reverse vaccinology approach offers three significant advantages to researchers. First, it allows the discovery of a broad spectrum of vaccine candidates that have not been detected before; secondly, it enables the identification of potential candidates that are difficult culture in the laboratory (Donati & Rappuoli, 2013) and, thirdly, it allows the stimulation of a precise immune response.

Reverse vaccinology starts with a high-throughput *in silico* analysis of the genome of a pathogen using GLIMMER, GeneMark or ORPHEUS for identifying open reading frames (ORF) (Figure 4.3). The alignment of coding sequences using homology search tools, such as BLASTN, BLASTX, TBLASTX (Altschul, 1997; Altschul et al., 1990), allows the functional annotation of genes and helps in the identification of the conserved sequences among different isolates of a pathogen (Movahedi & Hampson, 2008).

The outer-membrane proteins are considered potential vaccine targets because these proteins contain specific molecular patterns recognised by the host's immune system (Zagursky & Russell, 2001). The outer-membrane proteins can be identified using two different approaches. First, an integrated search in protein-domains and family databases, such as Pfam, PROSITE, Interpro Scan and ProDom, can help identify sequence motifs, which can, therefore, determine the surface-exposed proteins. Secondly, subcellular localisation prediction tools help predict the location of proteins in the cell (Bulashevska & Eils, 2006). The surface-exposed proteins are essential for interaction with host cells because they (Bhavsar et al., 2007; Simeone et al., 2009; Stavrinides et al., 2007)

- adhere to host cell
- damage host tissue
- resist host environmental stress
- subvert host's immune response

Several bioinformatics tools are available for the identification of surface-exposed or secreted proteins: the subcellular localisation of proteins can be determined using PSORTb 3.0 (Yu et al., 2010), CELLO (Yu et al., 2006), LocTree3 (Goldberg et al., 2014) and SOSUI (Imai et al., 2008). Outer membrane proteins with more than one transmembrane region are not considered good vaccine targets because they are hard to clone. The membrane-spanning regions are identified with the help of HMMTOP (Tusnady & Simon, 2001) and TMHMM (Krogh et al., 2001). Adhesion probability is also an essential aspect of identifying potential candidates as adhesins help the bacteria colonise. Adhesion probability can be predicted using SPAAN (Sachdeva et al., 2005). Adhesin proteins aid in the attachment of pathogens to the host's cell-surface receptors.
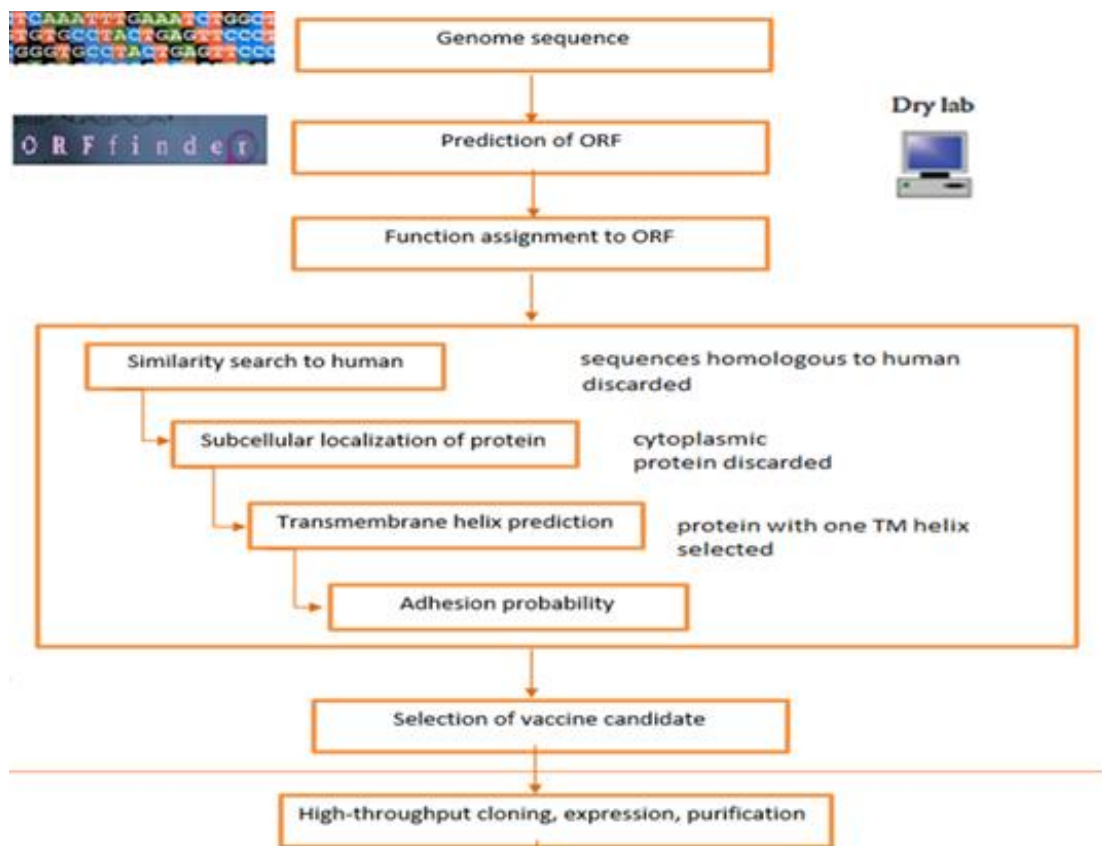
Figure 4. 3: A classical reverse vaccinology approach for identifying vaccine candidates

RV was used to analyse Men B (*Neisseria meningitidis* serogroup B), a causative agent of meningitis. The whole genome of *N. meningitidis* was analysed using reverse vaccinology to predict surface-exposed proteins that can act as potential vaccine candidates (Pizza, 2000). Five antigen candidates were selected to produce the multicomponent meningococcal serogroup B vaccine (4CMenB). 4CMenB has completed clinical trials (Donati and Rappuoli, 2013) and is approved for 30 countries (Doolan et al., 2014). In classical reverse vaccinology, only one genome is analysed for vaccine targets

(Figure 4.3). However, a single isolate of a pathogen fails to explain the genetic variability among different strains. This is a limitation of RV.

In our study, we will be using comparative proteomic analysis to address the limitation of RV. Comparative proteomic analysis helps in identification of conserved vaccine candidates that are broadly effective on different *Mycobacterium tuberculosis* strains and provide protection against drug resistance.

4.3.1.2 Immunoinformatics

Antigenic determinants, specific segments of an antigen, are known as epitopes. Epitopes comprise a short stretch of amino acids that are recognised by B-cell and T-cell receptors and specific antibodies. The epitopes are usually present on the surface of a protein and are more flexible than the rest of the protein. Multiple epitopes can be identified within a single antigenic protein. The major challenge faced in stimulating an effective immune response is the accurate prediction of the immunogenic site of an antigenic protein (You et al., 2010).

The availability of the genome sequence of a pathogen has advanced our understanding of immunology. The development of bioinformatics tools, and the analysis of antigenic proteins from reverse vaccinology that has helped the emergence of a new paradigm known as 'immunoinformatics'. The computational analysis of antigenic proteins for the prediction of T-cell and B-cell epitopes is defined as immunoinformatics. The field of immunoinformatics has advanced the research and discovery of subunit vaccines. It has also enhanced our knowledge of handling immunological data appropriately. The immunoinformatics research includes epitope mapping, designing epitope databases, virulence identification and prediction of MHC classes-I and -II binding HLA alleles. Immunoinformatics has attracted the attention of researchers working in the field of vaccine discovery and development.

Hossain et al. (2017) and Ortega-Tirado et al. (2020) used one protein for recognising promiscuous T-cell epitope and did not focus on crucial features of an epitope-based vaccine such as sequence conservancy, immunodominance, antigenicity, exclusion of self-peptides and multiple allelic interactions. The epitope-based vaccines comprising immunodominant epitopes helps in eliciting a swift and specific immune response. The emergence of new bioinformatics tools for predicting epitopes with high accuracy has reduced the time and costs of developing an epitope-based vaccine. Epitope-based vaccines are stable, specific and free from infectious agents. An epitope-based vaccine is anticipated to induce a robust and prolonged immune response against pathogens.

Epitopes are classified into B-cell epitopes and T-cell epitopes (Almeida et al., 2012).

(i) **B-cell epitopes** are the part of the antigen recognised by the B-cell receptor (BCR) and antibodies. They have been classified into linear or continuous epitopes, and conformational or discontinuous epitopes (Flower et al., 2010). A linear epitope is a short stretch of the peptides of a protein. In contrast, the conformational epitopes comprise amino acids folded in a tertiary structure (EL-Manzalawy & Honavar, 2010). The tools available for predicting B-cell epitopes are Bepitope (Larsen et al., 2006), ABCpred (Saha & Raghava, 2006b) and BCPREDS (Chen et al., 2007).

(ii) **T-cell epitopes** are fragments of the antigenic proteins of a pathogen formed in the host antigen-presenting cells (APC) by the proteolytic activity of proteasomes or endosomes, that are then exposed to the surface of APC with the help of MHC molecules, followed by interaction with TCR (Figure 4.4) of T cells leading to activation of T-cells for releasing cytokines (Sobolev et al., 2005). T-cell epitopes play an essential role in cell-mediated immunity against *Mycobacterium tuberculosis*. The most crucial problem in designing an epitope-based vaccine is identifying T-cell epitopes that can stimulate the T-cells of the host immune system (Schirle et al., 2001). T-cell activation is dependent on the presentation of epitopes by MHC molecules. MHC molecules are glycoproteins present on the antigen presenting cell's surface, and their main task is to present epitopes to the TCR of T-cells. Thus, epitopes are predicted based on their ability to bind with MHC class-I and class-II molecules, found in cytotoxic and helper T-cells, respectively. There are different tools available for the prediction of T-cell epitopes. Among the most popular tools for predicting MHC class-I binding T-cell epitopes are IEDB MHC-I (Zhang et al., 2008), NetCTLpan 1.1 (Stranzl et al., 2010), IEDB MHC-NP (Giguère et al., 2013), SYFPEITHI (Rammensee et al., 1999) and RANKPEP (Reche et al., 2004). The tools for MHC class-II are NetMHCIIpan 3.2 (Jensen et al., 2018), IEDB MHC-II (Wang et al., 2010), ProPred (Singh & Raghava, 2001) and TEPITOPE (Zhang et al., 2012).
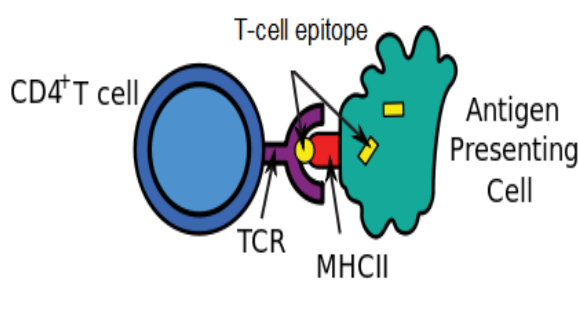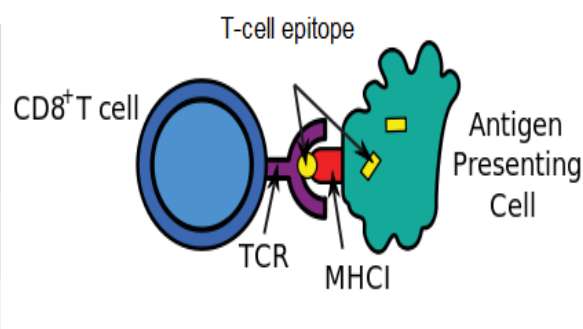
Figure 4. 4: T-cell epitopes binding to MHC-I and II molecules. Figure created using template from Biorender (https://biorender.com/)

4.3.2 Application of computational vaccinology to develop a conceptual and holistic framework for TB vaccine design

TB prevention is a long-term goal that can only be accomplished by developing a more effective vaccine than BCG. The current therapeutic and diagnostic strategies are inadequate to eliminate TB by 2050 (Ottenhoff & Kaufmann, 2012). At this juncture, new approaches for vaccine development must focus on host-pathogen interactions, the limitations of BCG, induction of a specific immune response that can prevent the reactivation of latent tuberculosis, prevent the spread of MDR-TB and XDR-TB (Ahsan, 2015), and provide broader coverage of immune response.

An extensive study of host-pathogen interactions, the distinctive strategies used by *Mycobacterium tuberculosis* for escaping host responses, the global pattern of drug resistant mutations, the limitations of BCG, challenges in developing an effective vaccine and the advantages of application of computational techniques in vaccine design helped us to create a conceptual framework for TB vaccine design. In our study, we used some crucial concepts for designing and developing an effective and robust TB vaccine compared to the conventional TB vaccine development approach as follows:

(1) *Mycobacterium tuberculosis* **genetic variability**: The vaccine prepared using a single strain or single antigen of *Mycobacterium tuberculosis* does not provide effective immunity to the host. Genetic variability among different strains of the immune evasive *Mycobacterium tuberculosis* increases the chances of drug resistance. Around 0.3 million people die each year from drug-resistant TB infections. Its high variability is one of the fundamental reasons behind the non-availability and less effective vaccines against tuberculosis. The practical solution to the TB bacteria polymorphism issue is to use highly conserved vaccine targets from the proteome of

Mycobacterium tuberculosis for developing a universal TB vaccine. Targeting highly conserved regions for their significant structural and functional roles in the life cycle of the pathogen would provide broad spectrum of protection against many *Mycobacterium tuberculosis* strains and drug resistance.

(2) **Identification of proteins involved in pathogenesis**: The proteome of *Mycobacterium tuberculosis* consists of proteins involved in the virulence, the progression of TB and escape from the host immune response. Some examples are secretory proteins, lipoproteins, PE (proline-glutamate) and PPE (proline-proline-glutamate) proteins. The secretory signal peptides are ubiquitous protein-sorting signals that help translocate a protein across the cytoplasmic membrane to cell wall and extracellular space in mycobacteria. Studies have shown that the secretory systems of *Mycobacterium tuberculosis* are involved in the virulence mechanism (Abdallah et al., 2007). There are two important secretion systems in *Mycobacterium tuberculosis*: the secretory pathway (Sec pathway) and the twin-arginine translocase pathway (TAT pathway). The Sec pathway helps in the translocation of proteins in their unfolded state. The Tat pathway translocate proteins in their folded state. The proteins involved in the TAT pathway help in the translocation of the phospholipase C enzyme in *Mycobacterium tuberculosis*. This enzyme exhibits a cytotoxic effect on the alveolar macrophage of the host (Assis et al., 2014).

Lipoproteins are an essential set of membrane proteins performing different functions in *Mycobacterium tuberculosis*, such as modulating the inflammatory response, translocating virulent factors in the macrophages, showing resistance to drugs, signal transduction, and the uptake of nutrients (Kovacs-Simon et al., 2011). These proteins constitute 2.5% of the predicted proteome of *Mycobacterium tuberculosis* and represent a significant virulent protein family (Sutcliffe & Harrington, 2004). Proteins belonging to the PE and PPE protein families present in the cell wall and cell membrane (Banu et al., 2002), are the main reason for developing antigen variability in *Mycobacterium tuberculosis* (Smith, 2003).

(3) **Surface-exposed protein as vaccine target**: As described above, most of the proteins involved in pathogenesis of TB are present in the cell membranes or cell walls. The surface proteins are easily accessible to the host immune system. For example, the extracellular proteins of *Mycobacterium tuberculosis* are involved in interactions with toll-like receptors (TLR-2 and TLR-4) present on the macrophages. Thus, surface-exposed proteins are potential vaccine targets to weaken *Mycobacterium tuberculosis* and eliminate TB infection.

(4) Heightened immune response: The triggering of a robust immune response is the most crucial challenge in vaccine development. The conventional TB vaccine approach, focused on the helper T-cell immune response, has not shown protective immunity against *Mycobacterium tuberculosis* (Rodo et al., 2019). Thus, an epitope-based TB vaccine containing helper-T cell, cytotoxic-T cell and B-cell epitopes can stimulate a specific cell-mediated and humoral immune response in the host to eliminate the chances of infection and re-infection of tuberculosis.

(5) Careful selection of vaccine candidates: Various studies have shown that all predicted T-cell epitopes are reliable MHC class-I and -II binders, but not all MHC binding epitopes are T-cell epitopes (Patronov & Doytchinova, 2013). Therefore, careful selection of the immunodominant T-epitopes, and determining their binding affinity with MHC molecules, is an extremely important step in vaccine design and development.

(6) Genetic variability of population: Another important concept in developing an effective TB vaccine is broad coverage of the human population. In most cases, a single strain or single antigen vaccine helps treat a small subset of patients in a specific region. The vaccine is then sometimes considered ethnically-biased in terms of protection. MHC molecules are highly polymorphic and more than a thousand HLA alleles for humans have been identified. The different HLA types express at different rates in different ethnicities around the world. Thus, the selection of T-cell epitopes binding to MHC HLA alleles which are most prevalent in the human population will help in designing a universal vaccine with optimal efficacy against TB.

(7) Safer vaccine candidates to reduce the probability of side effects: An epitope-based vaccine is safer than live-attenuated or heat-killed vaccines as they only contain fragments from antigenic proteins and not the live components of the pathogen. Thus, there is no risk of the reversion of an epitope-based vaccine to harmful elements. The elimination of cross-reactive and toxic epitopes that cause autoimmunity or hypersensitive reactions in the host makes the vaccine safer.

(8) Immunogenicity enhancement: The use of adjuvants in vaccines heightens the immune response (Pasquale et al., 2015). An adjuvant increases the vaccine's immunogenicity and helps induce a robust and long-lasting immune response in immunocompetent individuals.

We purposefully designed a reductionist pipeline method for developing an epitope-based TB vaccine from the proteome of *Mycobacterium tuberculosis* using the aforementioned concepts. The proposed method involves six phases (Figure 4.5). The first phase involved comparative proteomic analysis to

identify conserved proteins across 159 *Mycobacterium strains* and functional categorisation of these conserved proteins. In the second phase, reverse vaccinology was used on conserved proteins to analyse the surface-exposed, antigenic, non-allergic proteins involved in the pathogenesis of TB that were easily accessible to surveillance by the host immune system. In phase three, B-cell and T-cell epitopes were predicted using the immunoinformatics approach. T-cell epitope selection to a broad range of MHC HLA alleles is crucial for global coverage of the human population. The fourth phase comprised filtering the immunodominant epitopes that can induce protective and robust immunity in the host. The B-cell and T-cell epitopes predicted by the fourth phase are used for generating the final vaccine construct. In the fifth phase, using the structural vaccinology technique, the structural model of the new TB vaccine, with the adjuvant attached to it, is generated. The binding affinity of the TB vaccine with toll-like receptors in the host antigen presenting cell helps determine the behaviour of the vaccine inside the host body. The sixth phase involved analysing the immune response generated by the designed epitope-based TB vaccine using the C-ImmSim server. This step determines the efficacy of the developed TB vaccine inside the human body.

The details of the methodology and bioinformatics tools used in each step can be found in the next section.
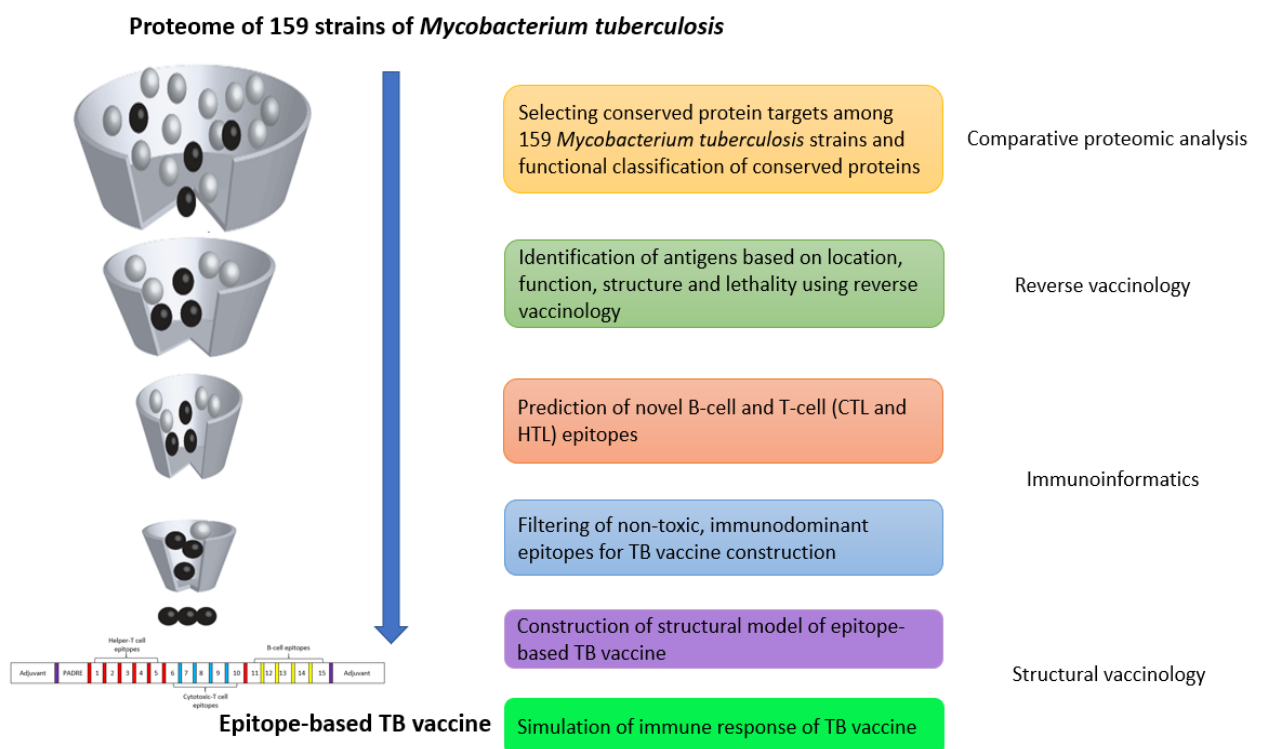
Figure 4. 5: Conceptual framework for constructing an epitope-based TB vaccine against *Mycobacterium tuberculosis* using comparative proteomic analysis, reverse vaccinology, immunoinformatics and structural vaccinology

4.4 Testing the framework on *Mycobacterium tuberculosis* to identify potential TB vaccine candidates

Our study developed a conceptual framework for identifying potential T-cell and B-cell epitopes for developing a universal vaccine for tuberculosis. This study utilises sequence conservancy, antigenicity, exclusion of self-peptides and allelic interactions, and other essential factors. This section of the chapter describes the step-by-step approach to the framework developed. In this study, we used several bioinformatics tools to identify ideal vaccine candidates by directly analysing the proteome of *Mycobacterium tuberculosis* strains that assisted in developing an epitope-based vaccine that can stimulate a specific and swift humoral and cell-mediated immune response.

4.4.1 Comparative proteomic analysis

A large number, 159, *Mycobacterium tuberculosis* strains were used to identify the conserved antigenic vaccine targets to solve the problem of antigen variability, drug resistance and limited coverage. These antigenic proteins can help in developing a universal vaccine for TB that is effective and harmless to the host.

4.4.1.1 Data retrieval and formatting

The proteome of the completely sequenced 159 strains of *Mycobacterium tuberculosis* was downloaded in FASTA format via the NCBI (**N**ational **C**enter for **B**iotechnology **I**nformation) Genome FTP site. The unwanted information and blank spaces within protein sequences were removed. The protein sequences in multiline FASTA format were converted into a single-line format.

4.4.1.2 Identification of conserved proteins

The **b**asic **l**ocal **a**lignment **s**earch **t**ool (BLAST) program performs sequence alignment for screening homologous sequences among similar or different species (Altschul, 1997; Altschul et al., 1990). Standalone BLAST was downloaded from the NCBI FTP site for performing a homology search among *Mycobacterium tuberculosis* strains. The selection of conserved proteins was made based on three criteria: (i) *Mycobacterium tuberculosis* H37Rv was used as reference proteome, (ii) only those proteins that were present in all 159 strains were considered, and (iii) protein sequences with more than 99% sequence similarity across all 159 strains were considered as conserved proteins and selected for further research.

4.4.1.3 Function assignment to selected conserved proteins

The resulting conserved proteins were categorised into eleven distinct functional categories given by Smith (2003). The eleven functional categories given by Smith are lipid metabolism; information

pathways; cell wall and cell processes; stable RNAs; IS elements and bacteriophages; PE and PPE proteins; intermediary metabolism and respiration; regulatory proteins; virulence, detoxification and adaptation; conserved hypothetical function; and proteins of unknown function. There are hypothetical proteins (HPs) as well as proteins with unknown functions in Smith's classification. These proteins might be involved in some essential metabolic activity or growth and development or pathogenicity. Some of the selected conserved proteins (from the previous step) might be HPs. Thus, it is crucial to predict the function of hypothetical proteins. In this step, we attempted to assign the functions to the HPs in two ways:

- domain prediction using Pfam, Interpro Scan and SMART (explained in the Materials and Methods section 3.3.4.2 of chapter-3), and
- homology search using BLAST

The analysis could provide a functional contribution of HPs in *Mycobacterium tuberculosis* and improve the understanding of the survival of *Mycobacterium tuberculosis*.

4.4.2 Identifying antigenic proteins using reverse vaccinology

Reverse vaccinology identifies outer-membrane antigenic and non-allergenic proteins that have unique characteristics, such as, the presence of signal peptides, membrane-spanning regions, lipoprotein signatures and adhesion probability from the proteome of *Mycobacterium tuberculosis.*

4.4.2.1 Subcellular localisation (SCL) prediction

The proteins present in the cell membrane and extracellular space are considered suitable TB vaccine candidates since they are easily accessible to the host's immune system. These proteins play an essential role in maintaining cell integrity. In this step, we used six SCL predicting tools: PSORTb v.3.0 (Yu et al., 2010), CELLO (Yu et al., 2006) and LocTree3 (Goldberg et al., 2014) that are the machine learning tools that use support vector machines (SVM) for SCL prediction; SOSUI (Imai et al., 2008) and pLoc_bal-mGpos (Xiao et al., 2019) that use protein sequence information, such as physiochemical properties for predicting SCL; and GramLocEN that uses a single and multi-label elastic net classifier to make SCL predictions (Wan et al., 2017). The proteins present in the cell membranes and extracellular spaces, predicted by four or more tools, were selected for further analysis.

4.4.2.2 Transmembrane helix prediction

Proteins with more than one transmembrane helix are difficult to clone and purify in the laboratory (Pizza, 2000). Thus, proteins having a single transmembrane helix were used to help design the vaccine. TMHMM, based on the hidden Markov model (HMM), allows the prediction of transmembrane regions and their orientation (Krogh et al., 2001).

4.4.2.3 Identification of functionally relevant proteins

The outer membranes and extracellular spaces of *Mycobacterium tuberculosis* consist of some essential secretory and lipoproteins that help in the translocation of a protein across the cytoplasmic membrane. In this step, we used the following tools for predicting the functionally-important proteins: (i) SignalP 4.1 server is a combination of several artificial neural networks (ANN) and predicts the presence and position of a signal peptide in the query sequence (Petersen et al., 2011); (ii) PRED-TAT is used for identifying TAT-signal peptides using HMM (Bagos et al., 2010); (iii) SecretomeP is an ANN-based method for predicting secretory proteins (Bendtsen et al., 2005), and proteins with a score of 0.5 or above were selected; and (iv) PRED-LIPO is an HMM-based method for predicting lipoproteins (Bagos et al., 2008).

4.4.2.4 Antigenicity prediction

The development of a vaccine for TB requires identification of antigenic proteins that can stimulate a precise immune response in the host. Thus, there is the need to identify proteins that have the capability of causing an infection in the host cell. We used three tools for predicting antigenic proteins. VaxiJen helps predict the antigenic properties based on the physiochemical properties of the protein (Doytchinova & Flower, 2007). VirulentPred is an SVM-based method (Garg & Gupta, 2008), whereas MP3 tool uses SVM and HMM (Gupta et al., 2014) to predict antigenicity within a protein. The proteins predicted as antigenic by all three methods were selected for further analysis.

4.4.2.5 Allergenicity prediction

The allergenicity determines whether the antigenic protein is allergic or non-allergic. Algpred is an SVM model that differentiates between allergenic and non-allergenic proteins based on the amino acid composition of protein(Saha & Raghava, 2006a).

4.4.2.6 Adhesion probability

Mycobacteria have adhesin proteins that help it attach to the host cell-surface receptors. The adherence of *Mycobacterium tuberculosis* to host cells generates membrane agitation and enables a robust interaction between *Mycobacterium tuberculosis* and the host (Vidal Pessolani et al., 2003). The adhesion probability of the protein can be predicted using an artificial neural network-based method known as SPAAN (**s**oftware for **p**redicting **a**dhesins and **a**dhesin-like proteins using **n**eural networks) (Sachdeva et al., 2005). The proteins with a probability score of 0.5 or above were selected for further analysis.

4.4.2.7 Interaction pathway analysis

Analysis of the protein-interaction pathways was performed to predict its functional significance in *Mycobacterium tuberculosis*. The protein-interaction network was analysed using the STRING database (Szklarczyk et al., 2019). Proteins having more than ten interacting partners with a confidence score of 0.4 or above were selected.

4.4.2.8 Homology to humans

For predicting the homology between human and *Mycobacterium tuberculosis,* a standalone BLAST was used to identify sequence similarity between them. Proteins with sequence similarity of more than 30% were eliminated from the research. The Kyoto Encyclopaedia of Gene and Genome (KEGG) (Kanehisa et al., 2017) was then used for identifying proteins involved in metabolic pathways common in humans and *Mycobacterium tuberculosis.* The proteins involved in the common pathways were removed. This step helps exclude proteins that can cause autoimmunity or any hypersensitive reactions in the host.

4.4.3   Prediction of B-cell and T-cell epitopes

The antigens identified using the reverse vaccinology approach were further analysed for predicting potential B-cell and T-cell epitopes for developing an effective vaccine against tuberculosis. To predict epitopes, datasets containing experimentally validated epitopes were used to train and test different machine learning models such as support vector machines, hidden Markov models and neural networks.

4.4.3.1 B-cell epitope prediction

B-cell epitopes play a vital role in initiating the humoral immune response. B-cell epitopes were predicted using three methods. The Bepipred tool of the **i**mmune **e**pitope **d**ata**b**ase (IEDB) uses a combination of a hidden Markov model and a propensity scale method for accurate prediction of linear B-cell epitopes (Larsen et al., 2006). ABCpred uses a recurrent neural network for predicting B-cell epitopes (Saha & Raghava, 2006b). BCPREDS is a server used for B-cell epitope prediction using the amino acid pair (AAP) antigenicity scale (Chen et al., 2007). The common epitopes predicted by all three methods were chosen.

4.4.3.2 T-cell epitope prediction

MHC molecules are an important class of proteins present on the surface of a cell and play an essential role in cell-mediated immunity. The important function of the MHC molecules is the presentation of fragmented or processed antigens to the appropriate T-cells of the immune system. The binding of a peptide with MHC molecule is essential for recognising an epitope by the T-cell receptors.

(i) MHC-I binding T-cell (cytotoxic-T cell) epitope prediction

MHC class I molecules are present on the surface of all nucleated cells in the human body. MHC class I molecules present the epitope to the cytotoxic T-lymphocyte (CTL). The triggering of CTL causes the APCs to undergo programmed cell death. Three methods were used in this step: IEDB MHC I (Zhang et al., 2008), NetCTLpan 1.1 (Stranzl et al., 2010) and IEDB MHC-NP (Giguère et al., 2013). In the IEDB MHC-I method, we used a reference file containing 27 MHC-I HLA alleles (HLA-A0101, HLA-A0201, HLA-A0203, HLA-A0206, HLA-A0301, HLA-A1101, HLA-A2301, HLA-A2402, HLA-A2601, HLA-A3001, HLA-A3002, HLA-A3101, HLA-A3201, HLA-A3301, HLA-A6801, HLA-A6802, HLA-B0702, HLA-B0801, HLA-B1501, HLA-B3501, HLA-B4001, HLA-B4402, HLA-B4403, HLA-B5101, HLA-B5301, HLA-B5701, HLA-B5801) for a broader human population coverage. The 9-mer epitopes with percentile rank lower than 0.5 were chosen. The lower value of the percentile rank indicates a higher affinity of the epitope towards the MHC molecule. NetCTLpan 1.1 uses an artificial neural network for predicting MHC-I epitopes. IEDB MHC-NP is a machine learning method that predicts naturally-processed peptides using MHC- I experimental data (Giguère et al., 2013). The epitopes common in all three prediction results were selected as T-cell MHC class I epitopes.

(ii) MHC-II binding T-cell (helper-T cell) epitope prediction

MHC class-II molecules are present on the surface of antigen-presenting cells (APC), such as macrophages and dendritic cells. The primary function of MHC-II is to present the antigen to the naïve T-helper cells. This interaction leads to the release of cytokines that help transform naïve-T- helper cells into effector T-cells or memory T-cells. We used 2 methods. The NetMHCIIpan 3.2 server supports the prediction of epitopes that bind to MHC class-II molecules. The method is based on a feed-forward artificial neural network that predicts all three MHC-II alleles HLA-DR, -DP, -DQ for humans covering 36 HLA-DR, 9 HLA-DP, 27 HLA-DQ molecules (Jensen et al., 2018). The second method used for prediction was the IEDB MHC-II binding server for predicting 15-mer MHC-II epitopes using the consensus prediction method (Wang et al., 2010). Thirty MHC-II HLA alleles used in the study are HLA-DRB1_0101, HLA-DRB1_0102, HLA-DRB1_0301, HLA-DRB1_0401, HLA-DRB1_0404, HLA-DRB1_0405, HLA-DRB1_0701, HLA-DRB1_0801, HLA-DRB1_0802, HLA-DRB1_0804, HLA-DRB1_0901, HLA-DRB1_1101, HLA-DRB1_1201, HLA-DRB1_1301, HLA-DRB1_1302, HLA-DRB1_1501, HLA-DRB1_1502, HLA-DRB3_0101, HLA-DRB4_0101, HLA-DRB5_0101, HLA-DQA1_0501/DQB1_0201, HLA-DQA1_0501/DQB1_0301, HLA-DQA1_0301/DQB1_0302, HLA-DQA1_04:01/DQB1_0402, HLA-DQA1_0101/DQB1_0501, HLA-DQA1_01:02/DQB1_06:02, HLA-DPA1_02_01/DPB1_01_01, HLA-DPA1_0103/DPB1_0201, HLA-DPA1_0301/DPB1_0402, HLA-DPA1_0201/DPB1_0501. The epitope with percentile ranks lower than 0.5 were chosen. The lower value of the percentile rank indicates a higher affinity of the epitope towards the MHC molecule. The epitopes commonly predicted by both methods were selected as T-cell MHC class II epitopes.

4.4.4   Filtering of epitopes

Filtering epitopes was undertaken to identify antigenic, non-toxic and non-allergenic epitopes with a high potential to initiate a robust immune response. The filtering process had the following steps:

4.4.4.1 Antigenicity and toxicity prediction of B-cell and T-cell epitopes

VaxiJen (Doytchinova & Flower, 2007) was used to predict the epitope having the potential to initiate an immune response. The epitopes with a VaxiJen score of 0.8 or above were selected. The toxicity and non-toxicity of an epitope can be determined using the SVM-based tool, ToxinPred (Gupta et al., 2013). Non-toxic epitopes were chosen for further analysis.

4.4.4.2 Hydrophilicity prediction

Hydrophilic epitopes are present on the surface of the antigenic proteins. The selection of hydrophilic epitopes would help in constructing a vaccine that would initiate a quick immune response. The **gr**and **av**erage of hydropathicit**y** (GRAVY) index score (Kyte & Doolittle, 1982) was predicted using ProtParam (Gasteiger et al., 2005).

4.4.4.3 Allergenicity prediction

Allergenicity determines whether the epitope is allergic or non-allergic. AllerTOP 2.0 predicts the allergenicity of epitopes based on the physiochemical properties of a protein sequence (Dimitrov et al., 2014). All allergenic epitopes were eliminated.

4.4.4.4 Population coverage analysis

Designing a universal vaccine for TB needs an analysis of the population coverage to minimise the risk of developing an ethnically-biased vaccine (Bui et al., 2006). The population coverage assessment uses the IEDB epitope analysis resource for a distinct population coverage. This analysis tool predicts the predictable population coverage, average number of epitope hits or HLA combinations recognised by diverse ethnic groups or populations, and the least number of epitope hits expected by 90% of the population (Bui et al., 2006). The final epitopes for a vaccine must have extensive population coverage worldwide.

4.4.5   *In silico* vaccine construction using structural vaccinology

After selecting epitopes from immunoinformatics analysis, structural vaccinology was implemented to design the structural model for the TB vaccine and analyse its interaction with toll-like receptors present in the host.

4.4.5.1 Vaccine structure construction and validation

1. **Vaccine design**: For designing a final vaccine sequence from the shortlisted epitopes, we used a new approach that had the following steps:

   (i) Selection of linkers: Flexible linkers help restore protein folding and maintain the interaction between protein domains. Figure 4.6 shows the linker used for designing the vaccine proteins.

      - The GPGPG linker was used to attach helper-T cell epitopes with cytotoxic-T epitopes and cytotoxic-T cell epitopes with B-cell epitopes.
      - For attaching helper-T cell epitopes with themselves, the GPGPG linker was used.
      - For attachment of cytotoxic-T cell epitopes, AAY was used.
      - For B-cell epitopes, KK was used.

      The 3D structure of each epitope with the attached linker was generated using PEPstrMOD (Singh et al., 2015).

   (ii) Pan DR epitope (PADRE) is a universal synthetic 13 amino acid peptide. The PADRE sequence was used to improve the efficacy of cell-mediated immune response. The attachment of PADRE with CTL epitopes would facilitate quick and robust interaction with Toll-like receptors in the host body. It would accomplish the goal of initiating the immune response. Thus, it was attached to the first helper-T cell epitope in the vaccine protein (Ghaffari-Nazari et al., 2015).

   (iii) For finalising the vaccine sequence, the 3D structure of the PADRE sequence with the GPGPG linker attached to its end was generated using PEPstrMOD. Next, molecular docking of the 3D structure of PADRE was completed with each structure of the helper-T cell epitopes (structure constructed in step (i)) using PatchDock (Schneidman-Duhovny et al., 2005) and FireDock (Mashiach et al., 2008). The helper-T cell epitope with strong binding with the structure of PADRE was selected as the first epitope in the final sequence design.

   (iv) Next, the PADRE+1st helper-T cell epitope structure was constructed using I-TASSER (Yang et al., 2015), and its compatibility was checked again with the remaining helper-T cell epitopes. This process was completed first for helper-T cell epitopes, then cytotoxic-T cell epitopes and finally the B-cell epitopes. This process continues until the last B-cell epitope was found for the vaccine sequence.

   (v) Adjuvants were used to enhance the immune response in the host. In this study, we used two adjuvants. At the N-terminal of protein, we used the 50S ribosomal protein L7/L12 and at the C-terminal of the vaccine, protein β-defensin (Tani et al., 2000) was used (Figure 4.6). The 50S ribosomal protein L7/L12 and β-defensin were attached to

vaccine protein obtained from above using an EAAAK linker (purple coloured box in Figure 4.6).
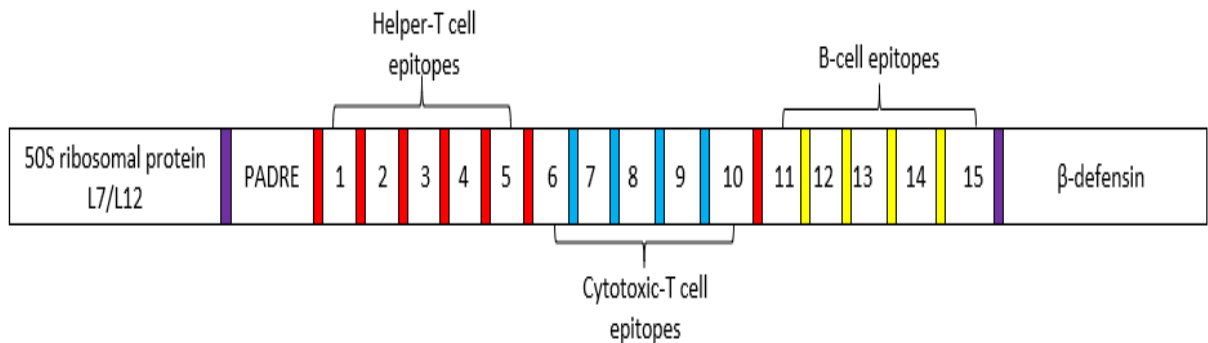
Figure 4. 6: The location of different linkers used in designing the final vaccine protein - purple coloured box: EAAAK linker, red: GPGPG, blue: AAY, yellow: KK. The adjuvants are connected to the vaccine protein using EAAAK. The GPGPG linker combined different types of epitopes

2. **Antigenicity and allergenicity prediction**: For predicting the antigenicity of the final vaccine construct, VaxiJen (Doytchinova & Flower, 2007) was used. Algpred, an SVM module, was used for determining the allergenicity of the vaccine protein (Saha & Raghava, 2006a).

3. **Physiochemical properties analysis**: ProtParam was used for computing the various physiochemical properties of the final TB vaccine. The properties include molecular weight (MW), the isoelectric point of a protein (pI), amino acid composition, extinction coefficient (Gill & von Hippel, 1989), instability index (II), estimated half-life (Bachmair et al., 1986), aliphatic index (Ikai, 1980) and the GRAVY index score (Kyte & Doolittle, 1982)

4. **Secondary structure prediction**: The secondary structure prediction of the vaccine construct used PSIPRED (Buchan & Jones, 2019). Knowledge of alpha-helix, beta-sheet and coils improves understanding of the structure of the vaccine construct.

5. **Tertiary structure construction and validation:** The tertiary structure of the final TB vaccine construct was generated using RaptorX (Wang et al., 2016). First, RaptorX performed a template search based on the similarity of the input sequence and then constructed a good quality structural model. The output structural model of the vaccine, given by RaptorX, was further refined using GalaxyRefine to minimise any distortions present in the structure (Ko et al., 2012). In the last step, RAMPAGE was used for determining the quality of the vaccine structural model (Lovell et al., 2003). If the vaccine structural model had more than 90% of its residues in the most favoured region, it was considered the best quality model and was used for further analysis.

4.4.5.2 Vaccine-TLR docking and dynamics

TLRs present on the surface of the antigen-presenting cells interact with the pathogen to initiate the innate immune response. As stated in chapter-2, *Mycobacterium tuberculosis* usually interacts with TLR-2, 4 and 6. Analysis of the interaction of TLRs with the final vaccine construct was undertaken by performing molecular docking. For the docking process, the 3D structures of TLR-2, TLR-4 and TLR-6 were retrieved from PDB with PDB ID 2Z7X, 4G8A and 4OM7, respectively. The Cluspro 2.0 (Kozakov et al., 2017) webserver was used for performing molecular docking of the vaccine construct-TLR. Cluspro 2.0 performs three steps: (i) performing rigid docking using the fast Fourier transform (FFT) correlation method, (ii) calculation of the root-mean-square deviation (RMSD) for docked clusters, and (iii) refinement of the structure using a Monte Carlo simulation algorithm (Kozakov et al., 2017). After docking, normal mode analysis (NMA) helped determine the mobility and the presence of any deformation in the interacting macromolecule. iMODS (López-Blanco et al., 2014) was used for determining the dynamic motion of the vaccine-TLR docked complex.

4.4.6 Analysis of the evoked immune response by the TB vaccine construct

The final, and most crucial, step in this study was analysing the immune response generated by the final vaccine construct. This helped in understanding the immunogenic nature of the vaccine. For estimating the immunogenic potential of the TB vaccine constructed, the C-ImmSim (Rapin et al., 2010) server was used. C-ImmSim server is an agent-based model that uses the position-specific scoring matrix (PSSM) and machine learning approaches, such as neural networks, for predicting epitopes, the immune interactions of epitopes and immune cells, and immune response activation (Rapin et al., 2010). The C-ImmSim server simulated three sections that corresponded to the three different anatomical regions inside the host body: (i) bone marrow, (ii) thymus, and (iii) lymph nodes (Rapin et al., 2010). For determining vaccine efficacy, the TB vaccine protein sequence in FASTA format was given as input to the server. The vaccine protein sequence was administered three times at an interval of four weeks. In C-ImmSim, one simulation step represented eight hours. So, we used 1050 simulation steps to predict the immune response for the TB vaccine for over a year.

4.5 Results

Here we present results from the vaccine development pipeline. Specifically, this study designed and tested a conceptual framework using computational vaccinology. This section describes the result of the holistic framework tested using several bioinformatics tools.

4.5.1 Comparative proteomic analysis

To develop a potent TB vaccine that would provide broad spectrum of protection against many *Mycobacterium tuberculosis* strains and drug resistance, we downloaded the complete proteome

sequence of 159 different *Mycobacterium tuberculosis* strains from the NCBI Genome FTP site (Appendix Table G. 1 provides information on the 159 *Mycobacterium tuberculosis* strains). *Mycobacterium tuberculosis* H37Rv (GenBank accession number NC_000962.3), having 3906 proteins, was taken as the reference proteome for identifying the conserved proteins within 159 strains. A protein was selected as a conserved target only if: (i) it was present in all 159 strains, (ii) it had a 100% query coverage compared to *Mycobacterium tuberculosis* H37Rv, and (iii) it had a more than 99% sequence identity among the 159 strains. Proteome comparison was made using a standalone BLAST and the results from sequence similarity were stored in Excel files. Out of 3906 proteins of H37Rv, 1982 proteins were conserved among 159 strains of *Mycobacterium tuberculosis*.

After performing sequence similarity among the 159 strains, we categorised each protein into a functional class, as given by Smith (2003). This step was carried out to determine the functional significance of conserved proteins and predict their role in bacterial survival strategy in the evolution of bacteria into many strains. Table 4.1 shows the distribution of proteins of conserved proteins in each functional category. When we compared the conserved proteins (1982) with the whole proteome (3906), we found a high percentage (73%) of conserved proteins involved in critical functional classes. For example, intermediary metabolism and respiration (28%), cell wall processes (19%), information pathways (8%), lipid metabolism and regulatory proteins (7%), and virulence, detoxification and adaptation (4%). This suggests that *Mycobacterium tuberculosis* does not support an evolutionary process in proteins playing a vital role in biological processes and functions, such as physiology, metabolism and translation of proteins. Instead, it experiments with or employs other less functionally essential proteins, thus not jeopardising the normal functioning and structural stability needed to survive the TB bacteria.

Twenty-four per cent of conserved hypothetical proteins (proteins with no assigned function) were found in the conserved protein list, which was almost half the whole proteome of H37Rv. We discovered an under-representation of proteins in functional categories, such as the PE and PPE proteins, IS elements and bacteriophages and proteins with unknown functions in the list of conserved proteins. Out of 167 PE and PPE proteins, only 21 proteins were found to be conserved. No protein was found in stable RNA.

Table 4. 1: Distribution of whole proteome (3906 proteins) of *Mycobacterium tuberculosis* H37Rv and the conserved proteins (1982 proteins) identified after performing sequence similarity within 159 strains into functional categories suggested by Smith (2003)

S. No	Functional classes	Whole proteome	Conserved proteins
I	Lipid metabolism	215	147
II	Information pathways	207	154
III	Cell wall and cell processes	506	373
IV	Stable RNAs	44	0
V	IS elements and bacteriophages	125	15
VI	PE and PPE proteins	167	21
VII	Intermediary metabolism and respiration	845	561
VIII	Regulatory proteins	188	135
IX	Virulence, detoxification and adaptation	91	85
X	Conserved hypothetical function	911	467
XI	Proteins of unknown function	607	24
		3906	1982

Next, we performed the functional annotation of the 467 conserved hypothetical proteins and 24 proteins with unknown functions to predict their functional roles in *Mycobacterium tuberculosis.* For function annotation, we used Pfam, SMART, Interpro Scan and BLASTp. Based on the sequence similarity of the query proteins with already annotated proteins, we can determine the function of hypothetical proteins. Out of the 467 conserved hypothetical proteins, we were able to annotate 259 proteins. No reference was found for the 24 proteins with unknown functions. Figure 4.7 shows the distribution of 259 conserved hypothetical proteins into different functional classes. A large number of proteins (n) were found to be involved in intermediary metabolism and respiration (n=125), followed by cell wall processes (n=43), regulatory proteins (n=38) and information pathways (n=33). The percentages of proteins belonging to lipid metabolism (n=12), IS elements and bacteriophages (n=4), and virulence, detoxification and adaptation (n=3), were very low. Only one protein was found to be from the PE and PPE protein family. No protein was found in stable RNAs.

Comparative proteomic analysis suggests that *Mycobacterium tuberculosis* prefers mutations in proteins that have less biological significance in the growth and development of TB bacteria. *Mycobacterium tuberculosis* safeguards the proteins involved in the normal functioning of bacteria, progression of TB disease and survival within the host.
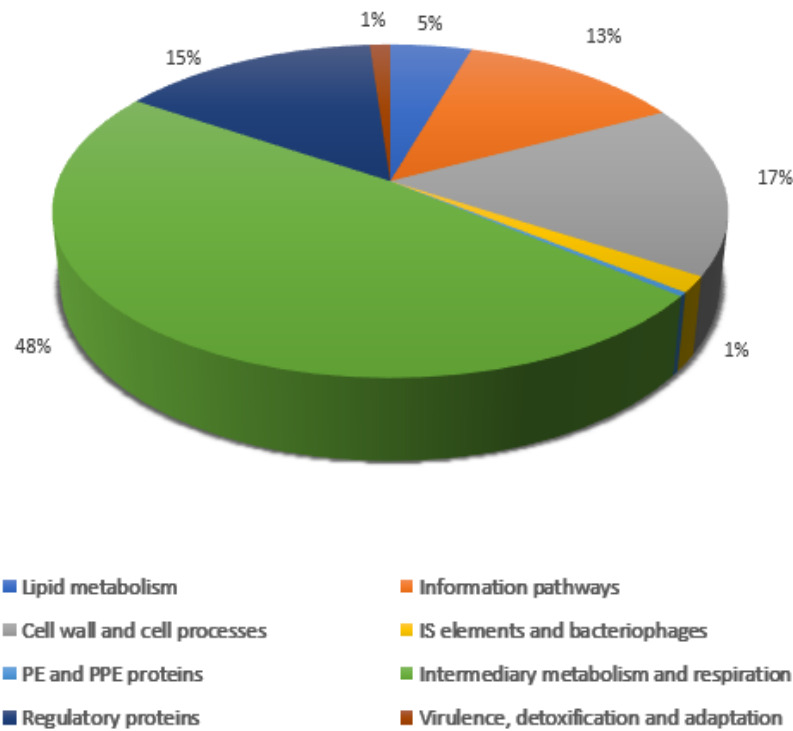
Figure 4. 7: Distribution of 259 conserved hypothetical proteins into functional categories as suggested by Smith (2003)

4.5.2 Identification of 24 antigenic proteins from 1982 conserved proteins using a reverse vaccinology approach

A reductionist reverse vaccinology process was performed on the 1982 conserved proteins to reduce the time and resources needed to identify potential antigenic TB proteins as vaccine candidates. Figure 4.8 shows the reverse vaccinology approach used for the identification of 24 TB antigens from 1982 proteins.

4.5.2.1 Prediction of the subcellular location of the 1982 conserved proteins

SCL prediction was undertaken to identify the location of each protein in *Mycobacterium tuberculosis.* The proteins present in the outer membrane and extracellular space are involved in membrane integrity and permeability, efflux mechanism and active transport of molecules. They are considered as suitable TB vaccine candidates since they are easily accessible to the host immune system. The proteins present in the cell membrane, cell wall and extracellular space of *Mycobacterium tuberculosis*, as predicted by four or more tools, were selected in this step. With the help of PSORTb v.3.0, CELLO, LocTree3, SOSUI, pLoc_bal-mGpos and GramLocEN, 525 proteins were filtered from 1982 conserved proteins. These 525 proteins were localized in the cell membrane, cell wall and extracellular space of *Mycobacterium tuberculosis*.
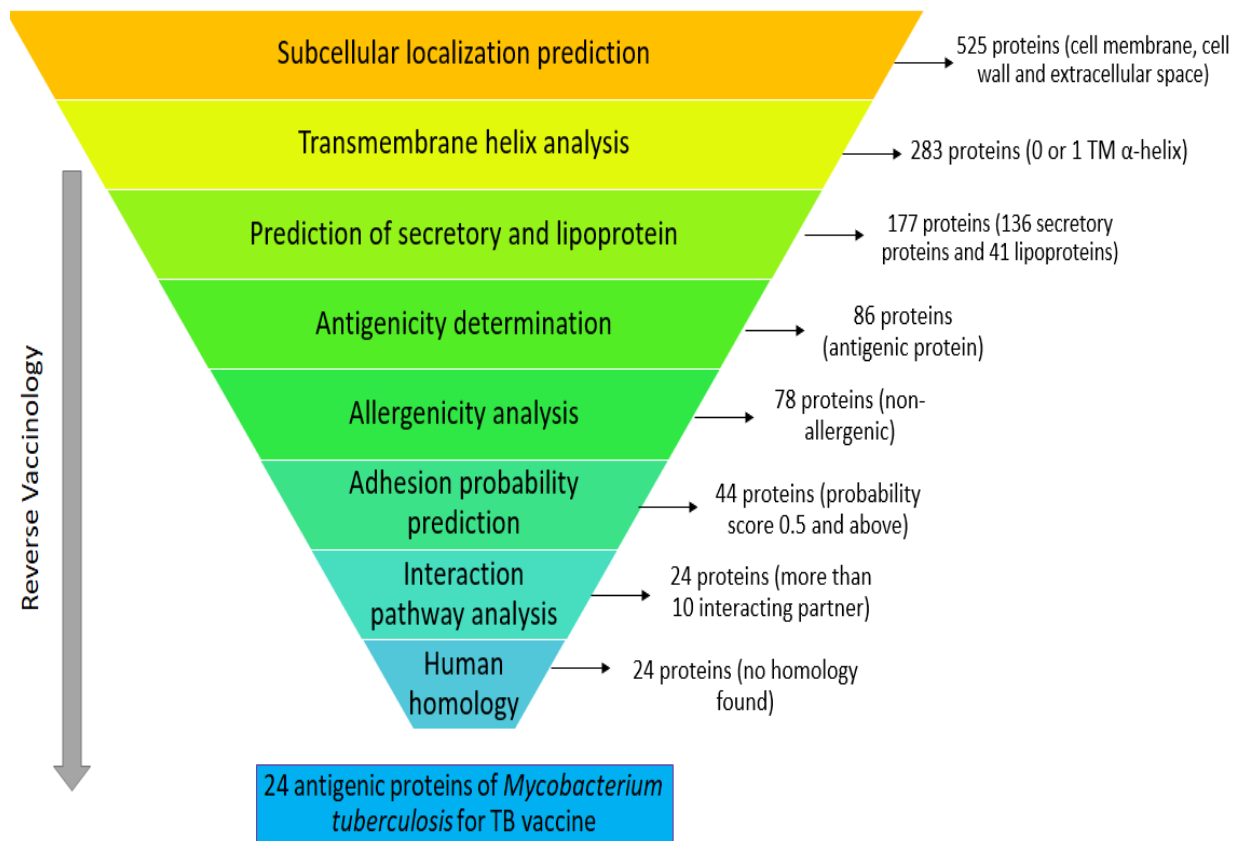
Figure 4. 8: Reverse vaccinology approach used for identification of 24 *Mycobacterium tuberculosis* antigens from 1982 conserved proteins

4.5.2.2 Exclusion of proteins having more than one transmembrane α-helix

Transmembrane integral proteins are membrane-spanning proteins and contain a transmembrane α-helix. Proteins with more than one transmembrane α-helix are difficult to clone and purify in the laboratory. Thus, proteins having a single transmembrane helix were intended to help in designing a potential vaccine. With the help of the TMHMM server, we found that out of 525 proteins, 242 proteins had more than one transmembrane α-helix. These 242 proteins were excluded from the study and the remaining 283 proteins were selected for further analysis.

4.5.2.3 Identification of secretory and lipoproteins

In *Mycobacterium tuberculosis,* secretory proteins and lipoproteins are considered important antigenic and immunogenic TB vaccine targets. Secretory signal peptides are ubiquitous protein-sorting signals that help translocate a protein across the cytoplasmic membrane in mycobacteria. There are two important secretion systems in *Mycobacterium tuberculosis:* the secretory pathway (Sec pathway) and the twin-arginine translocase pathway (TAT pathway) and they already have a highly conserved

process for protein secretion. Studies have shown that secretory proteins of *Mycobacterium tuberculosis* are involved in the virulence of mycobacteria (Abdallah et al., 2007). The Sec pathway helps in the translocation of proteins in their unfolded state. The TAT pathway translocates proteins in their folded state. We identified 119 proteins involved in the Sec pathway using SecretomeP and the SignalP 4.1 server. PRED-TAT identified 17 proteins involved in the TAT pathway. Out of 283 proteins, 136 proteins were found to be engaged in the secretory pathway of *Mycobacterium tuberculosis*.

Lipoproteins are an essential set of membrane proteins that perform different functions in *Mycobacterium tuberculosis*, such as modulating the host's immune response, translocation of virulent factors in the macrophages, signal transduction, and the uptake of nutrients. Therefore, with the help of PRED-LIPO, 41 lipoproteins were identified from the 283 proteins. Thus, a total of 177 (136 secretory proteins and 41 lipoproteins) proteins were selected for further analysis (Figure 4.8).

4.5.2.4 Selection of antigenic and non-allergic proteins as vaccine targets

A protein is considered a potential vaccine target only if it can initiate an immune response without causing considerable side effects inside the host body. For the prediction of antigenicity among the 177 conserved proteins, three bioinformatics tools were used. Concordance analysis by VaxiJen, VirulentPred and MP3 showed that 86 proteins were potentially highly antigenic compared to the remaining proteins (91 proteins).

Algpred was used to predict the allergenic or non-allergenic proteins. From the 86 antigenic proteins, 78 proteins were found to be non-allergenic. These 78 conserved proteins were further analysed for adhesion probability. SPAAN was used to identify the adhesin proteins among these proteins. Adhesin proteins help attach *Mycobacterium tuberculosis* to the host's cell-surface receptors. 44 proteins had a strong adhesion probability with a score of 0.5 or above. Hence, these 44 proteins were chosen to study *Mycobacterium tuberculosis* pathway interactions using the STRING database. Twenty-four conserved and antigenic proteins were found to have more than ten interacting partners with a confidence score of 0.4. Targeting a protein having a large number of interacting partners in the membrane region of *Mycobacterium tuberculosis* plays a crucial role in weakening cell membrane and cell wall, thus, leading to the death of the bacterial cell. The last step of reverse vaccinology involved the prediction of the homology between humans and *Mycobacterium tuberculosis.* The homologous proteins may initiate autoimmune reactions causing severe health problems in the host. We found no similarity among the 24 *Mycobacterium tuberculosis* proteins and humans (Figure 4.8).

A total of 24 conserved, membrane-spanning, antigenic, non-allergic proteins with high adhesion probability and non-homologous to humans were selected using the reverse vaccinology approach, as

shown in Table 4.2. Out of the 24 proteins, nine proteins were observed in the cell walls and the cell processes functional class. Six proteins belonged to the PE and PPE protein families. Three proteins were involved in lipid metabolism, whereas only one protein belonged to the virulence, detoxification, and adaptation class. The literature showed that these proteins play an essential role in escaping from the human immune response. PPE11 is involved in the pathogenesis of TB and macrophage persistence (Deng et al., 2015) and PPE15 inhibits the production of reactive nitrogen inside alveolar macrophages (Fishbein et al., 2015). Mazandu and Mulder (2012) predicted the role of PPE21 in the pathogenicity of *Mycobacterium tuberculosis*. PPE65 helps evade the adaptive immune response by obstructing the helper-T cell response (Khubaib et al., 2016). ESAT-6 regulates macrophage apoptosis (Roy et al., 2020). The membrane protein, MmpS4, plays an essential role in the growth of *Mycobacterium tuberculosis* under iron-deficiency conditions (Wells et al., 2013). Goulding et al. (2009) study suggests that immunogenic protein Mpt63 is involved in the virulence mechanism inside the host body. A recombinant vaccine constructed using Ag85A and Ag85B is in phase-I of a TB vaccine clinical trial (WHO Global Tuberculosis Report, 2020).

Thus, literature validates that we had selected some of the most crucial proteins involved in the survival mechanisms of *Mycobacterium tuberculosis* and the construction of an epitope-based TB vaccine from these 24 proteins would help generate a robust immune response in humans.

Table 4. 2: List of 24 surface-exposed, antigenic and non-allergic proteins identified by reverse vaccinology approach. Column 1- gene id, column 2- accession number of the protein, column 3-function of protein, column 4- functional class given by Smith (2003)

Gene	Accession No	Function	Functional class
Rv0129c	YP_177694.1	diacylglycerol acyltransferase/mycolyltransferase Ag85C	I
Rv0287	NP_214801.1	ESAT-6 like protein EsxG	III
Rv0451c	NP_214965.1	membrane protein MmpS4	III
Rv0453	YP_177727.1	PPE family protein PPE11	VI
Rv1039c	YP_177778.1	PPE family protein PPE15	VI
Rv1184c	NP_215700.1	hypothetical protein Rv1184c	VI
Rv1252c	NP_215768.1	lipoprotein LprE	III
Rv1548c	YP_177817.1	PPE family protein PPE21	VI
Rv1706c	YP_177828.1	PPE family protein PPE23	VI
Rv1886c	NP_216402.1	diacylglycerol acyltransferase/mycolyltransferase Ag85B	I
Rv1891	NP_216407.1	hypothetical protein Rv1891	-
Rv1906c	NP_216422.1	hypothetical protein Rv1906c	-
Rv1926c	NP_216442.1	immunogenic protein Mpt63	III
Rv1973	NP_216489.1	Mce associated membrane protein	III
Rv2376c	NP_216892.1	low molecular weight antigen MTB12	III
Rv2507	NP_217023.1	hypothetical protein Rv2507	III

Rv2518c	NP_217034.1	L,D-transpeptidase LdtB	III
Rv2873	NP_217389.1	cell surface lipoprotein	III
Rv3572	NP_218089.1	hypothetical protein Rv3572	-
Rv3621c	YP_177998.1	PPE family protein PPE65	VI
Rv3707c	NP_218224.1	hypothetical protein 3707c	VIII
Rv3803c	YP_178017.1	MPT51/MPB51 antigen	IX
Rv3804c	NP_218321.1	diacylglycerol acyltransferase/mycolyltransferase Ag85A	I
Rv3873	YP_178022.1	PPE family protein PPE68	VI

4.5.3 Immunoinformatics analysis for the prediction of B-cell and T-cell epitopes for TB vaccine construction

The most crucial step in designing an epitope-based TB vaccine is identifying the immunodominant B-cell and T-cell epitopes. Epitopes are the part of an antigen that binds to T-cells, B-cells or antibodies in the host body. The T-cells and B-cells are the principal constituents of the immune system. For predicting epitopes, immunoinformatics analysis is performed to predict the B-cell and T-cell (MHC-I and II-restricted epitopes) and then the most suitable vaccine candidates are filtered based on immunogenicity, toxicity allergenicity and hydrophilicity. The hydrophilic peptides are usually present on the surface of the protein.

4.5.3.1 B-cell epitopes for TB vaccine

B-cells are essential in providing humoral or antibody-mediated responses. The epitopes recognised by the antibody's paratopes (antigen/epitope-recognition site in antibodies) or B-cell receptors (BCR) are termed B-cell epitopes. In our study, we used Bepipred, ABCpred and BCPREDS for identifying the B-cell epitopes. From all three bioinformatics tools, a total of 280 B-cell epitopes were identified from the 24 antigenic proteins. These 280 epitopes had varying lengths of 2-40 amino acid residues. A B-cell epitope with a length of 8-20 residue was considered a suitable vaccine candidate. Concordance analysis was undertaken to identify B-cell epitopes commonly predicted by all three tools to improve prediction accuracy. In total, 74 B-cell epitopes of 8-20 residues length were found to be common among these 24 antigenic TB proteins. Table 4.3 provides the results of immunoinformatics analysis performed on the 24 TB antigens. The third and fourth columns of Table 4.3 show the B-cell epitopes predicted and selected for each TB protein.

Table 4. 3: Immunoinformatics analysis for prediction of B-cell epitopes. Column 3 shows the number of epitopes predicted by the Bepipred, ABCpred and BCPREDs server. Column 4 shows the number of epitopes commonly predicted by three servers. Columns 5, 6 and 7 show the total number of epitopes filtered by performing antigenicity and toxicity prediction, hydrophilicity prediction and non-allergenicity prediction, respectively from the 24 *Mycobacterium tuberculosis* antigenic proteins

Gene	Protein accession number	B-cell epitopes predicted	B-cell epitopes selected	Antigenicity & toxicity prediction	Hydrophilicity prediction	Non-Allergenicity analysis
Rv0129c	YP_177694.1	12	4	1	-	-
Rv0287	NP_214801.1	10	1	1	-	-
Rv0451c	NP_214965.1	11	1	1	1	1
Rv0453	YP_177727.1	15	8	2	2	2
Rv1039c	YP_177778.1	11	4	1	-	-
Rv1252c	NP_215768.1	9	4	1	1	1
Rv1548c	YP_177817.1	6	1	-	-	-
Rv1706c	YP_177828.1	10	5	1	-	-
Rv1886c	NP_216402.1	12	3	2	-	-
Rv1926	NP_216442.1	20	2	1	-	-
Rv1973	NP_216489.1	10	1	1	1	1
Rv2376c	NP_216892.1	13	2	1	-	-
Rv2518c	NP_217034.1	16	5	3	-	-
Rv2873	NP_217389.1	11	2	3	1	1
Rv3612	YP_177998.1	9	3	1	-	-
Rv3803c	YP_178017.1	5	4	4	-	-
Rv3804c	NP_218321.1	14	3	2	-	-
Rv3873	YP_178022.1	11	2	2	-	-
Rv1184c	NP_215700.1	15	4	2	-	-
Rv1891	NP_216407.1	17	3	1	1	1
Rv1906c	NP_216422.1	10	3	2	2	1
Rv2507	NP_217023.1	6	2	3	-	-
Rv3572	NP_218089.1	15	3	1	-	-
Rv3707c	NP_218224.1	12	4	1	-	-

After selecting 74 epitopes, a filtering process was carried out to investigate safe and immunodominant B-cell epitopes for vaccine construction. First, a prediction of the immunogenic nature of B-cell epitopes was made using the VaxiJen server and ToxinPred. A total of 38 non-toxic B-cell epitopes having a Vaxijen score higher than 0.8 were selected for further analysis. The selection of hydrophilic epitopes for constructing TB vaccine helps in swift interactions between a vaccine and the host cells. This enables a high chance of initiating a robust immune response in the host. The hydrophilicity prediction was made by calculating the GRAVY score using the ProtParam server. In

ProtParam analysis, 9 B-cell epitopes with negative GRAVY scores were selected. For further identification of non-allergenic epitopes, the ALLERTOP 2.0 bioinformatics tool was used. Out of 9 B-cell epitopes, one epitope was found to be allergenic and excluded from the study. After finishing the immunoinformatics analysis, a total of 8 B-cell epitopes were shortlisted for TB vaccine construction.

4.5.3.2 MHC-I restricted T-cell epitopes for TB vaccine

T-cells play a significant role in cell-mediated or cellular immunity. In cellular immunity, *Mycobacterium tuberculosis* is ingested by antigen-presenting cells, such as macrophages, dendritic cells, etc. The pathogen is then fragmented into smaller antigenic peptides. These are later presented to T-cell receptors (TCR) present on the surface of T-cells through the cell-surface attached major histocompatibility complex (MHC) molecule. Most T-cell epitope prediction tools are trained with experimentally validated 9-mer residues and 15-mer epitope residues for MHC-I and MHC-II restricted epitopes prediction.

The epitopes binding to MHC-I molecules are termed cytotoxic T-cell (CTL) epitopes or MHC-I restricted T-cell epitopes. In our study, we set the length of the CTL epitope at nine amino acid residues long. The prediction of CTL epitopes used IEDB MHC-I, NetCTLpan 1.1 and IEDB MHC-NP. In the IEDB server analysis, CTL epitopes with higher binding affinity and a percentile rank lower than 0.4 were selected. From the 24 TB antigens, there was a total of 469 CTL epitopes from all three servers. Two hundred and seven CTL epitopes were found to be commonly predicted by all three servers. The third and fourth columns of Table 4.4 show the CTL epitopes predicted and selected for each TB membrane-localized antigen.

Table 4. 4: Immunoinformatics analysis for the prediction of CTL epitopes or MHC-I restricted T-cell epitopes. Column 3 shows the number of epitopes predicted by the IEDB MHC-I, NetCTLpan1.1 and IEDB MHC-NP server. Column 4 shows the number of epitopes commonly predicted by three servers. Columns 5, 6 and 7 show the total number of epitopes filtered by performing antigenicity and toxicity prediction, hydrophilicity prediction and non-allergenicity prediction, respectively from the 24 *Mycobacterium tuberculosis* antigenic proteins

Gene	Protein accession number	CTL epitopes predicted	CTL epitopes selected	Antigenicity & toxicity prediction	Hydrophilicity prediction	Non-allergenicity analysis
Rv0129c	YP_177694.1	34	16	3	2	2
Rv0287	NP_214801.1	8	2	2	1	-
Rv0451c	NP_214965.1	12	5	4	2	1
Rv0453	YP_177727.1	35	15	4	1	-
Rv1039c	YP_177778.1	39	14	3	1	1
Rv1252c	NP_215768.1	7	3	1	1	-

Rv1548c	YP_177817.1	26	13	5	2	2
Rv1706c	YP_177828.1	44	17	4	-	-
Rv1886c	NP_216402.1	22	12	1	1	-
Rv1926	NP_216442.1	13	4	1	1	1
Rv1973	NP_216489.1	10	4	-	-	-
Rv2376c	NP_216892.1	11	6	1	1	1
Rv2518c	NP_217034.1	29	14	2	2	-
Rv2873	NP_217389.1	16	7	4	1	1
Rv3612	YP_177998.1	33	16	6	3	1
Rv3803c	YP_178017.1	24	11	3	-	-
Rv3804c	NP_218321.1	16	7	2	-	-
Rv3873	YP_178022.1	22	5	2	-	-
Rv1184c	NP_215700.1	15	7	2	2	1
Rv1891	NP_216407.1	8	4	1	1	-
Rv1906c	NP_216422.1	7	6	1	1	1
Rv2507	NP_217023.1	12	5	3	1	1
Rv3572	NP_218089.1	11	5	-	-	-
Rv3707c	NP_218224.1	15	9	3	1	1

For confirming the ability of CTL epitopes to initiate an immune response, antigenicity and toxicity analysis was performed using VaxiJen and Toxin Pred. The antigenicity and toxicity scores revealed that out of 207 CTL epitopes, 58 epitopes were highly immunogenic and non-toxic. These 58 epitopes were considered for further hydrophilic score calculations. In our study, we excluded the epitopes having positive GRAVY scores. Thus, we had 25 CTL epitopes that were hydrophilic in nature and with a negative GRAVY score. The allergenic nature of the epitopes was then determined using ALLERTOP 2.0. Out of 25 CTL epitopes, 11 were found to cause allergenic reactions inside the host body. These epitopes were excluded from the study. Finally, 14 CTL epitopes were chosen for constructing an epitope-based TB vaccine.

4.5.3.3 MHC-II restricted T-cell epitopes for a TB vaccine

MHC class-II molecules are present on macrophage and dendritic cells. The primary function of MHC-II is to present antigens/epitopes to the naïve T-helper cells. This interaction leads to the release of cytokines that help the development of naïve T-helper cells into effector or memory T-cells. The epitopes presented by MHC-II molecules are termed helper T-cells (HTL epitopes) or MHC-II restricted T-cell epitopes. We set the length of HTL epitopes at 15 amino acid residues. IEDB MHC-II and the NetMHCIIpan 3.2 server identified a total of 428 epitopes. Concordance analysis estimated 346 common HTL epitopes among the 24 *Mycobacterium tuberculosis* proteins. There was no common epitope found in one protein with accession number NP_214965.1.

Table 4. 5: Immunoinformatics analysis for the prediction of HTL epitopes or MHC-II restricted T-cell epitopes. Column 3 shows the number of epitopes predicted by the IEDB MHC-II and NetMHHCIIpan 3.2 server. Column 4 shows the number of epitopes commonly predicted by both servers. Columns 5, 6 and 7 show the total number of epitopes filtered by performing antigenicity and toxicity prediction, hydrophilicity prediction and non-allergenicity prediction, respectively from the 24 *Mycobacterium tuberculosis* antigenic proteins

Gene	Protein accession number	HTL epitopes predicted	HTL epitopes selected	Antigenicity & toxicity prediction	Hydrophilicity prediction	Non-allergenicity analysis
Rv0129c	YP_177694.1	21	13	3	-	-
Rv0287	NP_214801.1	13	9	8	-	-
Rv0451c	NP_214965.1	3	-	-	-	-
Rv0453	YP_177727.1	34	27	9	-	-
Rv1039c	YP_177778.1	46	39	36	-	-
Rv1252c	NP_215768.1	7	5	2	-	-
Rv1548c	YP_177817.1	40	32	21	-	-
Rv1706c	YP_177828.1	30	25	16	-	-
Rv1886c	NP_216402.1	21	18	14	2	2
Rv1926	NP_216442.1	1	1	-	-	-
Rv1973	NP_216489.1	13	10	9	-	-
Rv2376c	NP_216892.1	31	25	17	-	-
Rv2518c	NP_217034.1	3	2	2	-	-
Rv2873	NP_217389.1	27	23	21	-	-
Rv3612	YP_177998.1	32	27	25	1	1
Rv3803c	YP_178017.1	4	4	4	-	-
Rv3804c	NP_218321.1	24	18	14	2	2
Rv3873	YP_178022.1	18	15	14	-	-
Rv1184c	NP_215700.1	7	7	4	-	-
Rv1891	NP_216407.1	15	12	10	-	-
Rv1906c	NP_216422.1	1	1	1	-	-
Rv2507	NP_217023.1	10	10	10	1	-
Rv3572	NP_218089.1	11	11	6	-	-
Rv3707c	NP_218224.1	16	12	4	-	-

Table 4.5 shows the results of the immunoinformatics analysis on the 24 *Mycobacterium tuberculosis* proteins. Columns 5, 6 and 7 in Table 4.5 show the outcomes of the antigenicity, toxicity, hydrophilicity and allergenicity analysis. Out of the 250 immunogenic and non-toxic HTL epitopes, only six were found to be hydrophilic. All remaining HTL epitopes were excluded from the study and finally, five HTL epitopes remained after evaluating the allergenic nature of epitopes.

After selecting highly immunogenic and exclusion of cross-reactive epitopes, we had 27 epitopes (8 B-cell epitopes, 14 CTL epitopes and 5 HTL epitopes) from 18 (out of 24) antigenic *Mycobacterium*

tuberculosis proteins. The selected epitopes from the 18 proteins that can evoke a strong cellular and humoral immune response in the host are listed in Table 4.6. Three proteins were found to have both B-cell and T-cell epitopes. No suitable vaccine candidates were found in six proteins: NP_214801.1, YP_177828.1, NP_217034.1, YP_178017.1, YP_178022.1 and NP_218089.1. Out of these six proteins, the functional significance of the three proteins (NP_217034.1, NP_218089.1 and YP_178017.1) was unknown in *Mycobacterium tuberculosis*.

Table 4. 6: Final 27 epitopes (8 B-cell epitopes, 14 CTL epitopes and 5 HTL epitopes) selected from 18 *Mycobacterium tuberculosis* antigens for constructing an epitope-based vaccine for tuberculosis. Length represents the length of an epitope or number of amino acid residues in an epitope, start and end represents starting and ending position of an epitope in a protein, and epitope represents one letter code of amino acid residues

S. No	Accession	B-cell epitope				CTL epitope				HTL epitope				Immune response
		length	start	end	epitope	length	start	end	epitope	length	start	end	epitope	
1	YP_177694.1					9	72	80	FQGGGPHAV					Cellular immunity
						9	210	218	LAMNDSGGY					
2	NP_214965.1	12	119	130	KVRAERVSNEVN	9	124	132	RVSNEVNAY					**Cellular and humoral immunity**
3	YP_177727.1	20	181	201	SNAQSQHSSSNNSGGADPVD									Humoral immunity
		20	424	444	RHQARRRRRAAAKERGNADE									
4	YP_177778.1					9	365	373	SAAKGTGAY					Cellular immunity
5	NP_215768.1	20	25	45	TLTGCGSGDSTVAKTPEATP									Humoral immunity
6	YP_177817.1					9	457	465	NSATTSTGW					Cellular immunity
						9	564	572	HTGTNNSGY					
7	NP_216402.1									15	46	60	LPVEYLQVPSPSMGR	Cellular immunity
										15	47	61	PVEYLQVPSPSMGRD	
8	NP_216442.1					9	99	107	RTADGINYR					Cellular immunity
9	NP_216489.1	10	127	136	ITVGKDAPTT									Humoral immunity
10	NP_216892.1					9	93	101	RIADHKLKK					Cellular immunity
11	NP_217389.1	20	162	182	QASPSRIDGTHQTLQGADLT	9	36	44	SPKPATSPA					**Cellular and humoral immunity**
12	YP_177998.1					9	398	406	RYGFKPTVI	15	361	375	SNGLGAAAAAEGSTH	Cellular immunity
13	NP_218321.1									15	49	63	LPVEYLQVPSPSMGR	Cellular immunity
										15	50	64	PVEYLQVPSPSMGRD	
14	NP_215700.1					9	308	316	LQPQIDAAY					Cellular immunity
15	NP_216407.1	20	20	40	PVAGADPQRYDGDVPGMNYD									Humoral immunity

16	NP_216422.1	20	30	50	AGADPEPAPTPKTAIDSDGT	9	4	12	KPAPSPAAA					**Cellular and humoral immunity**
17	NP_217023.1					9	70	78	HEASPTQQL					Cellular immunity
18	NP_218224.1					9	9	17	GTGTPTGDY					Cellular immunity

4.5.3.4 Population coverage analysis

MHC molecules are highly polymorphic and different HLA types are expressed at different rates in different ethnicities around the world. The prevalence of MHC-I and II HLA alleles in the diverse ethnic groups of the world helped in determining the most likely set of T-cell epitopes. Tables 4.7 and 4.8 show the CTL epitopes and HTL epitopes and their corresponding MHC-I and MHC-II HLA alleles.

Table 4. 7: Potential CTL epitopes with their respective MHC-I HLA alleles

S. No	Accession No	CTL epitopes			
		start	end	epitope	MHC-I HLA allele
1	YP_177694.1	72	80	FQGGGPHAV	HLA-A*02:01, HLA-B*39:01, HLA-B*53:01, HLA-A*02:06, HLA-A*01:01
		210	218	LAMNDSGGY	HLA-A*01:01, HLA-A*26:01, HLA-B*15:01, HLA-B*53:01, HLA-B*44:03, HLA-B*35:01, HLA-A*30:02
2	NP_214965.1	124	132	RVSNEVNAY	HLA-A*01:01, HLA-A*26:01, HLA-B*15:01, HLA-B*53:01, HLA-B*44:03, HLA-B*35:01, HLA-B*57:01, HLA-A*02:01, HLA-A*30:02
3	YP_177778.1	365	373	SAAKGTGAY	HLA-A*01:01, HLA-A*26:01, HLA-B*15:01, HLA-B*44:03, HLA-B*53:01, HLA-B*35:01, HLA-B*57:01, HLA-A*30:02
4	YP_177817.1	457	465	NSATTSTGW	HLA-B*58:01, HLA-B*53:01, HLA-B*44:03, HLA-B*57:01, HLA-A*01:01
		564	572	HTGTNNSGY	HLA-A*01:01, HLA-A*26:01, HLA-B*44:03, HLA-B*35:01, HLA-A*02:01
5	NP_216442.1	99	107	RTADGINYR	HLA-B*53:01, HLA-B*44:03, HLA-B*57:01, HLA-A*31:01, HLA-A*68:01, HLA-A*02:01, HLA-A*01:01
6	NP_216892.1	93	101	RIADHKLKK	HLA-A*03:01, HLA-B*53:01, HLA-A*02:01, HLA-B*44:03, HLA-A*11:01, HLA-A*02:01
7	NP_217389.1	36	44	SPKPATSPA	HLA-B*07:02, HLA-B*53:01, HLA-B*35:01, HLA-A*01:01, HLA-A*26:01, HLA-B*44:03, HLA-B*35:01
8	YP_177998.1	398	406	RYGFKPTVI	HLA-A*24:02, HLA-B*53:01, HLA-B*07:02, HLA-B*44:03, HLA-A*01:01, HLA-A*26:01
9	NP_215700.1	308	316	LQPQIDAAY	HLA-B*15:01, HLA-B*44:03, HLA-B*53:01, HLA-B*35:01, HLA-A*01:01, HLA-A*26:01
10	NP_216422.1	4	12	KPAPSPAAA	HLA-B*53:01, HLA-A*02:01, HLA-B*07:02, HLA-A*01:01, HLA-A*02:01
11	NP_217023.1	70	78	HEASPTQQL	HLA-B*40:01, HLA-B*35:01, HLA-B*44:03, HLA-B*07:02, HLA-A*01:01, HLA-A*02:01
12	NP_218224.1	9	17	GTGTPTGDY	HLA-A*01:01, HLA-B*53:01, HLA-B*44:03, HLA-B*35:01, HLA-A*30:02, HLA-A*02:01

Table 4. 8: Predicted HTL epitopes interacting with their MHC-II HLA alleles

S. N o	Accession No	HTL epitopes			
		start	end	epitope	MHC-II HLA allele
1	NP_216402.1	46	60	LPVEYLQVPSPSMGR	HLA-DRB1*01:01,HLA-DRB1*15:01,HLA-DRB1*03:01,HLA-DRB1*04:05,HLA-DRB1*07:01,HLA-DRB1*09:01
		47	61	PVEYLQVPSPSMGRD	HLA-DRB1*04:01, HLA-DRB1*01:01, HLA-DRB5*01:01, HLA-DRB1*11:01, HLA-DRB1*09:01
2	YP_177998.1	361	375	SNGLGAAAAAEGSTH	HLA-DQA1*04:01, HLA-DRB1*01:01, HLA-DRB3*02:02, HLA-DRB1*11:01, HLA-DRB5*01:01
3	NP_218321.1	49	63	LPVEYLQVPSPSMGR	HLA-DRB1*01:01,HLA-DRB1*15:01,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DRB1*11:01,HLA-DRB1*13:02,HLA-DRB1*07:01,HLA-DRB1*08:02,HLA-DRB1*09:01
		50	64	PVEYLQVPSPSMGRD	HLA-DRB1*01:01,HLA-DRB1*15:01,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DRB1*11:01,HLA-DRB1*13:02,HLA-DRB1*07:01,HLA-DRB1*08:02,HLA-DRB1*09:01

To design a universal vaccine against TB, there is a need to analyse population coverage of the selected T-cell epitopes to minimize the risk of developing an ethnically-biased vaccine (Bui et al., 2006). Thus, the population coverage assessment of the predicted CTL and HTL epitopes and their corresponding MHC-I and MHC-II HLA alleles used the IEDB epitope analysis resource. Among 15 regions, the results exhibited maximum population coverage in Europe (99.74%), closely followed by North America (99.54%), East Asia (99.44%), West Indies (98.14%), South Asia (97.09%), Oceania (96.94%), Southeast Asia (96.52%), Northeast Asia (95.53%), North Africa (95.06%) and others (Figure 4.9). The lowest coverage, compared to the others, was found in Central Africa (89.09%), followed by South Africa (83.76%). As a result, 99.16% of the population of the world was covered by the predicted CTL and HTL epitopes from the 18 antigens of *Mycobacterium tuberculosis*. This analysis supports our strategy of developing a universal TB vaccine.
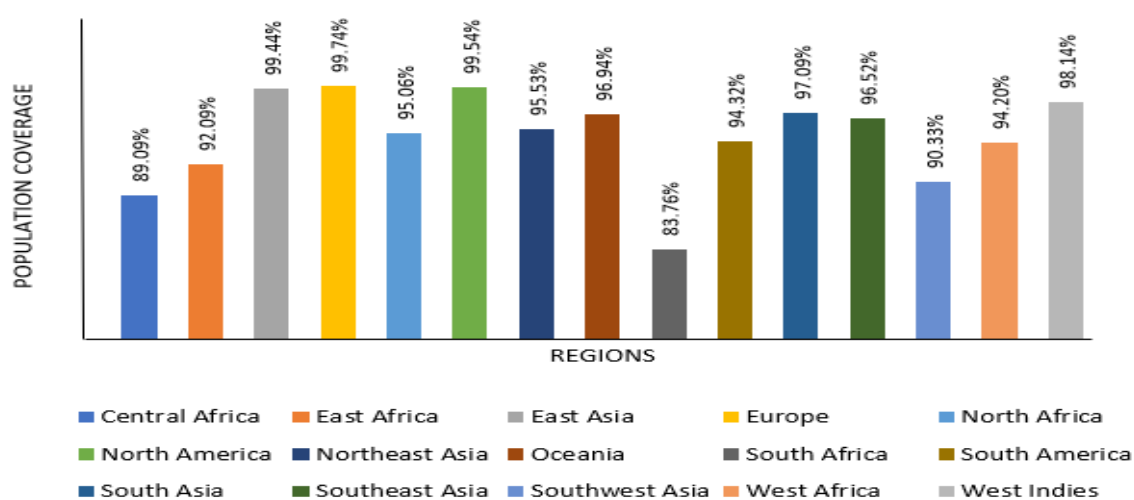
Figure 4. 9: Population coverage analysis predicted using IEDB-AR for 14 CTL epitopes and 5 HTL epitopes with their respective MHC HLA alleles. The percentage of population coverage was calculated for 15 regions covering the globe

4.5.4   Designing a structural model for a TB vaccine

After selecting the final TB vaccine epitopes, the TB vaccine sequence was designed and the vaccine structural model was constructed to determine its efficacy of interaction inside the host.

4.5.4.1 Vaccine sequence design

The epitopes mentioned in Table 4.9 were used for designing the TB vaccine protein sequence.

Table 4. 9: Final 27 epitopes for TB vaccine construction

Epitope No.	CTL epitopes (E)	Epitope No.	HTL epitopes(E)	Epitope No.	B-cell epitopes(E)
1	FQGGGPHAV	15	LPVEYLQVPSPSMGR	20	KVRAERVSNEVN
2	LAMNDSGGY	16	PVEYLQVPSPSMGRD	21	SNAQSQHSSSNNSGGADPVD
3	RVSNEVNAY	17	SNGLGAAAAAEGSTH	22	RHQARRRRRAAAKERGNADE
4	SAAKGTGAY	18	LPVEYLQVPSPSMGR	23	TLTGCGSGDSTVAKTPEATP
5	NSATTSTGW	19	PVEYLQVPSPSMGRD	24	ITVGKDAPTT
6	HTGTNNSGY			25	QASPSRIDGTHQTLQGADLT
7	RTADGINYR			26	PVAGADPQRYDGDVPGMNYD
8	RIADHKLKK			27	AGADPEPAPTPKTAIDSDGT
9	SPKPATSPA				
10	RYGFKPTVI				
11	LQPQIDAAY				
12	KPAPSPAAA				
13	HEASPTQQL				
14	GTGTPTGDY				

The structure of the PADRE sequence and the 27 epitopes attached with the linker was constructed using PEPstrMOD. First, the affinity of the PADRE sequence with the five HTL epitopes was determined using PatchDock and FireDock. The PADRE-E17 (HTL epitope) was discovered to be the best combination compared to the other four HTL epitopes. The structure of PADRE-E17 was then constructed using the I-TASSER server. The binding compatibility of PADRE-E17 was determined for the remaining HTL epitopes. E-15 was found to have strong binding to PADRE-E17. The structure of the PADRE-E17-E15 combination was then generated using I-TASSER. This process was first undertaken for the HTL epitopes, followed by CTL epitopes and then the B-cell epitopes. Figure 4.10(i) shows the final TB vaccine sequence obtained through the extensive analysis of the TB epitope combinations.

The adjuvant 50S ribosomal protein L7/L12 at the N-terminal and β-defensin at the C-terminal were added to the vaccine protein sequence with the help of EAAAK linker. With the two adjuvants and linkers, the final length of the TB vaccine sequence was 629 amino acid residues (Figure 4.10 (ii)).
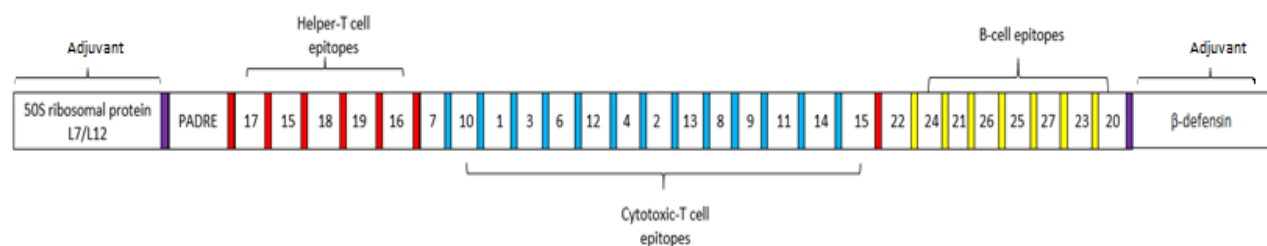
(i)

MAKLSTDELLDAFKEMTLLELSDFVKKFEETFEVTAAAPVAVAAAGAAPAGAAVEAAEEQSEFDVILEAAGDKKIGV
IKVVREIVSGLGLKEAKDLVDGAPKPLLEKVAKEAADEAKAKLEAAGATVTVKEAAAKAKFVAAWTLKAAAGPGPG
SNGLGAAAAAEGSTHGPGPGLPVEYLQVPSPSMGRGPGPGLPVEYLQVPSPSMGRGPGPGPVEYLQVPSPSMG
RDGPGPGPVEYLQVPSPSMGRDGPGPGRTADGINYRAAYRYGFKPTVIAAYFQGGGPHAVAAYRVSNEVNAYAA
YHTGTNNSGYAAYKPAPSPAAAAAYSAAKGTGAYAAYLAMNDSGGYAAYHEASPTQQLAAYRIADHKLKKAAYS
PKPATSPAAAYLQPQIDAAYAAYGTGTPTGDYAAYNSATTSTGWGPGPGRHQARRRRAAAKERGNADEKKITV
GKDAPTTKKSNAQSQHSSSNNSGGADPVDKKPVAGADPQRYDGDVPGMNYDKKQASPSRIDGTHQTLQGADL
TKKAGADPEPAPTPKTAIDSDGTKKTLTGCGSGDSTVAKTPEATPKKVRAERVSNEVNEAAAKGIINTLQKYYCRV
RGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKK

(ii)

Figure 4. 10: Epitope-based vaccine sequence construction scheme: (i) Schematic representation of TB vaccine construct consisting of B-cell and T-cells epitopes joined by flexible linkers and adjuvants at N- and C-terminals of the vaccine protein. The epitope combinations were found after determining the binding affinity with each of them using PatchDoack and FireDock, and (ii) TB vaccine protein sequence. The adjuvant sequence is highlighted in green, the PADRE sequence in pink, the flexible linkers in purple (EAAAK), red (GPGPG), blue (AAY) and yellow (KK). The HTL epitopes are highlighted in orange, the CTL epitopes are black, and the B-cell epitopes are brown

4.5.4.2 Assessment of properties of TB vaccine construct

After designing the TB vaccine sequence, its antigenicity, allergenicity and physiochemical properties were evaluated. The antigenicity was predicted using the VaxiJen server. The high antigenic score of 0.9126 showed the great immunogenic potential of the constructed TB vaccine. The AlgPred server indicated that the vaccine designed is non-allergenic to the host.
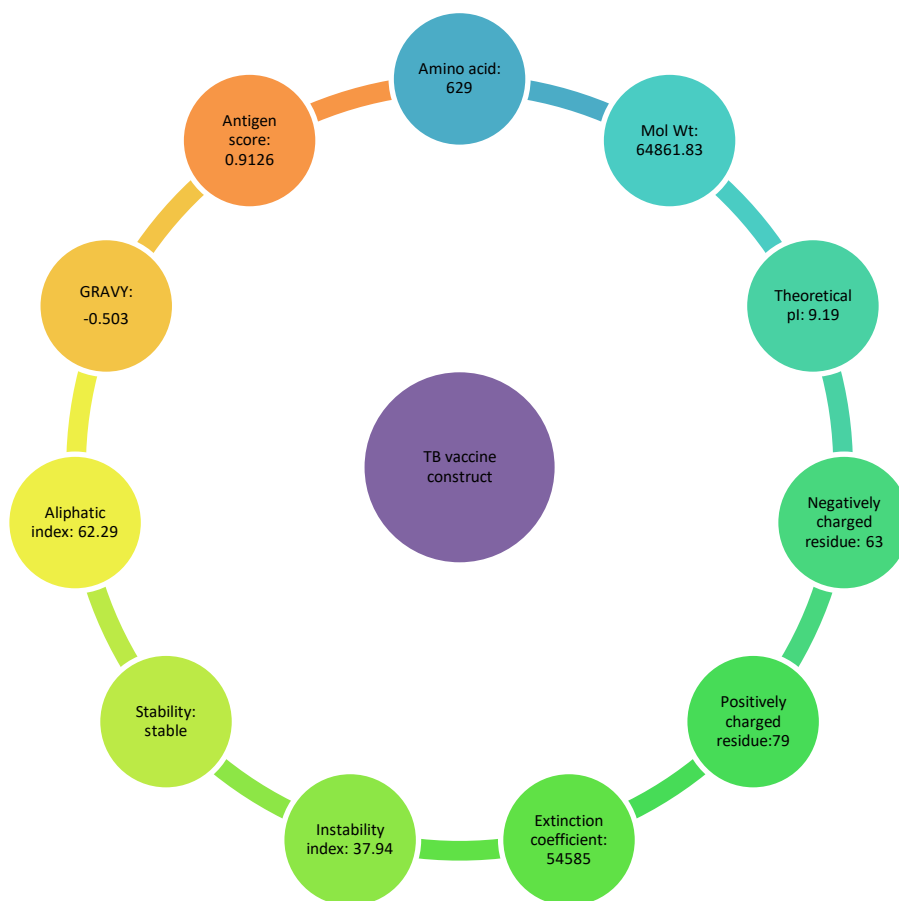
Figure 4. 11: Physiochemical properties of the final TB vaccine construct

ProtParam was used for predicting the physiochemical properties of the designed TB vaccine sequence. Proteins with MW less than 100 kDa indicate the possibility of experimentally studying the proteins in the laboratory for vaccine or drug development. Thus, proteins having a molecular weight of more than 100 kDa were not considered as potential vaccine targets. The vaccine sequence comprised 629 amino acid residues with a molecular weight of 64.86 kDa. The TB vaccine construct was composed of 9071 atoms and its molecular formula was $C_{2837}H_{4507}N_{817}O_{895}S_{15.}$ The isoelectric point (pI or pH(I)) is the pH at which a molecule carries no net electrical charge. The pI was assessed to be 9.19, indicating the vaccine construct to be slightly basic. *In vivo* half-life predicts the time taken by half of the amount of protein to disappear within a cell. A half-life between 2 to 100 hours is considered suitable for the degradation of protein, depending on the nature of the protein. The estimated half-life was found to be 30 hours in mammalian reticulocytes (*in-vitro*). An instability index with a value of 37.94 represented the stable nature of the TB vaccine. The higher value of the aliphatic index (62.29) indicated the higher thermostability of the protein. The GRAVY score of the vaccine was found to be -0.503. The negative value of the estimated GRAVY indicated the hydrophilic nature of the vaccine. After evaluating all the properties, we discovered that our designed TB vaccine had all the characteristics of a promising vaccine candidate required for initiating an effective immune response inside the host.

4.5.4.3 Structural model of the TB vaccine

Secondary structure is crucial for maintaining the three-dimensional structural stability of a protein. The secondary structure of the TB vaccine was constructed using PSIPRED. According to PSIPRED, 213 amino acid residues were involved in forming alpha-helixes, constituting 33.86% of the overall TB vaccine sequence. Only 66 residues participated in beta-strand formation (10.49%), whereas 350 amino acid residues formed the coils (55.64%) in the whole TB vaccine sequence. Secondary structural arrangement showed that the vaccine protein was highly flexible. The higher flexibility of the TB vaccine facilitates smooth interactions with the immune cells.
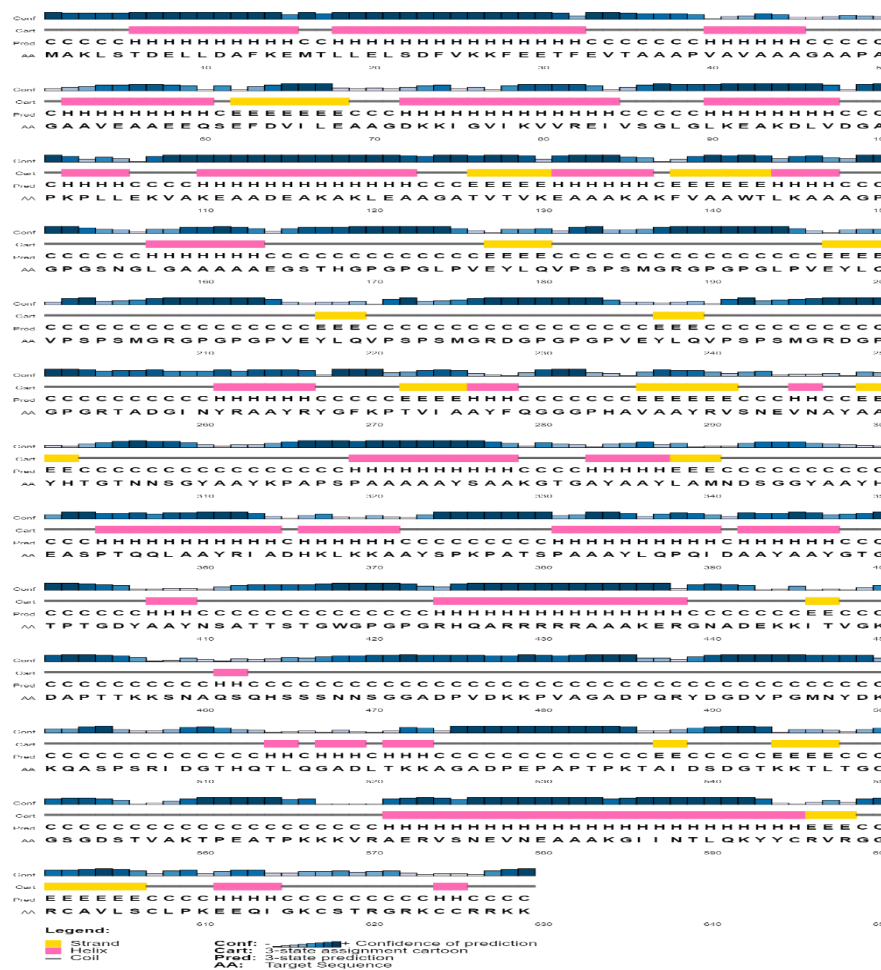
Figure 4. 12: Secondary structure prediction of the TB vaccine sequence using PSIPRED. The sequence comprised alpha-helixes (33.86%), beta-strands (10.49%) and coils (55.64%)

The tertiary structural model of the TB vaccine was constructed using the RaptorX server. RaptorX performed homology modelling and 1DD3_A PDB entry was used as a template for generating a structural model. All 629 amino acid residues were modelled with only 10% being in the disordered region. In the constructed model, 61% of the residue were exposed, 17% medium and 21% were buried

in folded conformation. Before refining the structural model, the Ramachandran plot showed that 87.6% of the amino acid residues were in the most favoured region. Thus, refinement of the structural model was carried out using the GalaxyRefine server to improve the quality of the predicted structural model. Figure 4.13 (i) shows the 3D structural model of the TB vaccine after performing the refinement step. The refined structural model had 92.3% amino acid residues in the most favoured region (Figure 4.13 (ii)).

Other parameters evaluated after refinement were GDT-HA (**g**lobal **d**istance **t**est- **h**igh **a**ccuracy), MolProbidity, clash score, RMSD and poor rotamers. GDT-HA score determines the backbone quality of the generated 3D structural models (Kopp et al., 2007). GDT-HA score ranges from 0 to 1 (Alapati et al., 2020). MolProbity score indicates the log-weighted combination of clash score, percentage of not favoured amino acid residues in Ramachandran plot and percentage of bad side-chain rotamers (Chen et al., 2010). The lower numerical value of MolProbilty score represents better quality of the three-dimensional structure. The results for the parameters evaluated after refinement were GDT-HA score of 0.9112, MolProbidity of 2.027, a clash score of 10.7, a RMSD 0.52 and poor rotamers with the value of 0.7.

(i)   (ii)
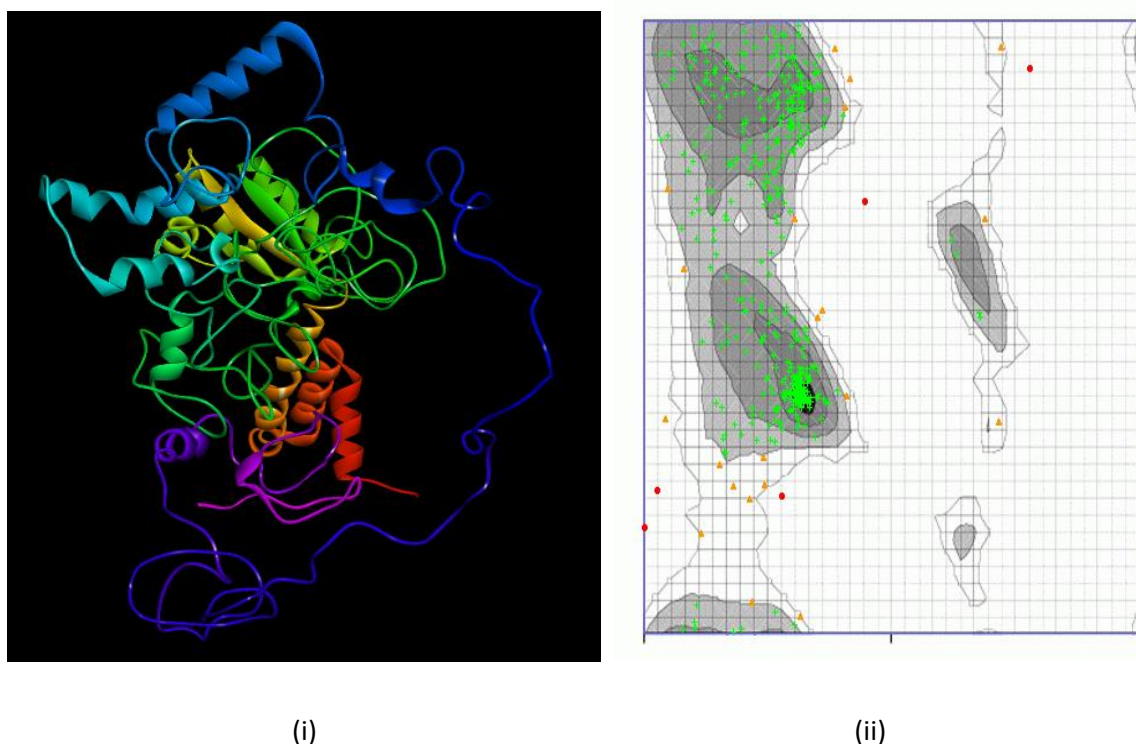
Figure 4. 13: Refined structural model of TB vaccine construct: (i) The 3D structure is coloured in rainbow colour from violet to red (from N-terminus to C-terminus), and (ii) the Ramachandran plot of the refined structure showing 92.3% amino acid residues in the most favoured region (green crosses)

4.5.4.4 Interaction of TB vaccine construct with toll-like receptors

The epitopes present on the surface of the pathogen interact with toll-like receptors (TLRs) present on the surface of the antigen-presenting cells (APCs), causing activation of the adaptive immune system. For understanding the interaction of the TB vaccine with TLRs, molecular docking was performed. The strong binding of both molecules will initiate the process of immune response generation. The ClusPro 2.0 server was used for performing molecular docking between the TB vaccine construct and TLR-2, TLR-4 and TLR-6 with PDB id 2Z7X, 4G8A and 4OM7, respectively (Figure 4.14). The output of the docking process displayed 30 clusters (0-29) for each docked complex. Cluster 0 of TLR-2 and the TB vaccine docked complex was found to have the lowest binding energy of -1172.8 kcal/mol. This cluster involved the highest number of members in the docking interaction, i.e., 66 amino acid residues of the vaccine interacting with TLR-2. Cluster 2 of the TLR-4 docked complex had the lowest binding energy of -1241.1 kcal/mol, with 52 residues involved in the interaction. For TLR-6 and the TB vaccine, cluster 0 had the lowest binding affinity of -1002.2 kcal/mol, with 38 members involved in the docking interaction. The docking results revealed that the designed TB vaccine had strong interaction with the selected TLRs and would accomplish the goal of initiating the immune response.

(i)   (ii)
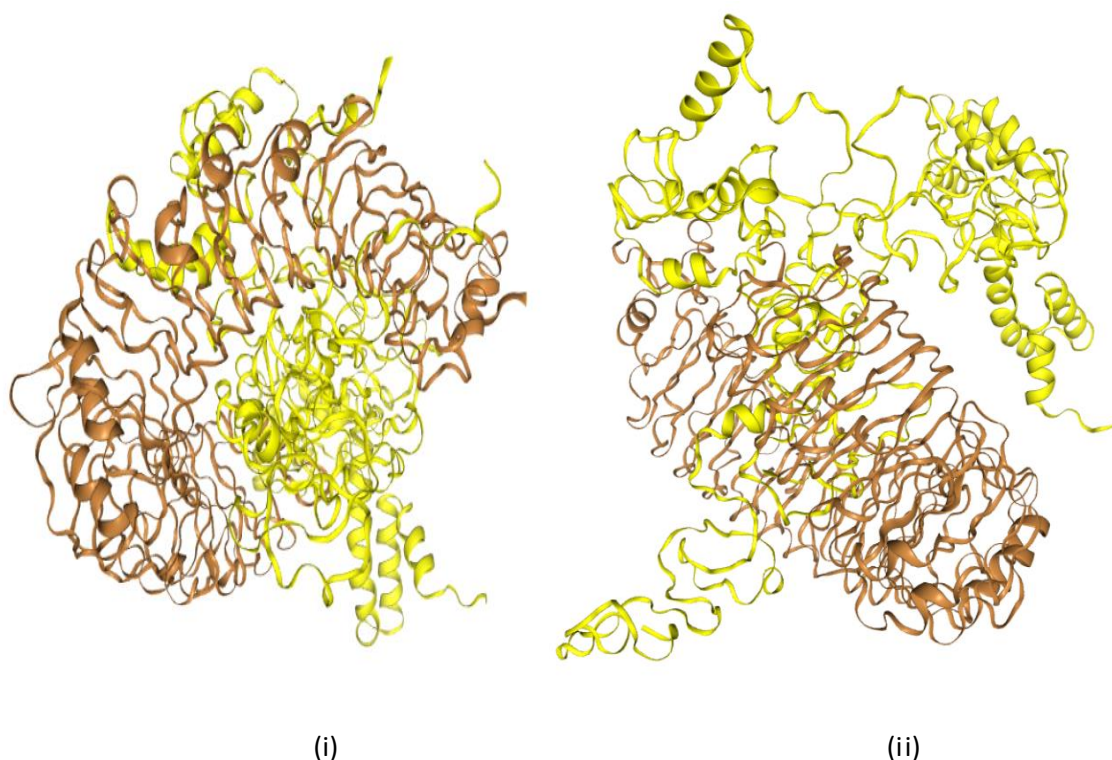
(iii)

Figure 4. 14: The docked complex of toll-like receptors (brown colour) with the TB vaccine construct (yellow colour). TB vaccine construct docked with: (i) TLR-2, (ii) TLR-4 and (iii) TLR-6

Normal mode analysis (NMA) was carried out to determine the stability of the docked complex and mobility of the TB vaccine residues within the complex on a large scale. The iMODS server was used to perform a dynamic simulation of the docked complex of the vaccine with TLR-2, TLR-4 and TLR-6. The presence of deformation in the docked complex of the TB vaccine and TLR will not activate the immune system. Deformability is the ability of amino acid residues to deform at a particular position within the protein. Analysis of the trajectory of the docked complex helps determine the occurrence of any deformation in the complex. The eigenvalue provided by the docked complex represents the energy required for deforming the complex. The eigenvalues found for the TB vaccine with TLR-2, TLR-4 and TLR-6 complex were 1.018829e-5, 1.304754e-05 and 2.501926e-5, respectively. The resulting eigenvalues indicated low chances of deformation for the TB vaccine-TLR docked complex. Thus, this analysis supports that our TB vaccine designed has the potential to generate a quick immune response.

(i)

(ii)

(iii)
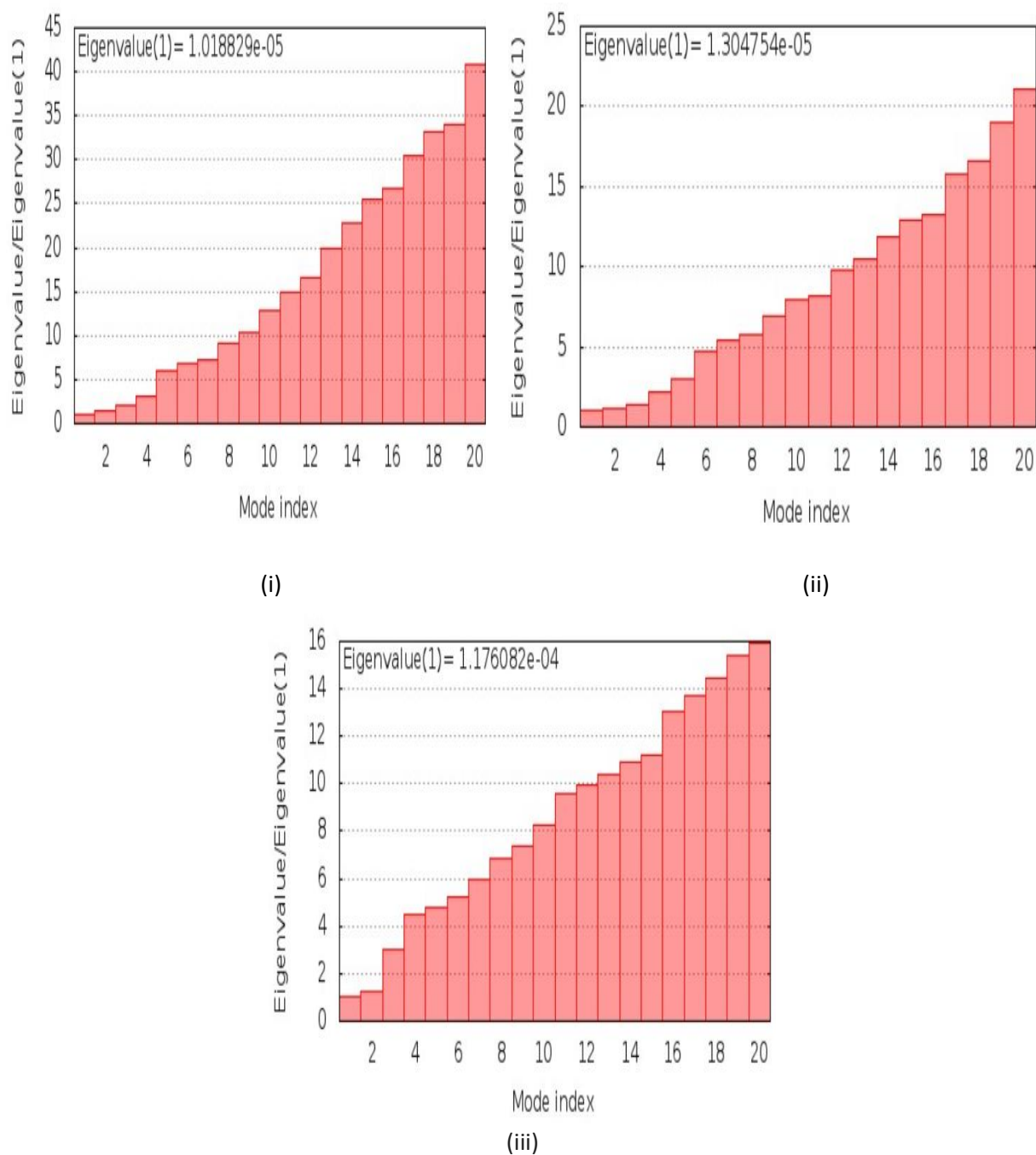
Figure 4. 15: Results of molecular dynamic simulation performed using iMODS. The eigenvalue of the docked complex of TB vaccine and toll-like receptors: (i) TB vaccine-TLR2, (ii) TB vaccine-TLR4, and (iii) TB vaccine-TLR6

4.5.5 Study of the immune response profile of the TB vaccine construct for predicting vaccine efficacy

Vaccines are cost-effective pharmaceutical products that have played an important role in eliminating and eradicating infectious diseases. The main goal of this study was to develop a cost-effective epitope-based TB vaccine that can provide long-lasting protection and herd immunity, reduce mortality rates,

fight against antimicrobial resistance and enhance the speed and strength of the host immune response against *Mycobacterium tuberculosis.* In the previous step, we designed an epitope-based TB vaccine and predicted its interaction with antigen-presenting cells. It is also essential to determine the efficacy of our epitope-based TB vaccine by studying the immune response it generated. We used the C-ImmSim server for examining the immune response profile of our vaccine construct. The vaccine construct was administered three times at different time intervals. The fixed time interval between each vaccine dose was four weeks.

After exposure to the vaccine construct, we found that the immune response generated by the host immune system was higher after every dose. Figure 4.16 (i) shows that the secondary and tertiary immune responses are higher than the primary response. The repeated exposure of TB epitopes via injections caused an increase in immunoglobulin activity (IgM, IgG1+IgG2, IgG1, IgG2 and IgM+IgG) after each exposure with rapid antigen clearance from the host. This shows that the host body is eliciting a robust immune response after encountering TB antigens. The rise in the population of active B-cells after each injection indicated an effective humoral immune response against *Mycobacterium tuberculosis* (Figure 4.16, (ii)). Several B-cell isotypes (IgM, IgG1 and IgG2) and memory B-cells were observed (Figure 4.16 (iii)). The presence of B-cells throughout the year indicates the formation of memory B-cell for a more extended period. This revealed the long-lasting humoral response against TB.

(i) (ii)

B cell population (cells per mm^3)

(iii)
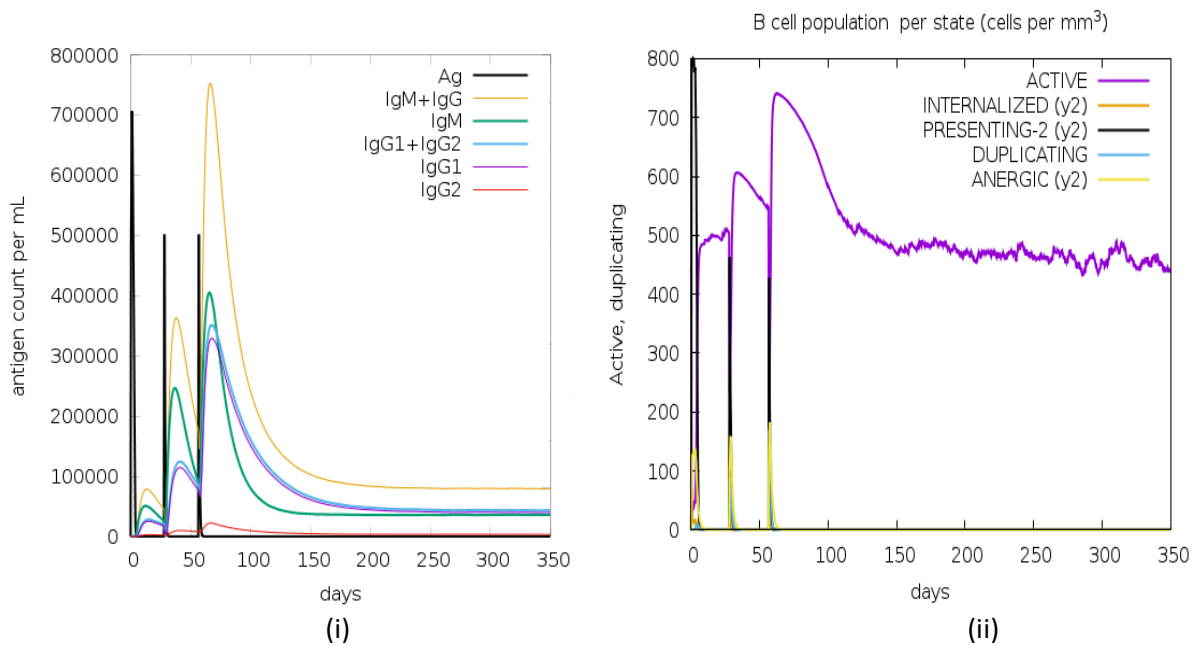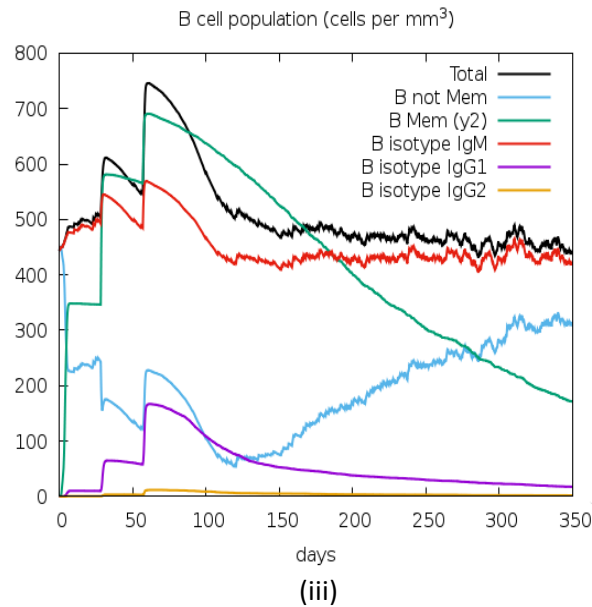
Figure 4. 16: Humoral immune response profile of the epitope-based TB vaccine construct: (i) generation of immunoglobulins upon exposure to TB vaccine construct, (ii) number of active B-cell populations per state of antigen exposure, and (iii) evolution of B-cells into different isotype populations after the administration of three TB injections

The number of active T-cells (helper T-cells and cytotoxic T-cells) significantly increased after the second and third exposures to the TB vaccine construct and slowly decreased while eliminating the antigen (Figure 4.17 (i), (iii)). The elevated level of CTL and HTL cells also activated the generation of memory CTL and HTL cells (Figure 4.17 (ii), (iv)). The memory cells generated in this process would play a crucial role in providing immune protection against *Mycobacterium tuberculosis*. The host body restores these memory cells and responds quickly on re-encounter with *Mycobacterium tuberculosis* by generating a robust immune response.

The activation of T-cells led to an increase in the concentration of cytokines and interleukins (Figure 4.17 (v)). The repeated exposure to the three TB injections maintained the high level of IFN- γ required for eliminating the TB antigens. IL-4, IL-10, IL-12, IL-23 and TNF-α are the key players needed to recruit mononuclear cells from nearby blood vessels to contain a TB infection. Thus, generation of all these cytokines would lead to the activation of macrophages and proliferation of helper T and cytotoxic T-cells required for complete the elimination of *Mycobacterium tuberculosis* antigens present inside the host body. A Simpson index (D) indicates the diversity in response generated. The lower value of the Simpson index implies greater diversity in the immune response of the host. The output of the immune simulation shows that immunization from our vaccine has generated required cytokines and interleukins required to eliminate TB antigens from the host body and activated sufficient number of B-cells and T-cells for long-lasting humoral and cell-mediated immune response.

TH cell population per state (cells per mm³)

(i)

TH cell population (cells per mm³)

(ii)

TC cell population per state (cells per mm³)

(iii)

TC cell population (cells per mm³)

(iv)

(v)
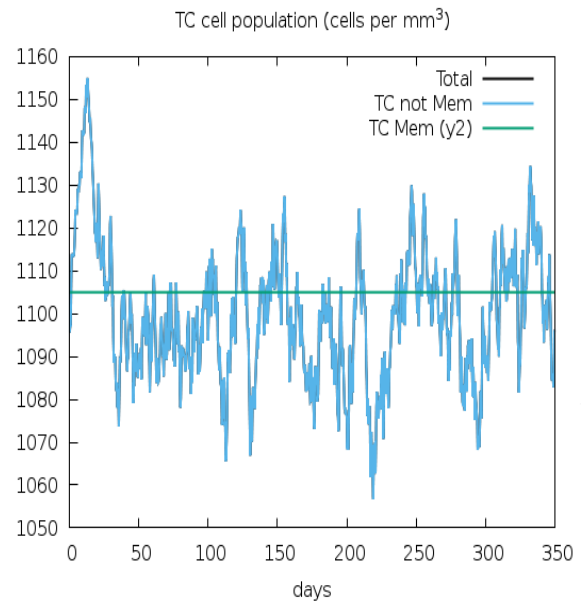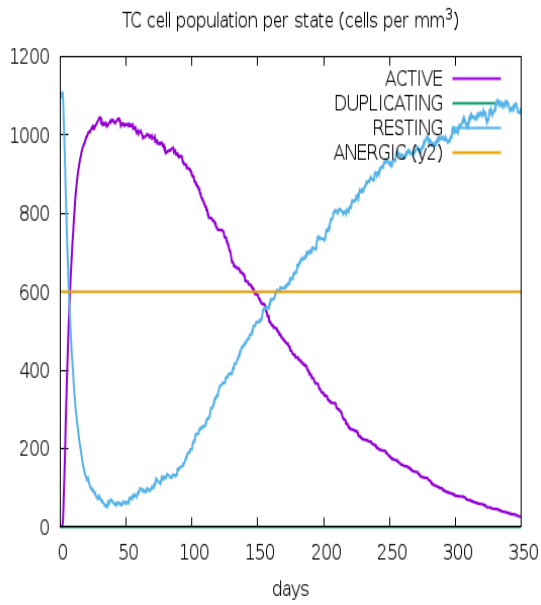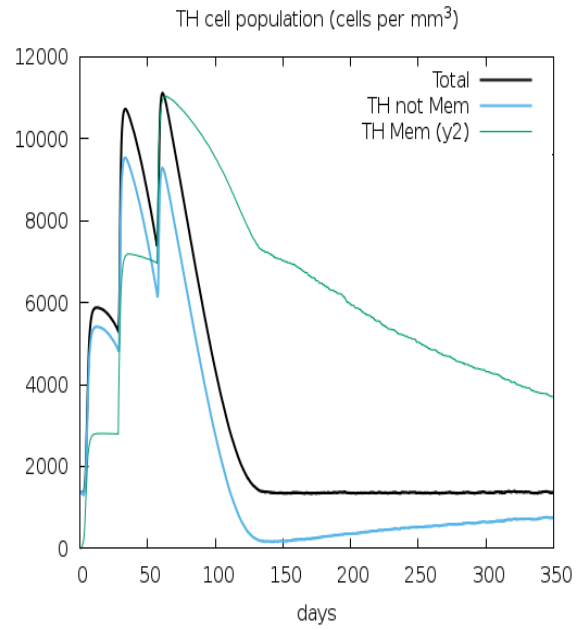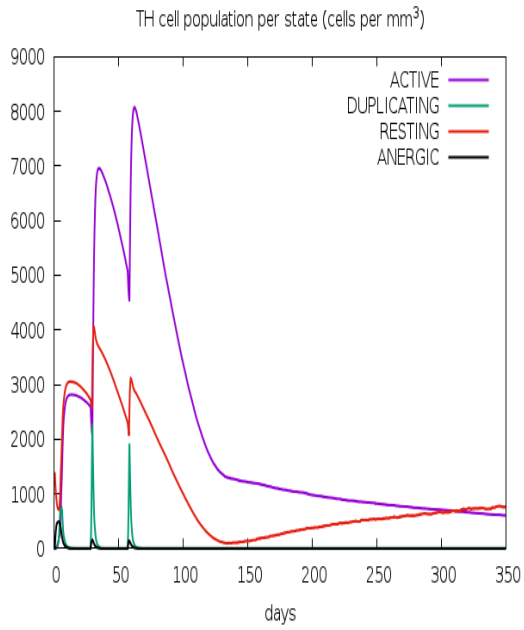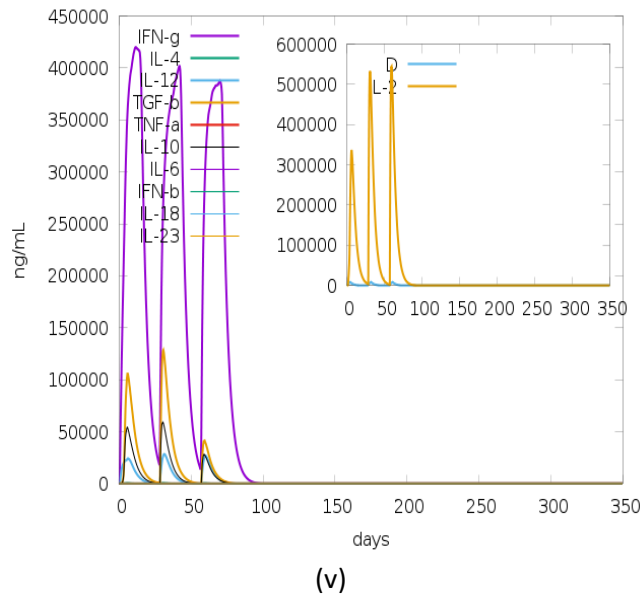
Figure 4. 17: Cell-mediated immune response profile of the epitope-based TB vaccine construct: (i) population of helper T-cells per state upon exposure to TB vaccine construct, (ii) population of memory helper T-cell per state, (iii) population of cytotoxic T-cells per state after exposure to the antigen, (iv) population of memory cytotoxic T-cell per state, and (v) production of different cytokines and interleukins with Simpson index (D) in three subsequent responses

The results obtained from stable interactions of the TB vaccine construct, TLRs and immune response simulation shows that our designed epitope-based TB vaccine would successfully initiate a swift and robust immune response against *Mycobacterium tuberculosis*. Thus, we have successfully constructed an *in-silico* epitope-based TB vaccine having all characteristics of the best possible vaccine.

4.6 Chapter Summary

Tuberculosis is an evolving deadly disease caused by the highly pathogenic *Mycobacterium tuberculosis.* TB is a highly contagious bacterial infection that spreads from an infected person to a healthy person through air by the inhalation of airborne droplet nuclei. Despite advances in the medical science, TB remains the cause of death of more than 1.4 million people every year. The World Health Organization considered tuberculosis a global threat with a significant mortality and morbidity rates. The BCG vaccine and drug therapy are the two most important countermeasures created by humans against TB. The evolution of *Mycobacterium tuberculosis* has led to the expansion of survival strategies and the emergence of drug-resistant TB strains that make drugs ineffective through various mechanisms. For example, an impermeable cell wall prevents drug entry into the cells, or a mutation in the target protein can lead to inactivation of drug molecules with the help of bacterial enzymes. Therefore, vaccination is still a promising strategy to protect the human population and bring down the incidence rate of TB. BCG is the only licensed vaccine available that is prepared from a live-

attenuated strain of *Mycobacterium bovis.* The low efficacy of BCG, the reemergence of the disease in immunocompetent individuals and drug-resistant *Mycobacterium tuberculosis* strains have generated an urgent requirement for a powerful and effective TB vaccine.

Today, after 100 years of research in the field of TB vaccines, the process of designing and developing an ideal universal TB vaccine is still hindered by several challenges, such as pathogen polymorphism, eliciting a precise immune response against TB, genetic diversity of the human population, hypersensitivity, safety and time. In the present study, we designed an epitope-based TB vaccine by developing and testing a conceptual framework to address the research challenges of the conventional vaccine development process.

With the aim of developing an effective TB vaccine that would elicit a robust immune response inside the host, we focussed on selecting the highly immunodominant epitopes from the conserved and surface-exposed antigenic proteins of *Mycobacterium tuberculosis* for constructing the TB vaccine. In recent years, bioinformatics tools and software have offered many significant breakthroughs in computational vaccinology. Our study incorporates different branches of bioinformatics such as comparative proteomic analysis, reverse vaccinology, immunoinformatics and structural vaccinology, for identifying potential vaccine candidates and designing an epitope-based TB vaccine.

Using a single strain or single antigen does not offer a complete picture of the genetic diversity of *Mycobacterium tuberculosis*. Thus, we used the proteome of 159 completely sequenced strains of *Mycobacterium tuberculosis* to cover the diversity and identify conserved proteins among those strains. *Mycobacterium tuberculosis H37Rv* was used as a reference proteome. By performing comparative proteomic analysis using a standalone BLAST, we found that out of 3906 proteins of H37Rv, 1982 proteins were conserved among the 159 strains of *Mycobacterium tuberculosis*. The functional classification of the 1982 conserved proteins was undertaken for understanding the impact of evolution on the biological processes and functions, such as physiology, metabolism and translation of proteins of *Mycobacterium tuberculosis*. We observed a high percentage (73%) of conserved proteins involved in critical functional classes, such as intermediary metabolism and respiration, cell wall processes, information pathways, lipid metabolism and regulatory proteins and virulence, detoxification and adaptation. This suggests that *Mycobacterium tuberculosis* experiments by mutating fewer essential proteins, thus, not risking the normal functioning and structural stability needed for its survival.

After the identification of conserved proteins, a reductionist reverse vaccinology process was performed. Reverse vaccinology is a low-cost technique, entirely feasible for use on the plethora of

genomic or proteomic data being generated. Reverse vaccinology helps identify surface-exposed, antigenic proteins having unique characteristics like signal peptides, membrane-spanning regions, lipoprotein signatures and adhesion probabilities that are crucial for the survival of *Mycobacterium tuberculosis* and non-homologous to the host proteome or host metabolic pathways. Thus, we performed comprehensive reverse vaccinology analysis using several bioinformatics tools with reliable accuracy to scrutinize the ideal vaccine candidates by directly analysing the conserved proteins of the *Mycobacterium tuberculosis* H37RV strain. Out of 1982 conserved proteins, a total of 24 membrane-spanning, antigenic and non-allergic proteins were selected for the reverse vaccinology approach. These 24 shortlisted proteins were further examined for determining their efficacy in developing an epitope-based vaccine that can stimulate a specific and swift humoral and cell-mediated immune response.

The epitope-based vaccine provides a powerful new strategy for pathogen-specific immunity. Here, we have used the immunoinformatics approach for identifying safe and immunodominant epitopes that could stimulate an innate, humoral and cell-mediated immune response in the host. From an extensive repertoire of TB epitopes predicted from immunoinformatics analysis, we shortlisted 27 epitopes (CTL epitopes-14, HTL epitopes-5 and B-cell epitopes-8) from 18 antigenic *Mycobacterium tuberculosis* proteins. These 27 epitopes were highly immunogenic, non-toxic and non-allergenic to the host. Population coverage analysis showed that 99.16% of the world's population were covered by the predicted CTL and HTL epitopes from the 18 antigens of *Mycobacterium tuberculosis*. This analysis strengthens the confidence in our strategy of developing a universal TB vaccine.

Structural vaccinology then helped in developing an *in-silico* vaccine. The design of the vaccine sequence was based on a new concept introduced in this research. The docking analysis of the 27 filtered epitopes with every other epitope created distinct combinations of epitopes. The combinations that had a strong binding affinity to one another were used for constructing vaccines. The epitopes were attached with the help of flexible linkers (GPGPG, AAY and KK). After arranging the epitopes based on their binding affinity, the final vaccine sequence was completed using two adjuvants, 50S ribosomal protein L7/L12 and β-defensin attached at the N- and C-terminals, respectively, with the help of the EAAAY linker. After designing the TB vaccine sequence, the antigenicity, allergenicity and physiochemical properties were evaluated. After assessing all the properties, we discovered that our developed TB vaccine had all the characteristics of a promising vaccine candidate required for initiating an effective immune response inside the host.

The structural model of the TB vaccine was constructed using RaptorX. The model was then refined to improve the quality of the vaccine's structure. The structural analysis of the model of the TB vaccine

construct highlighted the structural integrity with more than 92% amino acid residues in the most favourable region. Then, molecular docking and dynamics simulation of TB vaccine construct and TLRs (TLR-2, 4 and 6) was carried out. A strong interaction between the pathogen and host TLR can help in initiating a strong immune response. The results of our docking analysis suggested stable and robust interactions between the TB vaccine construct and TLRs with binding energies -1172.8 kcal/mol (TB vaccine-TLR2), -1241.1 kcal/mol (TB vaccine-TLR4) and -1002.2 kcal/mol (TB vaccine-TLR6). The results of an analysis predicted a minimum level of deformability between the vaccine construct and TLRs. This indicates that the constructed epitope-based TB vaccine would have a strong interaction inside the host, thus activating the macrophages, further leading to the activation of cell-mediated and humoral immunity. Finally, evaluating our constructed TB vaccine's immune response profile helped predict the vaccine's efficacy in generating a strong and specific humoral and cell-mediated immune response. The production of B-cells, T-cells and cytokines after exposure to the TB antigen showed the vaccine's efficacy in generating an immune response. Extensive analysis of the results suggests that the epitope-based TB vaccine we developed has a high potential of evoking a specific immune response to provide broad immune protection against many *Mycobacterium tuberculosis* strains. A further trial in the laboratory in suitable *in vitro* and *in vivo* models is recommended to validate our prediction of safety, efficacy and immunogenicity.

Chapter 5

Identification of therapeutic vaccine and drug targets for bovine tuberculosis treatment

In chapter 5, we accomplished objective 3 of our study. We had briefly discussed the need for efficient treatments for bovine TB in chapter 1. In this chapter, we will use the framework developed in chapter-4 for designing a bovine TB vaccine. We also developed a method for identifying potential anti-bovine TB drug targets and tested the framework using bioinformatics tools for providing a solution to reduce the burden of bovine TB worldwide. An overview of bovine tuberculosis is provided in section 5.1. In Section 5.2, the control measures used in bovine TB, problems with them and formulation of the framework for potential therapeutics, are discussed. Section 5.3 explains the step-by-step process of the computational framework developed and the bioinformatics tools used for identifying vaccine and drug targets for bovine TB. In section 5.4, the results of the extensive computational analysis are discussed. Finally, a summary of the chapter is presented in the last section (Section 5.5).

5.1   Bovine Tuberculosis

Bovine tuberculosis is a chronic infectious disease that primarily affects cattle, but other livestock, such as deer, goats, horses and sheep, are also affected by the bacteria (Kuria, 2019). Bovine TB is a zoonotic disease that can spread to humans directly by inhaling aerosols or, indirectly, by ingesting unpasteurised milk. Twenty to thirty per cent of the global livestock population is potentially affected by bovine TB, leading to annual economic losses of more than USD 3 billion globally (Kuria, 2019). There is significant loss of livestock and trade restrictions due to bovine TB, e.g., in New Zealand, the beef and dairy industries are at potential risk of TB (Price-Carter et al., 2018). Bovine TB also decreases milk production by four and 20 per cent worldwide.

Bovine TB is caused by the etiological agent *Mycobacterium bovis.* It is a rod-shaped, intracellular, aerobic, gram-positive and a slow-growing bacterium. The *Mycobacterium bovis* cell wall comprises covalently linked peptidoglycans, arabinogalactans, non-peptidoglycan amino acids and a glucan (Petit et al., 1975). The cell structure and metabolism of *Mycobacterium bovis* is similar to *Mycobacterium tuberculosis*. The main *in-vitro* difference that occurs in *Mycobacterium bovis* is a point mutation in *pykA* that affects the binding of the Mg^{2+} cofactor with pyruvate kinase in the final step of glycolysis. Pyruvate kinase cannot catalyse the reaction and the glycolytic metabolites are not transferred to oxidative metabolism. Thus, *Mycobacterium bovis* rely on fatty acids or amino acids for growth and development in the laboratory (Garnier et al., 2003).

The genome of *Mycobacterium bovis* is 99.95% identical to *Mycobacterium tuberculosis* (Garnier et al., 2003). A fully virulent strain, *Mycobacterium bovis* AF2122/97, was isolated from an infected cow in 1997 (Garnier et al., 2003). The complete information on the genome sequence of *Mycobacterium bovis* AF2122/97 was published in 2003. The *Mycobacterium bovis* genome contains a single, circular chromosome containing 4,345,492 base pairs with a high G+C content of 65.6% (Garnier et al., 2003). Around 4043 genes are present, encoding 3988 proteins in *Mycobacterium bovis* AF2122/97. Developments in high-throughput experimental methods have made completed genomes available for different *Mycobacterium bovis* strains.

Cattle are the primary host of infection, but *Mycobacterium bovis* can infect many hosts, including humans as well as domestic and wild mammals (Kuria, 2019). The reported susceptible domestic species are sheep, goats, pigs, deer, horses, camel, cats, dogs and ferrets (The Centre for Food Security and Public Health, Iowa State University). Wild mammals infected by *Mycobacterium bovis* are bears, African buffalo, elephants, raccoons, primates, possums, rhinoceros, foxes and rodents (Thomas et al., 2021). There is little information available about the susceptibility of birds to *Mycobacterium bovis.* A maintenance host can be defined as a species endemic to infection that transmits the disease to other animals by direct contact (Haake & Levett, 2015). There are several maintenance hosts for *Mycobacterium bovis,* such as brush-tail possums (New Zealand), badgers (United Kingdom and Ireland), bison and elk (Canada) (Thomas et al., 2021).

Mycobacterium bovis is usually transmitted by the inhalation or ingestion of droplet nuclei. The infection can also be transmitted through other bodily fluids such as urine, saliva, milk and colostrum. The transmission of *Mycobacterium bovis* depends on several factors such as the frequency of excretion, infective dose through coughing, period of communication with the infected host and the host's susceptibility (Good & Duignan, 2011). Humans get infected by *Mycobacterium bovis* by ingesting unpasteurised milk and raw or undercooked meat (Kuria, 2019). A person working in dairy farms or slaughter-houses is more susceptible to TB. The infected person can transmit the infection to other people, but the transmission rate of bovine in humans is low (Davies, 2006). It has been seen that transmission of TB from humans to animals is rare (The Centre for Food Security and Public Health, Iowa State University).

Bovine TB is found globally, but eradication programmes have helped eliminate or nearly eliminate this chronic disease from domestic animals in several countries. The countries reported as TB-free include Iceland, Denmark, Sweden, Switzerland, Norway, Canada, Singapore, Australia and Finland (The Centre for Food Security and Public Health, Iowa State University). The World Organization for

Animal Health (OIE) has a total of 181 member countries. In 2017, 78 countries reported the prevalence of bovine TB of the 181 OIE countries (World Organization for Animal Health, 2018). Reis et al. (2020) studies showed the prevalence of bovine TB is more in Ireland (22.87%) and the United Kingdom (16.43%), compared to France (5.03%) and Spain (3.05%). The disease prevalence is still high in Africa and some parts of Asia. Srinivasan et al., 2018 estimated that 21.8 million cattle are infected with bovine TB in India. Currently, several eradication programmes focussing on eliminating the maintenance hosts are being implemented in the United Kingdom, the U.S.A, Mexico, Japan and New Zealand. The precise estimation of the prevalence of bovine TB remains poorly understood at the global level due to poor diagnosis of livestock and domestic animals, and the lack of surveillance data from most countries in the world. *Mycobacterium bovis* infections detected in wildlife can cause severe health implications to other organisms living in the same ecosystem. Moreover, bovine tuberculosis has zoonotic potential, raising health concerns for the public (Renwick et al., 2007).

The main factors associated with the prevalence of bovine tuberculosis are the sex of the species, breed of cattle and living conditions (Biffa et al., 2012; O'Reilly & Daborn, 1995; Torres-Gonzalez et al., 2013). Male badgers have a high risk of transmitting the infection to cattle (Nugent et al., 2018). Poor sanitation and no isolation of infected cattle from the herd on the farm can lead to exposure of *Mycobacterium bovis* to humans. Out of 10 million incidence cases in humans in 2019, WHO estimated 0.14 million cases were zoonotic TB caused by *Mycobacterium bovis*, with 11,400 human deaths (WHO Global Tuberculosis Report, 2020; World Organization for Animal Health, 2019).

The immune response generated by cattle against *Mycobacterium bovis* is similar to the human immune response against *Mycobacterium tuberculosis* because of the similarity of the mammalian immune system and cells. In cattle, a primary infection results in a lesion in the nasopharynx and upper parts of the lungs (Cousins, 2001). The alveolar macrophage is the primary host for intracellular growth of *Mycobacterium bovis* and performs phagocytosis. In phagocytosis, the alveolar macrophage ingests *Mycobacterium bovis* forming a phagosome, which later, together with a lysosome, forms a phagolysosome complex. The formation of phagolysosome causes the breakdown of *Mycobacterium bovis* into smaller fragments with the help of hydrolytic enzymes present in the lysosome. In the lung, alveolar macrophages play a vital role in interacting with innate and acquired immune cells. After an initial infection, the alveolar macrophages present the bacterial fragments (antigens) to T-helper cells leading to activation of cell-mediated immunity. The release of anti-inflammatory cytokines IL-4, IL-5 IL-10 and IL-13 by T-helper cells promotes the activation of B-lymphocytes leading to antibody production. The B-cells and T-cells surround the site of primary infection and form a mass of tissue called a granuloma.

In the initial stages of bovine TB infection, cattle are asymptomatic, so it can take a few months to years for any sign or symptom to be observed. As the infection progresses, the infected animal shows clinical signs that include weakness, fluctuating fevers, prominent lymph nodes, a moist cough with increased breathing rate, loss of weight, emaciation, diarrhoea and induration of the udder. The standard approaches for diagnosing bovine tuberculosis include skin testing, microscopic examination, culturing of bacteria and the nucleic acid recognition method (Kuria, 2019). Microscopic analysis of bovine bacteria obtained from clinical samples is performed by staining direct smears using the Ziehl-Neelsen stain. Definitive diagnosis of *Mycobacterium bovis* is made by culturing bacteria in the laboratory for at least eight weeks. A tuberculin skin test is the primary screening test for diagnosing TB in cattle. A small amount of a purified-protein derivative is injected into the skin to measure the adaptive immune response in the tuberculin test. Polymerase chain reaction (PCR) assays confirm infections in cattle by comparing interferon-gamma levels in blood samples.

5.2 Control measures for eradicating bovine TB

The techniques involved in controlling bovine tuberculosis in wildlife are restricted. The isolation of infected animals is not an option in hugely populated or low-income countries. Pasteurisation of milk is not compulsory in India. Thus, bovine TB in cattle also impacts human health. It is also important to remember that getting infected from eating meat of the infected animal is less likely, but the risk is still there. In New Zealand, possums serve as the significant wildlife reservoir for *Mycobacterium bovis* infection and, since 1994, there has been a marked increase in the implementation of possum poisons (Corner & Norton, 2003). This resulted in a 70% reduction in the prevalence of *Mycobacterium bovis*-infected cattle herds from June 1994 to June 2001. Animal test-and-slaughter schemes implemented in several countries have successfully reduced the prevalence of bovine tuberculosis. Still, such expensive control programmes have raised economic burdens and increasing opposition by the farmers (Bennett, 2009; Torgerson & Torgerson, 2010).

Currently, there is no effective treatment available for bovine TB due to its infectious nature and drug resistance of *Mycobacterium bovis*. The available treatment of bovine TB mainly depends on the health status of the infected species. Antibiotic therapy can be used for animal species living in captivity, but this is not reliable for herd or free-grazing animals. *Mycobacterium bovis* is considered naturally resistant to pyrazinamidase (first-line TB drug) (Nakajima et al., 2010). First-line human TB drugs for treating livestock are also ineffective and costly as the treatment requires six to nine months of daily medication. BCG vaccine is another option available for treatment of the disease, but it shows limited efficacy in cattle. BCG is prepared from a live-attenuated strain of *Mycobacterium bovis.* It is produced by using living *Mycobacterium bovis* that has been weakened or attenuated under specific laboratory

conditions so, while it cannot cause the disease, it retains the ability to generate an immune response within the host. BCG has been around for almost a hundred years and was first used in 1921.

As the lives of humans and animals are interrelated, it is essential to implement specific therapeutic measures to reduce the burden of bovine TB in the world. The prevention of bovine TB is a long-term goal that can only be accomplished by developing more effective therapeutics against *Mycobacterium bovis*. Our study aims to identify potential vaccine and drug targets for bovine TB treatment using a number of bioinformatics approaches. The current explosion in bioinformatics has revolutionized the field of vaccine and drug development and will provide new tools that facilitate identification of potential vaccine and drug targets without the need to culture the pathogens in the laboratory. Information about the genome, transcriptome or proteome of *Mycobacterium bovis* can help in the identification of novel therapeutic candidates.

Vaccination is still a promising strategy to protect livestock and reduce the incidence rate of bovine TB. The goal of vaccinating cattle against tuberculosis is to prevent the establishment of the infection in these animals. Interest in the development and use of TB vaccines for cattle can be rekindled with an improved understanding of the protective immunity against TB. The development of a new vaccine against bovine TB is considered as a control option to address two important issues: preventing the establishment of the disease, and the elimination of zoonotic bovine TB. The goal of our study is to design an improved vaccine that can provide protection against bovine TB disease. Generating effective memory cells against bovine TB would be a crucial step towards tackling the disease worldwide. Epitope-based vaccine is a powerful strategy for this. The development of epitope-based bovine TB vaccine containing B-cell and T-cell (MHC-I and MHC-II restricted) epitopes could elicit a humoral and cell-mediated immune response. For designing an epitope-based bovine TB vaccine, we used the method developed in chapter-4. Section 5.3.1 provides the details of the method's application for *Mycobacterium bovis*.

For identifying the therapeutic drug candidates for bovine TB treatment, we developed a new bioinformatics approach to answer pathogenicity and drug resistance against bovine TB. The emergence of antimicrobial resistance (AMR) has reduced the efficacy of antibiotics in treating diseases. To overcome the AMR issue, we used an approach that identifies conserved and pathogenic drug targets to design better drug therapeutics against bovine TB. Various bioinformatics approaches can be used for examining the proteome of *Mycobacterium bovis* to identify potential drug targets that can facilitate drug development for bovine TB treatment. A subtractive proteomic approach was used to determine the conserved, essential, antigenic and druggable targets having unique metabolic pathways in *Mycobacterium bovis* (Figure 5.1). The detailed methods for vaccine development and

drug target identification are explained in section 5.3.1 and 5.3.2, respectively. The strategy developed for identifying drug targets is generic and can be used for other zoonotic infectious diseases.

5.3 Method for identifying potential therapeutic bovine TB vaccine and drug candidates

5.3.1 Epitope-based bovine TB vaccine design and development

This study designed an epitope-based bovine TB vaccine to stimulate a specific and swift humoral and cell-mediated immune response. For this, we used the conceptual framework developed in chapter-4. This section of the chapter briefly describes the step-by-step approach for designing a bovine TB vaccine. We used several bioinformatics tools to identify ideal vaccine candidates by directly analysing the proteome of the *Mycobacterium bovis* strain.

5.3.1.1 Subtractive reverse vaccinology analysis

Our study used 11 strains of *Mycobacterium bovis* to identify the conserved antigenic vaccine targets to solve antigen variability, drug resistance, and to provide a broad coverage. Reverse vaccinology was used to determine the outer-membrane antigenic and non-allergenic proteins with unique characteristics, such as, signal peptides, membrane-spanning regions, lipoprotein signatures and adhesion probability, from the proteome of *Mycobacterium bovis.* These antigenic proteins would help in developing a vaccine for bovine TB.

The proteome of the completely sequenced 11 strains of *Mycobacterium bovis* was downloaded in FASTA format via the NCBI Genome FTP site. Blank spaces and unwanted information were removed and multi-line protein sequences were converted into single line FASTA format. Standalone BLAST (Altschul 1997; Altschul et al., 1990) was downloaded from the NCBI FTP site for performing a homology search. *Mycobacterium bovis* AF2122/97 was used as a reference proteome to identify conserved proteins with more than 99% sequence similarity across all 11 strains. The conserved protein sequences identified were selected for analysis using reverse vaccinology.

The outer membrane and extracellular space of *Mycobacterium bovis* consist of some essential secretory and lipoproteins that help in the translocation of proteins across the cytoplasmic membrane. The proteins present in the cell membrane and extracellular space were predicted using six SCL predicting tools: PSORTb v.3.0 (Yu et al., 2010), CELLO (Yu et al., 2006), LocTree3 (Goldberg et al., 2014), SOSUI (Imai et al., 2008), pLoc_bal-mGpos (Xiao et al., 2019) and GramLocEN (Wan et al., 2017). Those selected were analysed for the presence of a transmembrane-helix. Proteins having a single transmembrane helix were chosen for designing a bovine TB vaccine. TMHMM server was used to predict transmembrane regions and their orientation (Krogh et al., 2001). Next, we used SignalP 4.1

(Petersen et al., 2011) and SecretomeP (Bendtsen et al., 2005) to predict secretory proteins. PRED-TAT (Bagos et al., 2010) was used for identifying TAT-signal peptides and lipoproteins were identified using PRED-LIPO (Bagos et al., 2008).

The development of a vaccine for bovine TB requires identifying an antigenic protein that can stimulate a precise immune response in the host. VaxiJen (Doytchinova & Flower, 2007), VirulentPred (Garg & Gupta, 2008) and MP3 (Gupta et al., 2014) tools were used for determining the antigenicity of the selected proteins. The proteins predicted antigenic by all three methods were chosen for further analysis. Algpred (Saha & Raghava, 2006a) was then used to determine the allergenic nature of the selected antigenic proteins. Non-allergenic proteins were chosen. The adhesin proteins of mycobacteria help in its attachment to host cell-surface receptors. The adhesion probability was predicted with the help of SPAAN (Sachdeva et al., 2005). The proteins with a probability score of 0.5 or above were selected for further analysis. For predicting homology, BLASTp was used for identifying homologous proteins between cattle and *Mycobacterium bovis.* Proteins with a sequence identity of 30% or more and a bit score of more than 100 should be eliminated. The homologous proteins that could cause autoimmunity or any hypersensitive reactions in the host were excluded from the study.

5.3.1.2 Prediction of B-cell and T-cell epitopes

The antigens identified using the reverse vaccinology approach were further analysed for predicting potential B-cell and T-cell epitopes for developing an effective vaccine against bovine tuberculosis.

5.3.1.2.1 B-cell epitope prediction

B-cell epitopes play a vital role in initiating the humoral immune response. ABCpred was used for the prediction of the B-cell epitopes. ABCpred uses a recurrent neural network for B-cell epitope prediction (Saha & Raghava, 2006b). The length of B-cell epitopes was set to 20 amino acid residues

5.3.1.2.2 T-cell epitope prediction

MHC molecules are an important class of proteins present on the surface of antigen-presenting cells. The function of the MHC molecules is to present the fragmented or processed antigen to the appropriate T-cell (HTL or CTL) of the immune system. This initiates the cell-mediated immune response and secretion of the cytokines to eliminate the infection from the host.

(i) Prediction of MHC-I binding T-cell (cytotoxic-T cell) epitopes

MHC class I molecules are present on the surface of all nucleated cells in the host body. The MHC class I molecules present the epitope to cytotoxic T-lymphocytes (CTL). The selection of epitopes that bind strongly to MHC-I is the most crucial step in the vaccine design process. It helps to predict the most antigenic and immunodominant epitopes required for initiating the immune response. To predict

MHC-I restricted T-cell epitopes, we used the following bovine MHC-I HLA alleles: BoLA-D18.4, BoLA-AW10, BoLA-JSP.1, BoLA-HD6, BoLA-T2a, BoLA-T2b and BoLA-T2c. We used two methods in this step: IEDB MHC I (Zhang et al., 2008) and NetMHCpan 4.1 server (Reynisson et al., 2020). In the IEDB MHC I method, the prediction was made using an artificial neural network for determining the binding affinity of epitopes with bovine HLA alleles. Nine-mer epitopes with a percentile rank lower than 1.0 were chosen. The lower value of percentile rank indicates a higher affinity of epitope towards the MHC molecule. The NetMHCpan 4.1 server predicts the peptides binding to MHC class-I molecule. The method uses an artificial neural network trained with 201 different MHC alleles of humans, mice, cattle, primates and swine. The technique has shown a strong preference for the 9-mer peptide (Reynisson et al., 2020).

(ii) MHC-II binding T-cell (helper-T cell) epitope prediction

MHC class-II molecules are present on the surface of antigen-presenting cells, such as macrophages and dendritic cells. The primary function of MHC-II is to present antigens to the naïve T-helper cells. This interaction leads to the release of cytokines that help develop naïve-T- helper cells into effector T-cells or memory T-cells. The method used for the prediction of HTL epitopes was the IEDB MHC II server. This helped predict 15-mer MHC-II epitopes using the consensus prediction method (Wang et al., 2010). The HLA alleles of the mouse model were used in this step as no information was available for bovine MHC-II molecules. The epitopes with a percentile rank lower than 1.0 were chosen.

5.3.1.3 Filtering of epitopes

B-cell and T-cell epitopes were filtered to identify immunodominant epitopes that were antigenic, non-toxic and non-allergenic. VaxiJen (Doytchinova & Flower, 2007) was used to predict the epitope having the potential to initiate an immune response. Epitopes with a VaxiJen score of 0.7 or above were selected. Non-toxic epitopes were predicted using the ToxinPred (Gupta et al., 2013). AllerTOP 2.0 predicts the allergenicity of epitopes based on the physiochemical properties of a protein's sequence (Dimitrov et al., 2014). The allergenic epitopes were excluded from the study. The hydrophilic epitopes are present on the surface of the antigenic proteins. The selection of hydrophilic epitopes would help in constructing a vaccine that would initiate a quick immune response. The grand average of hydropathicity (GRAVY) (Kyte & Doolittle, 1982) was predicted using ProtParam.

5.3.1.4 Bovine TB vaccine construction using structural vaccinology

After selecting the immunodominant B-cell and T-cell epitopes from immunoinformatic analysis, structural vaccinology was implemented to design a structural model for the bovine TB vaccine and analyse its interaction with MHC molecules and the toll-like receptors present in the host (cattle).

5.3.1.4.1 Vaccine design

To design a final bovine TB vaccine sequence from the shortlisted epitopes, we joined the HTL, CTL and B-cell epitopes with the help of flexible linker, AAY. An adjuvant 50s ribosomal protein L7/L12 was attached using an EAAAK linker at the N-terminal of the bovine TB vaccine protein sequence.

5.3.1.4.2 Analysis of antigenicity and physiochemical properties

VaxiJen (Doytchinova & Flower, 2007) was used to predict the antigenicity of the final vaccine construct. Algpred, an SVM model, was used for determining the allergenicity of the vaccine protein (Saha & Raghava, 2006a). ProtParam was used for computing the various physiochemical properties of the final bovine TB vaccine. These properties include molecular weight (MW), an isoelectric point of the protein (pI), amino acid composition, extinction coefficient (Gill & von Hippel, 1989), instability index (II), estimated half-life (Bachmair et al., 1986), aliphatic index (Ikai, 1980) and the grand average of hydropathicity (GRAVY) (Kyte & Doolittle, 1982)

5.3.1.4.3 Vaccine structural model construction

Prediction of the secondary structure of the vaccine construct consisting of alpha-helices, beta-sheet and coil (Buchan & Jones, 2019) PSIPRED (Buchan & Jones, 2019). The tertiary structure of the final bovine TB vaccine construct was generated using RaptorX (Wang et al., 2016). First, RaptorX performed a template search based on the similarity of the input sequence and then constructed a good quality structural model. The output structural model of the vaccine, given by RaptorX, was further refined using GalaxyRefine to minimise any distortions present in the structure (Ko et al., 2012). In our study, we used PROCHECK and Verify 3D for structure validation. PROCHECK is a suite of programs used for checking the stereochemical quality of the protein modelled. It analyses the overall structural geometry and residues using residue geometry (Laskowski et al., 1993). Verify 3D works by determining the compatibility of the 3D atomic model with its amino acid sequences by assigning a structural class based on location and environment (alpha-helices, beta-sheets, loops) (Colovos & Yeates, 1993). If the vaccine structural model had more than 90% residues in the most favoured region, it was considered the best quality model and was used for further analysis.

5.3.1.5 Vaccine-TLR docking and dynamics

Bovine toll-like receptors are present on the surface of antigen-presenting cells (macrophages or dendritic cells). TLRs interact with the pathogen/vaccine antigen for initiating the innate immune response. The structures of bovine TLRs were not present in the PDB database. The protein sequence of bovine used for structure construction was Q95LA9, Q9GL65 and B5T278 for TLR-2, TLR-4 and TLR-6. I-TASSER server was used for generating the 3D structure of TLRs. I-TASSER is an online platform for

automated protein structure predictions from amino acid sequences. The I-TASSER suite pipeline consists of four essential steps (Yang et al., 2015):

- Threading template prediction: after submission of the query sequence, template proteins of similar folds are retrieved from the PDB library by LOMETS (**lo**cal **me**ta **t**hreading **s**erver)
- Iterative Monte-Carlo simulation for structural assembly: continuous fragments expunged from the PDB templates are assembled into full-length models using Monte Carlo simulations and loops are constructed using *ab initio* structure modelling.
- Selection of model and refinement: simulation is performed to remove steric clashes between atoms of the amino acid residues. The global topology of the constructed model is improved and the final structures built by optimising the hydrogen bonds.
- Function annotation: function is inferred by comparing the structural models with known proteins.

The final output of I-TASSER includes top ten structural models for bovine TLR along with top ten alignments and top ten PDB structures that are used in model building and are the closest to the modelled structure.

For performing molecular docking analysis, the PatchDock (Schneidman-Duhovny et al., 2005) and FireDock (Mashiach et al., 2008) servers were used. The PatchDock server computes the docking transformations of the receptors and ligands based on their molecular shape complementarity. The results of PatchDock were then presented to FireDock for further refinement of the docking results and predicting the global binding energy of receptor (TLR) and ligand (bovine vaccine).

5.3.2   Drug target identification of *Mycobacterium bovis* using subtractive proteomic analysis
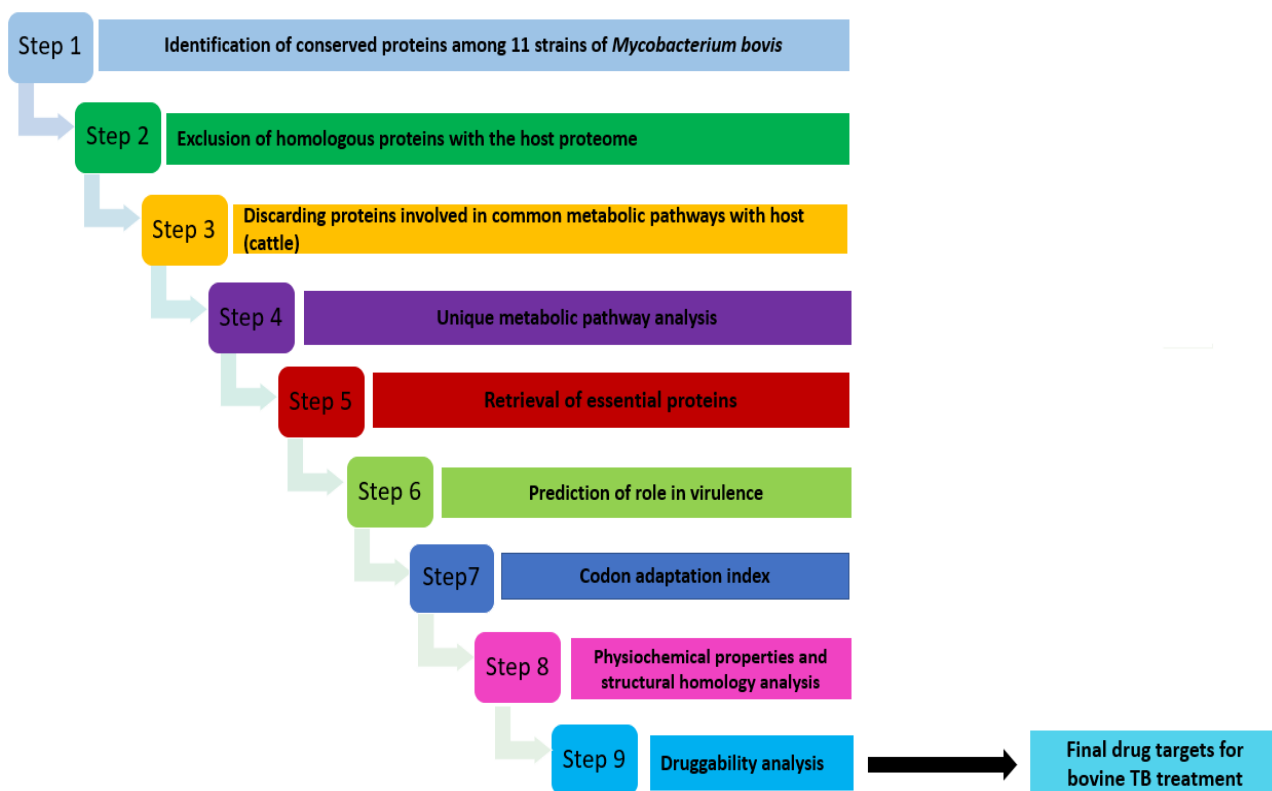
Figure 5. 1: Comprehensive subtractive proteomic analysis for the identification of potential drug targets in *Mycobacterium bovis*

5.3.2.1 Exclusion of proteins homologous with the host proteome

The conserved proteins identified by subtractive reverse vaccinology analysis, present among 11 strains of *Mycobacterium bovis*, were first tested for homology with the host proteome. The drugs targeting homologous proteins may cause side effects or hypersensitive reactions in the host. To identify homologous proteins between *Mycobacterium bovis* and the host (cattle), BLASTp was used (Boratyn et al., 2013). BLASTp was performed using a non-redundant protein sequence database with a threshold E-value of e10-4. *Bos* (taxid:9903), *Bos tauurus* (taxid:9913) and *Bos indicus* (taxid:9915) were used as reference organisms for performing sequence similarity. BLASTp performs local sequence alignment and measures sequence similarities based on maximum segment pairs (MSP) (Altschul et al., 1990). The results of BLASTp consist of homologous and non-homologous sequences. The proteins with sequence identity and a bit score of more than 30% and 100, respectively, were excluded from the study.

5.3.2.2 Eliminating proteins involved in common metabolic pathways with the host

Comparative metabolic pathway analysis was carried out to identify proteins involved in metabolic pathways common in cattle and *Mycobacterium bovis.* The Kyoto Encyclopaedia of Gene and Genome (KEGG) (Kanehisa et al., 2017) was used for performing this step. A manual comparison was made and the proteins involved in the common pathways were removed.

5.3.2.3 Unique metabolic pathway analysis

For the treatment of infectious disease, a drug target should be uniquely present in the pathogen. The selection of proteins involved in the specific pathways of *Mycobacterium bovis* was made using KEGG automatic annotation server (KAAS) (Moriya et al., 2007). BLAST was selected for performing the search in KAAS. *Mycobacterium bovis* AF2122/97 was chosen as a reference from the organism list. A manual search was carried out to choose proteins unique to the *Mycobacterium bovis* AF2122/97 strain, while the remaining proteins were excluded from the study.

5.3.2.4 Retrieval of the essential proteins of *Mycobacterium bovis*

In this step, we identified proteins that played a vital role in biological processes and functions, such as physiology, metabolism and developmental processes within *Mycobacterium bovis*. These proteins help in the regular functioning and survival of the disease-causing *Mycobacterium bovis*. For the identification of essential proteins, we used the Database of Essential Genes (DEG). The database contains information about essential genes and their proteins obtained from different experimental methods (Zhang, 2004). For a query sequence, a similarity search was performed against the essential proteins present in DEG with the help of BLASTp. The parameters for the search were set as:

- e-value 10-50
- sequence identity more than 30%
- bit score >100
- query coverage 100%

5.3.2.5 Prediction of the role of proteins in virulence

The proteome of *Mycobacterium bovis* consists of proteins involved in virulence, the progression of bovine TB and escaping the host's immune response (Garnier et al., 2003). Virulent proteins play a crucial role in infectious pathways, such as adherence, colonisation, invasion, and help in escaping from the host's immune response. Drugs targeting the virulent pathway of the pathogen would surely help in eliminating the disease from the host body. Thus, there is a need to identify proteins that play a major role in causing the infection in the host cell. We used two tools for predicting virulent proteins: VirulentPred that is an SVM-based method (Garg & Gupta, 2008) and MP3 tool that uses SVM and

HMM (Gupta et al., 2014). Proteins predicted virulent by both methods were selected for further analysis.

5.3.2.6 Codon adaptation index

The codon adaptation index (CAI) is a measure of synonymous codon usage bias (Puigbò et al., 2008). CAI predicts the translational efficiency and level of expression of a gene. The EMBOSS server was used to calculate the codon adaptation index for predicting the expression level of the selected gene targets of *Mycobacterium bovis* inside the host body. Proteins that are highly expressed are considered potential drug targets. To calculate CAI, the nucleotide sequences of the selected proteins were obtained from the NCBI Gene database. The gene sequences were then submitted to EMBOSS server for predicting the CAI of each gene. The genes with a CAI score of more than 0.7 were selected for further analysis.

5.3.2.7 Analysis of physiochemical properties of selected drug targets

ProtParam was used for computing various physiochemical properties of the selected drug targets in *Mycobacterium bovis*. The properties include molecular weight (MW), an isoelectric point of a protein (pI), amino acid composition, extinction coefficient (Gill & von Hippel, 1989), instability index (II), estimated half-life (Bachmair et al., 1986), aliphatic index (Ikai, 1980) and the grand average of hydropathicity (GRAVY) (Kyte & Doolittle, 1982). A potential drug target protein should have a molecular weight lower than 100 kDa. Isoelectric point is the pH of the solution at which the net electrical charge of the protein becomes zero (Pergande & Cologna, 2017). The instability index determines the stable nature of the protein in a laboratory test tube. A protein with an instability index of lower than 40 is considered as potential drug target. The extinction coefficient predicts the absorption of light at a particular wavelength by the target protein and helps in determining the protein concentration. The GRAVY score determines the hydrophilic or hydrophobic nature of a protein.

5.3.2.8 Structural homology analysis

The selected proteins were then analysed for the availability of their crystallographic X-ray structure in the PDB database. For this, BLASTp was performed using a PDB database with a threshold E-value of e10-6. The results of BLASTp consist of homologous and non-homologous sequences. The proteins with sequence identity and a bit score of more than 30% and 100, respectively, were selected in the study. The proteins with no template structure available for generating the 3D structural model were eliminated.

5.3.2.9 Prediction of druggability potency of the selected proteins

Druggability is defined as the ability of a protein to bind to a drug molecule. The druggability of the selected proteins was predicted by performing a sequence similarity search in the DrugBank database. DrugBank is a trustworthy database providing information on the available drugs and their drug targets (Knox et al., 2011). DrugBank currently contains information on 9591 drugs with 2037 FDA approved drugs (Food and Drug Administration). Over 6000 experimental drug entries are present in DrugBank. The e-value for the similarity search was set to 10-20. The drug target proteins for whom homologous proteins are present in the DrugBank database are druggable proteins. The protein with no hits found against any protein sequence is termed a novel drug target.

5.4   Results

5.4.1   Developing an epitope-based bovine TB vaccine

Bovine TB is a chronic infectious disease caused by the etiological agent, *Mycobacterium bovis*. Cattle are considered the main reservoir of bovine TB compared to other domestic livestock. Isolating the infected animal and slaughtering are performed for reducing transmission among other animals in the herd. The use of BCG has not provided a sufficient level of protection. Several attempts have been made to develop a live-attenuated or heat-killed vaccine against bovine TB in cattle, (Buddle, 2010; Buddle et al., 2018; Palmer & Thacker, 2018; Parlane & Buddle, 2015)r, 2018; Parlane & Buddle, 2015). As described in the previous section, in this study, we used the conceptual framework developed in chapter-4 for designing an epitope-based bovine TB vaccine containing B-cell and T-cell (HTL and CTL) epitopes that could elicit a humoral and cell-mediated immune response. This section describes the results of the holistic framework tested for bovine TB using several bioinformatics tools.

5.4.1.1 Identification of nine conserved antigens through the subtractive reverse vaccinology approach

Developing a vaccine providing a broad spectrum of protection against many *Mycobacterium bovis* strains and also combating drug resistance required the identification of conserved proteins within different strains of bovine TB bacteria. In this study, we used *Mycobacterium bovis* AF2122/97 as a reference strain with 3988 proteins present in its proteome. First, we downloaded complete proteome sequences of 11 strains of *Mycobacterium bovis* that have been completely sequenced using the NCBI Genome FTP site. A standalone BLAST was then used for a proteome comparison of 11 strains and the results were stored in Excel files. Manual comparison of the results revealed that 1163 proteins were conserved with more than 99% sequence similarity and 100% query coverage among the 11 strains of *Mycobacterium bovis.*

To develop an effective bovine TB vaccine, we used a subtractive reverse vaccinology approach to identify conserved, surface-exposed antigenic proteins with critical functional characteristics, such as signal peptides and membrane-spanning regions, non-homologous, lipoprotein signatures, adhesion probability and non-toxic from the proteome of *Mycobacterium bovis* AF2122/97. First, we predicted the localization of each conserved protein and selected proteins present in the cell membrane, cell walls and extracellular spaces of *Mycobacterium bovis*. As we know, the proteins present in the cell membrane, cell wall, and extracellular space are involved in membrane integrity and permeability, efflux mechanisms, and active transport of molecules; thus, they were considered suitable candidates for vaccine development. For subcellular localization prediction, we used six bioinformatics tools PSORTb v.3.0, CELLO, LocTree3, SOSUI, pLoc_bal-mGpos and GramLocEN. Of the 1163 conserved proteins, 273 proteins, commonly predicted by more than four SCL tools, were selected for further analysis. In the next step, proteins having more than one transmembrane α-helix were excluded from the study, as they were challenging to clone and purify in the laboratory. Then, using the TMHMM server, we shortlisted 142 proteins with 0 or 1 transmembrane α-helix.

The presence of signal peptides and a lipoprotein signature within a protein makes it a vital antigenic and immunogenic vaccine target. The Sec (secretory) and TAT signal peptides are ubiquitous protein-sorting signals that help in the translocation of a protein across the cell membrane in *Mycobacterium bovis*. For the prediction of secretory signal peptides, the SecretomeP and SignalP 4.1 servers were used. A total of 43 conserved and surface-exposed proteins showed the presence of Sec signal peptides. Next, PRED-TAT was used to identify tat signal peptides and 12 proteins were found to be involved in the TAT pathway. Thus, out of 142 proteins, 55 proteins (Sec and TAT proteins) were engaged in the secretory pathway of *Mycobacterium bovis*. Lipoproteins, which are essential sets of membrane proteins performing important functions in *Mycobacterium bovis,* were predicted using PRED-LIPO. Twenty-eight lipoproteins were identified from 142 proteins. Thus, we selected a total of 83 proteins (55 secretory proteins and 28 lipoproteins) for further analysis.

A protein that can activate an immune response against pathogenic bacteria without causing considerable side effects inside the host body is considered a potential vaccine target. The VaxiJen, VirulentPred and MP3 servers were used for determining the antigenicity of 83 proteins. Concordance analysis of the three antigen servers identified 21 highly antigenic proteins that can initiate a strong immune response with or without side effects. Thus, we identified the allergenic nature of the selected antigens. Algpred was used to predict the allergenic or non-allergenic proteins. From 21 antigenic proteins, 18 proteins were found to be non-allergenic. These 18 conserved, surface-exposed, antigenic and non-allergenic proteins were further analysed for adhesion probability. Adhesin proteins help the attachment of *Mycobacterium bovis* to the host's cell-surface receptors. Hence, we used SPAAN to

identify adhesin proteins among the shortlisted proteins. Of the 18 proteins, only nine proteins had a strong adhesion probability score of a 0.5 or above.

The last step of subtractive reverse vaccinology involves a study of the homology between the host (cattle) and *Mycobacterium bovis.* Homologous proteins can initiate hypersensitive reactions inside the host body, causing severe health problems or side effects. For homology analysis, BLASTp was used for determining sequence similarity among host and pathogen. We used *Bos tauurus* (taxid:9913) and *Bos indicus* (taxid:9915) as reference organisms for performing sequence similarity. We found no similarity among the nine proteins of *Mycobacterium bovis* and cattle. Finally, a total of nine conserved, membrane-spanning, antigenic and non-allergic proteins were selected with the help of a reverse vaccinology approach from the proteome of *Mycobacterium bovis,* as shown in Table 5.1.

Table 5. 1: List of nine conserved, surface-exposed, antigenic and non-allergic proteins identified by reverse vaccinology approach. Column 1- accession number of protein, column 2- the function of protein, columns 3-10 show the results of the reverse vaccinology process for the shortlisted proteins

Accession number	Function	Prediction of signal peptide	TMHMM	MP3	VirulentPred	VaxiJen	VaxiJen Score	AlgPred	Adhesion probability
YP_009357464.1	SECRETED ANTIGEN 85-C FBPC (85C) (ANTIGEN 85 COMPLEX C) (AG58C) (MYCOLYL TRANSFERASE 85C) (FIBRONECTIN-BINDING PROTEIN C)	Tat signal peptide predicted	Pred Helix=1	Pathogenic	Virulent	Probable Antigen	0.4402	Non-allergen	0.812
YP_009358272.1	ppe family protein ppe14	Sec signal peptide predicted	Pred Helix=0	Pathogenic	Virulent	Probable Antigen	0.6802	Non-allergen	0.592
YP_009359011.1	pe family protein pe17	Sec signal peptide predicted	Pred Helix=0	Pathogenic	Virulent	Probable Antigen	0.4152	Non-allergen	0.78
YP_009359070.1	ppe family protein ppe23	Sec signal peptide predicted	Pred Helix=0	Pathogenic	Virulent	Probable Antigen	0.4637	Non-allergen	0.752
YP_009361194.1	SECRETED MPT51/MPB51 ANTIGEN PROTEIN FBPD (MPT51/MPB51 ANTIGEN 85 COMPLEX C) (AG58C) (MYCOLYL TRANSFERASE 85C) (FIBRONECTIN-BINDING PROTEIN C) (85C)	Sec signal peptide predicted	Pred Helix=1	Pathogenic	Virulent	Probable Antigen	0.4515	Non-allergen	0.747
YP_009361195.1	SECRETED ANTIGEN 85-A FBPA (MYCOLYL TRANSFERASE 85A) (FIBRONECTIN-BINDING PROTEIN A) (ANTIGEN 85 COMPLEX A)	Tat signal peptide predicted	Pred Helix=1	Pathogenic	Virulent	Probable Antigen	0.5259	Non-allergen	0.77

YP_009358321.1	POSSIBLE LIPOPROTEIN LPRP	Lipoprotein	Pred Helix=0	Pathogenic	Virulent	Probable Antigen	0.5052	Non-allergen	0.598
YP_009358349.1	POSSIBLE CONSERVED EXPORTED PROTEIN	Sec signal peptide predicted	Pred Helix=1	Pathogenic	Virulent	Probable Antigen	0.5621	Non-allergen	0.757
YP_009361095.1	CONSERVED HYPOTHETICAL PROTEIN	Sec signal peptide predicted	Pred Helix=0	Pathogenic	Virulent	Probable Antigen	0.4189	Non-allergen	0.619

5.4.1.2 Prediction of B-cell and T-cell epitopes from the nine conserved antigens using an immunoinformatics approach

An ideal prophylactic bovine TB vaccine should stimulate a robust immune response by the generation of memory cells to eliminate an infection in the future. For this, there is a need to identify immunodominant B-cell and T-cell epitopes that would initiate humoral and cellular immunity inside the host. Hence, we performed immunoinformatic analysis for screening the effective B-cell and T-cell epitopes required for designing an epitope-based bovine TB vaccine.

5.4.1.2.1 Prediction of B-cell epitopes

B-cell epitopes help in generating a humoral or antibody-mediated immune response in the host's body. We used ABCpred for identifying B-cell epitopes from the nine antigenic proteins of *Mycobacterium bovis*. The length of the B-cell epitope was set to 20 amino acid residues. A total of 251 B-cell epitopes were found from the nine antigenic proteins. Table 5.2 provides the results of immunoinformatics analysis performed on these nine bovine TB antigens. The second column of Table 5.2 shows the B-cell epitopes predicted for each bovine TB antigenic protein.

Table 5. 2: Results of immunoinformatics analysis for predicting B-cell epitopes. Columns 2, 3, 4 and 5 show the results for each step performed for filtering the immunodominant B-cell epitopes from the nine *Mycobacterium bovis* antigenic proteins

Accession number	B-cell epitopes predicted	Antigenic & non-toxic	Non-allergenic	Hydrophilicity
YP_009357464.1	31	11	9	3
YP_009358272.1	34	12	10	2
YP_009359011.1	25	4	2	0
YP_009359070.1	29	5	1	0
YP_009361194.1	23	8	8	1
YP_009361195.1	31	13	12	4
YP_009358321.1	16	3	3	3
YP_009358349.1	34	15	11	1
YP_009361095.1	28	9	4	2

First, the immunogenic nature of B-cell epitopes was identified using the VaxiJen server and ToxinPred. For investigating safe and immunodominant epitopes, a filtering process was performed on the 251 B-cell epitopes. Of the 251 epitopes, 80 B-cell epitopes were found to be antigenic and non-toxic. Next, the identification of non-allergenic epitopes was performed with the help of the ALLERTOP 2.0 server. A total of 60 non-allergenic epitopes were shortlisted for further analysis of their hydrophilicity. The hydrophilic residues present on the surface of antigenic proteins facilitate quick interactions with the host's immune cells. The GRAVY score was calculated using the ProtParam server. In the ProtParam analysis, we discovered 16 B-cell epitopes with negative GRAVY scores. No potential B-cell epitopes were found in the YP_009359011.1 and YP_009359070.1 proteins. After finishing the immunoinformatics analysis, a total of 16 B-cell epitopes were selected for the construction of the bovine TB vaccine (Table 5.3).

Table 5. 3: Shortlist of 16 B-cell epitopes from *Mycobacterium bovis* antigens required for constructing an epitope-based vaccine for bovine tuberculosis

Start	End	Sequence	VaxiJen	VaxiJen score	ALLERTOP 2.0	ToxinPred	GRAVY
1. YP_009357464.1							
83	103	LDGLRAQDDYNGWDINTPAF	Antigen	1.0794	Non-allergenic	Non-toxin	-0.79
275	295	GLTLRTNQTFRDTYAADGGR	Antigen	1.132	Non-allergenic	Non-toxin	-0.94
287	307	TYAADGGRNGVFNFPPNGTH	Antigen	0.7017	Non-allergenic	Non-toxin	-0.79
2. YP_009358272.1							
285	305	GNAGGLGPTQGHPLSSATDE	Antigen	1.064	Non-allergenic	Non-toxicity	-0.71
381	401	LAARGTTGGGGTRSGTSTDG	Antigen	3.0539	Non-allergenic	Non-toxicity	-0.65
5. YP_009361194.1							
131	151	NRGLAPGGHAAVGAAQGGYG	Antigen	0.8762	Non-allergenic	Non-toxin	-0.17
6. YP_009361195.1							
64	84	DIKVQFQSGGANSPALYLLD	Antigen	1.3288	Non-allergenic	Non-toxin	-0.01
114	134	PVGGQSSFYSDWYQPACGKA	Antigen	0.7002	Non-allergenic	Non-toxin	-0.58
281	301	IKFQDAYNAGGGHNGVFDFP	Antigen	1.3875	Non-allergenic	Non-toxin	-0.42
311	331	GAQLNAMKPDLQRALGATPN	Antigen	0.7957	Non-allergenic	Non-toxin	-0.505

7. YP_009358321.1							
Start	End	Sequence	VaxiJen	VaxiJen score	ALLERTO P 2.0	ToxinPre d	GRAVY
156	176	SQIRDDGVTINLVNGDNRGP	Antigen	0.8999	Non-allergenic	Non-toxin	-0.83
181	201	NTGCHLPAAWRTAPPPLNMR	Antigen	0.7017	Non-allergenic	Non-toxin	-0.545
194	214	PPPLNMRPANDPDVHYPYLY	Antigen	1.3269	Non-allergenic	Non-toxin	-0.985
8. YP_009358349.1							
Start	End	Sequence	VaxiJen	VaxiJen score	ALLERTO P 2.0	ToxinPre d	GRAVY
316	336	DQSGATYPIAWEIEIERIGL	Antigen	1.2462	Non-allergenic	Non-toxin	-0.135
9. YP_009361095.1							
Start	End	Sequence	VaxiJen	VaxiJen score	ALLERTO P 2.0	ToxinPre d	GRAVY
223	243	GEMSIRQIDGQTVLSYFNAS	Antigen	0.7201	Non-allergenic	Non-toxin	-0.185
305	325	IFVSQWDTRARQNGPYRVIQ	Antigen	0.7626	Non-allergenic	Non-toxin	-0.735

5.4.1.2.2 Prediction of MHC-I restricted (CTL) T-cell epitopes

The antigens of *Mycobacterium bovis* are engulfed by the antigen-presenting cells and then fragmented into smaller antigenic peptides called epitopes. These epitopes are later presented to T-cell receptors (TCR) present on the surface of T-cells through cell-surface attached MHC molecules. The epitopes binding to MHC-I molecules are termed cytotoxic T-cell (CTL) epitopes, or MHC-I restricted T-cell epitopes and play an essential role in cellular immunity. In our study, the length of CTL epitopes was set to nine amino acid residues long. For predicting MHC-I restricted T-cell epitopes, we used the following bovine MHC-I HLA alleles: BoLA-D18.4, BoLA-AW10, BoLA-JSP.1, BoLA-HD6, BoLA-T2a, BoLA-T2b and BoLA-T2c.

IEDB MHC-I and the NetMHCpan 4.1 server were used for predicting CTL epitopes. We identified a total of 140 CTL epitopes from nine *Mycobacterium bovis* antigens. Of the 140 epitopes, 78 CTL epitopes were found to be commonly predicted by both servers. Table 5.4 shows the total predicted CTL epitopes and those common to both servers (first and second columns) and epitopes filtered in subsequent steps for each *Mycobacterium bovis* membrane-localized antigen.

Table 5. 4: Results of immunoinformatics analysis for predicting CTL epitopes. Columns 2, 3, 4, 5 and 6 show the results for each step performed for filtering the immunodominant CTL epitopes from the nine *Mycobacterium bovis* antigenic proteins

Accession number	Total CTL epitopes predicted	CTL epitopes (commonly predicted by IEDB MHC-I and NetMHCpan 4.1 server)	Antigenic & non-toxic	Non-allergenic	Hydrophilicity
YP_009357464.1	15	9	2	0	0
YP_009358272.1	23	16	5	5	1
YP_009359011.1	13	5	1	0	0
YP_009359070.1	24	11	4	4	1
YP_009361194.1	15	9	4	4	2
YP_009361195.1	13	9	1	1	0
YP_009358321.1	7	5	2	2	1
YP_009358349.1	19	8	1	1	0
YP_009361095.1	11	6	4	4	3

To validate the CTL epitopes' ability to initiate a cellular immune response, antigenicity and toxicity analysis was completed using VaxiJen and Toxin Pred. The antigenicity and toxicity scores revealed that out of 78 CTL epitopes, only 24 epitopes were highly immunogenic and non-toxic. These 24 epitopes were then considered for allergenicity prediction using ALLERTOP 2.0. The three CTL epitopes found to cause allergenic reactions inside the host body were excluded from the study. A hydrophilic score calculation was then undertaken using the ProtParam server. In our study, we excluded the epitopes having positive GRAVY scores. Thus, we had eight CTL epitopes, hydrophilic in nature, and with a negative GRAVY score. YP_009357464.1, YP_009359011.1, YP_009361195.1 and YP_009358349.1 antigens had no potential immunodominant CTL epitopes. Finally, these eight CTL epitopes were chosen for constructing an epitope-based TB vaccine (Table 5.5).

Table 5. 5: The eight shortlisted CTL epitopes from *Mycobacterium bovis* antigens required for constructing an epitope-based vaccine for bovine tuberculosis

Start	End	Length	Sequence	VaxiJen	VaxiJen score	ToxinPred	ALLERTOP 2.0	GRAVY
2. YP_009358272.1								
Start	End	Length	Sequence	VaxiJen	VaxiJen score	ToxinPred	ALLERTOP 2.0	GRAVY
14	22	9	RMYSGPGPE	Antigenic	1.1523	Non-toxic	Non-allergenic	-1.356
4. YP_009359070.1								
Start	End	Length	Sequence	VaxiJen	VaxiJen score	ToxinPred	ALLERTOP 2.0	GRAVY
50	58	9	RMYSGPGSG	Antigenic	1.582	Non-toxic	Non-allergenic	-0.922
5. YP_009361194.1								
Start	End	Length	Sequence	VaxiJen	VaxiJen score	ToxinPred	ALLERTOP 2.0	GRAVY
82	90	9	NAMNTLAGK	Antigenic	0.7311	Non-toxic	Non-allergenic	-0.3
258	266	9	NQYRSVGGH	Antigenic	0.9887	Non-toxic	Non-allergenic	-1.489
7. YP_009358321.1								
Start	End	Length	Sequence	VaxiJen	VaxiJen score	ToxinPred	ALLERTOP 2.0	GRAVY
68	76	9	AMIAKYSPQ	Antigenic	0.7577	Non-toxic	Non-allergenic	-0.122
9. YP_009361095.1								
Start	End	Length	Sequence	VaxiJen	VaxiJen score	ToxinPred	ALLERTOP 2.0	GRAVY
308	316	9	SQWDTRARQ	Antigenic	1.2331	Non-toxic	Non-allergenic	-2.233
228	236	9	RQIDGQTVL	Antigenic	0.9928	Non-toxic	Non-allergenic	-0.4
145	153	9	YQDGRQTQI	Antigenic	1.826	Non-toxic	Non-allergenic	-1.822

5.4.1.2.3 Prediction of MHC-II restricted (HTL) T-cell epitopes

MHC class-II molecules present on the surface of the macrophage and dendritic cells present antigens/epitopes to the naïve T-helper cells. The epitopes presented by MHC-II molecules are called helper T-cell epitopes (HTL epitopes) or MHC-II restricted T-cell epitopes. The interaction of epitopes with helper T-cells leads to the activation of cellular immunity with a further release of cytokines to eliminate the infection. The IEDB MHC II server was used to predict 15-mer HTL epitopes of *Mycobacterium bovis* using the consensus prediction method. HLA alleles of the mouse model were used in this step as no information was available for bovine MHC-II molecules. The lower percentile rank indicates a higher affinity of an epitope towards the MHC molecule. Thus, the epitope with

percentile ranks lower than 1.0 were chosen. A total of 145 HTL epitopes were predicted from the nine antigens of *Mycobacterium bovis* using the IEDB MHC II server.

Table 5. 6: Results of immunoinformatics analysis for predicting HTL epitopes. Columns 2, 3, 4 and 5 show the results for each step performed for filtering the immunodominant HTL epitopes from the nine antigenic proteins of *Mycobacterium bovis*

Accession number	HTL epitopes predicted	Antigenic & non-toxic	Non- allergenic	Hydrophilicity
YP_009357464.1	20	5	1	0
YP_009358272.1	36	19	12	0
YP_009359011.1	21	17	8	0
YP_009359070.1	31	20	15	0
YP_009361194.1	12	9	7	0
YP_009361195.1	12	6	5	1
YP_009358321.1	4	2	2	0
YP_009358349.1	6	5	2	1
YP_009361095.1	3	0	0	0

Table 5.6 shows the results of immunoinformatics analysis on the nine antigens of *Mycobacterium bovis*. Columns 3, 4 and 5 of Table 5.6 show the outcome of the antigenicity, toxicity, allergenicity and hydrophilicity analysis, respectively. A total of 83 immunogenic and non-toxic HTL epitopes were discovered from 145 epitopes with the help of VaxiJen and ToxinPred. The allergenicity of 83 HTL epitopes were then predicted using the ALLERTOP 2.0 server. Fifty-two non-allergenic HTL epitopes were selected for further analysis. Finally, we had only two HTL epitopes after calculating the hydrophilicity of the selected 52 non-allergenic HTL epitopes (Table 5.7).

Table 5. 7: Two shortlisted HTL epitopes from *Mycobacterium bovis* antigens required for constructing an epitope-based vaccine for bovine tuberculosis

6. YP_009361195.1								
Start	End	Length	Sequence	VaxiJen	VaxiJen score	ToxinPred	ALLERTOP 2.0	GRAVY
66	80	15	KVQFQSGGANSPALY	Antigen	1.1385	Non-toxic	Non-allergenic	-0.353
8. YP_009358349.1								
Start	End	Length	Sequence	VaxiJen	VaxiJen score	ToxinPred	ALLERTOP 2.0	GRAVY
16	30	15	TLVAFWWWQRPRTNA	Antigen	0.8592	Non-toxic	Non-allergenic	-0.487

5.4.1.3 Designing the vaccine sequence and an assessment of properties

After selecting highly immunogenic and excluding cross-reactive epitopes, we had 26 epitopes (HTL epitopes-2, CTL epitopes-8 and B-cell epitopes-16) from the nine antigenic proteins of *Mycobacterium bovis* (Table 5.8). The selected epitopes would evoke a potent cellular and humoral immune response inside the host body.

Table 5. 8: Final shortlisted 26 epitopes for bovine TB vaccine construction

HTL epitopes	
1	KVQFQSGGANSPALY
2	TLVAFWWWQRPRTNA
CTL epitopes	
3	RMYSGPGPE
4	RMYSGPGSG
5	NAMNTLAGK
6	NQYRSVGGH
7	AMIAKYSPQ
8	SQWDTRARQ
9	RQIDGQTVL
10	YQDGRQTQI
B-cell epitopes	
11	TYAADGGRNGVFNFPPNGTH
12	GLTLRTNQTFRDTYAADGGR

13	LDGLRAQDDYNGWDINTPAF
14	LAARGTTGGGGTRSGTSTDG
15	GNAGGLGPTQGHPLSSATDE
16	NRGLAPGGHAAVGAAQGGYG
17	GAQLNAMKPDLQRALGATPN
18	IKFQDAYNAGGGHNGVFDFP
19	DIKVQFQSGGANSPALYLLD
20	PVGGQSSFYSDWYQPACGKA
21	SQIRDDGVTINLVNGDNRGP
22	PPPLNMRPANDPDVHYPYLY
23	NTGCHLPAAWRTAPPPLNMR
24	DQSGATYPIAWEIEIERIGL
25	IFVSQWDTRARQNGPYRVIQ
26	GEMSIRQIDGQTVLSYFNAS

The sequence of the bovine TB vaccine was designed using a simple strategy. First, the adjuvant 50s ribosomal protein L7/L12 at the N-terminal was added to the vaccine protein sequence with the help of the EAAAK linker (Figure 5.10 (ii)). We then joined the HTL, CTL and B-cell epitopes using flexible linker AAY. With one adjuvant, one EAAAY linker, 25 AAY linkers and 26 epitopes (HTL epitopes-2, CTL epitopes-8 and B-cell epitopes-16), the final length of the bovine TB vaccine sequence was 632 amino acid residues.
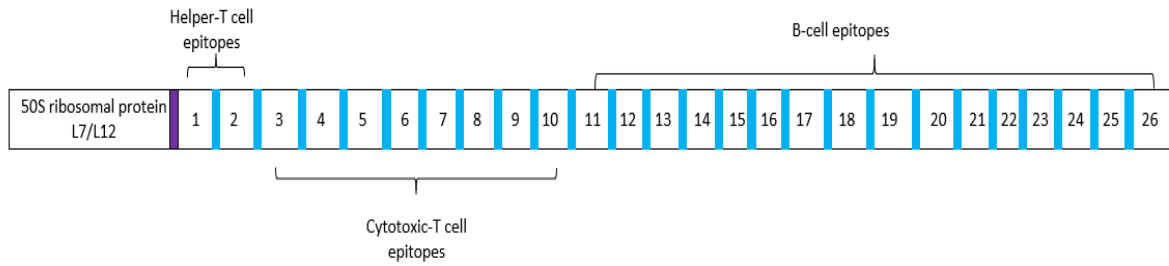
(i)

MAKLSTDELLDAFKEMTLLELSDFVKKFEETFEVTAAAPVAVAAAGAAPAGAAVEAAEEQSEFDVILEAAGDKKIGV
IKVVREIVSGLGLKEAKDLVDGAPKPLLEKVAKEAADEAKAKLEAAGATVTVKEAAAKKVQFQSGGANSPALYAAYT
LVAFWWWQRPRTNAAAYRMYSGPGPEAAYRMYSGPGSGAAYNAMNTLAGKAAYNQYRSVGGHAAYAMIAK
YSPQAAYSQWDTRARQAAYRQIDGQTVLAAYYQDGRQTQIAAYTYAADGGRNGVFNFPPNGTHAAYGLTLRTN
QTFRDTYAADGGRAAYLDGLRAQDDYNGWDINTPAFAAYLAARGTTGGGGTRSGTSTDGAAYGNAGGLGPTQ
GHPLSSATDEAAYNRGLAPGGHAAVGAAQGGYGAAYGAQLNAMKPDLQRALGATPNAAYIKFQDAYNAGGGH
NGVFDFPAAYDIKVQFQSGGANSPALYLLDAAYPVGGQSSFYSDWYQPACGKAAAYSQIRDDGVTINLVNGDNR
GPAAYPPPLNMRPANDPDVHYPYLYAAYNTGCHLPAAWRTAPPPLNMRAAYDQSGATYPIAWEIEIERIGLAAYI
FVSQWDTRARQNGPYRVIQAAYGEMSIRQIDGQTVLSYFNAS

(ii)

Figure 5. 2: Epitope-based vaccine sequence construction scheme. (i) Schematic representation of bovine TB vaccine construct consisting of B-cell and T-cell epitopes joined together by flexible linkers and an adjuvant at N-terminal of the vaccine protein, (ii) TB vaccine protein sequence. The adjuvant sequence is highlighted in green with flexible linkers in purple (EAAAK) and blue (AAY). HTL epitopes are highlighted in orange, with CTL epitopes in black and the B-cell epitopes in brown

After designing the bovine TB vaccine sequence, its antigenicity, allergenicity and physiochemical properties were evaluated. The VaxiJen server was used for predicting the antigenicity of the vaccine sequence. A high antigenic score of 0.82 showed the great immunogenic potential of the epitope-based bovine TB vaccine. The AlgPred server indicated that the designed vaccine was non-allergenic to the host (cattle).

Figure 5. 3: Physiochemical properties of final bovine TB vaccine construct

ProtParam was used for predicting the physiochemical properties of the epitope-based bovine TB vaccine sequence designed. The vaccine sequence comprised 632 amino acid residues with a molecular weight of 66.85 kDa. The bovine TB vaccine construct comprised 9071 atoms and its molecular formula was $C_{2981}H_{4509}N_{823}O_{912}S_{12}$. The theoretical pI was assessed to be 5.77. The estimated half-life was found to be 30 hours in mammalian reticulocytes (*in-vitro*). The instability index with a value of 32.35 represented the stable nature of the TB vaccine. The aliphatic index predicted the relative volume occupied by aliphatic side chains in a protein. A value for the aliphatic index greater than 30 indicates that the protein is thermodynamically stable. The higher the value of the aliphatic index (68.56 in this study), the higher the thermos-stability of the protein. The GRAVY score of the vaccine was found to be -0.302. The negative value of the estimated GRAVY score indicated the hydrophilic nature of the vaccine. After evaluating all the properties, we discovered that the epitope-based bovine TB vaccine we designed had all the attributes of a promising prophylactic vaccine candidate required for initiating a robust immune response inside the host.

5.4.1.4 Constructing the structural model of the epitope-based bovine TB vaccine

The secondary structure of the bovine TB vaccine was evaluated using PSIPRED. According to PSIPRED, 239 amino acid residues were involved in forming alpha-helices, constituting 37.82% of the overall bovine TB vaccine sequence. Only 99 amino acid residues participated in the formation of beta-strand (15.66%), whereas 294 residues formed the coils (46.52%) of the whole epitope-based bovine TB vaccine sequence (Figure 5.12).
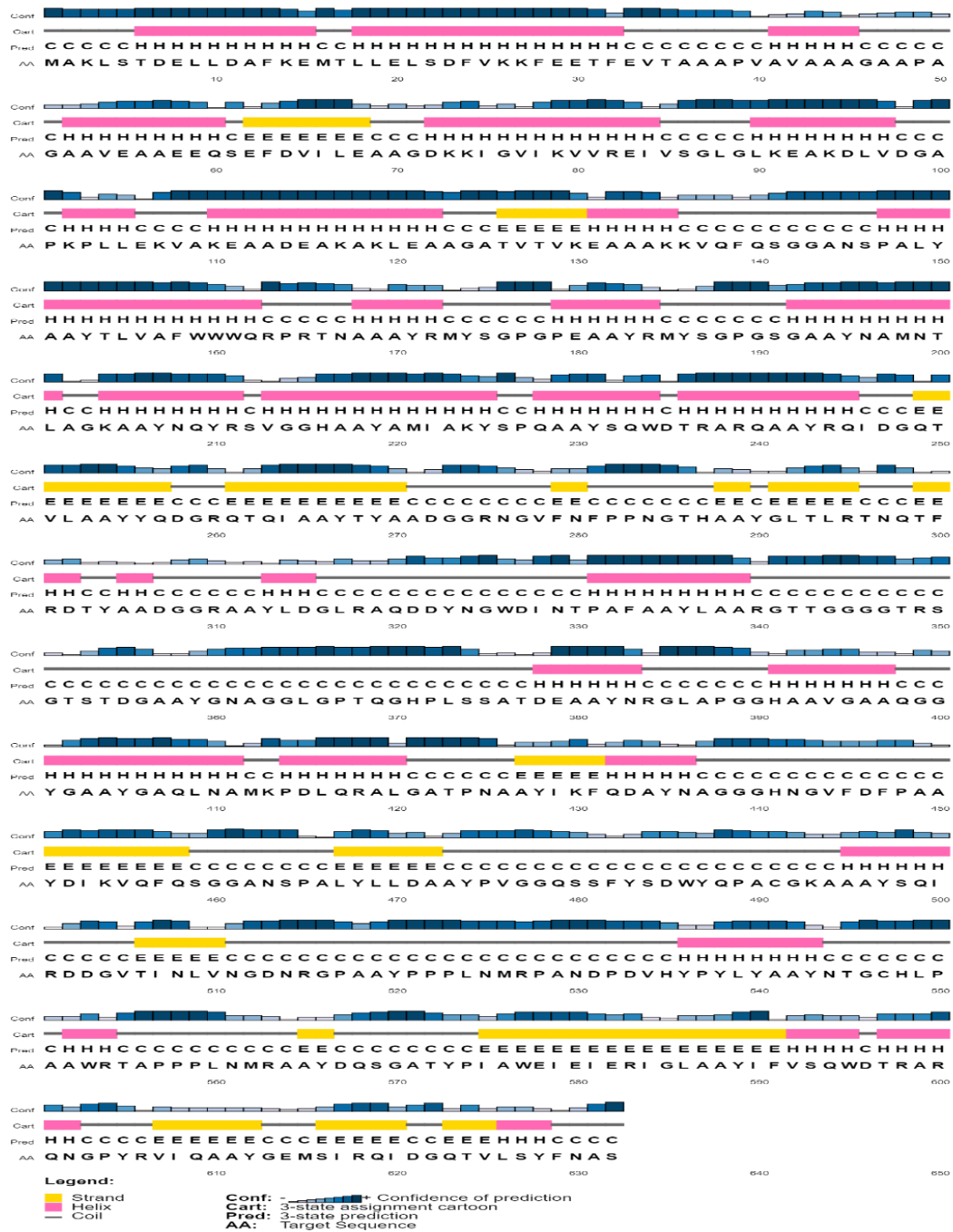
Figure 5. 4: Secondary structure prediction for the bovine TB vaccine sequence using PSIPRED. The sequence comprised alpha-helices (37.82%), beta-strands (15.66%) and coils (46.52%)

The RaptorX server was used for constructing a tertiary structure model of the epitope-based bovine TB vaccine. It performed homology modelling using 6VHD_A PDB entry as a template for generating a structural model. All 632 amino acid residues were modelled with only 8% in the disordered region. The constructed model was 39% exposed, 29% medium and 31% buried in folded conformation. Before refining the structural model, the Ramachandran plot showed that 85.4% of amino acid residues were in the most favoured region. Thus, the refinement of the structural model was undertaken using the GalaxyRefine server to improve the quality of the predicted structural model. Figure 5.13 (i) shows the 3D structural model of the bovine TB vaccine after refining. The refined structural model had 91.1% amino acid residues in the most favoured region (Figure 5.13 (ii)). Other parameters evaluated after refinement were GDT-HA score of 0.9568, MolProbidity 3.365 and RMSD 0.412.
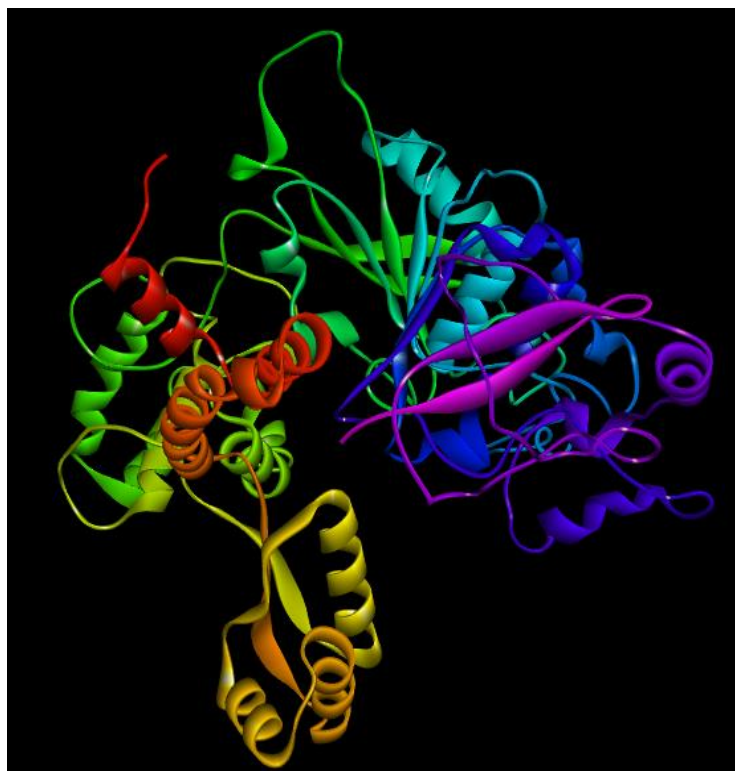
(i)
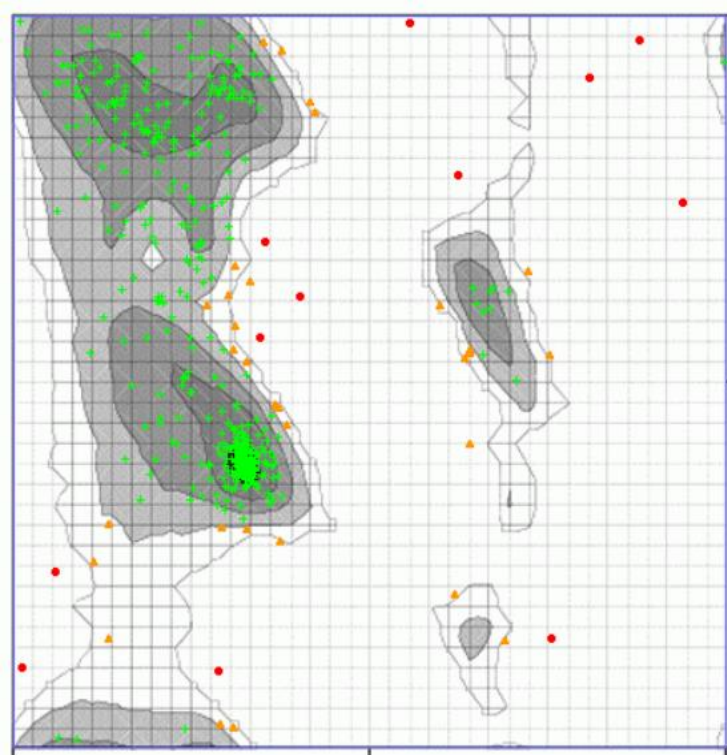
(ii)

Figure 5. 5: Refined structural model of bovine TB vaccine construct. (i) The 3D structure is coloured in rainbow colours, from violet to red (from N-terminus to C-terminus), and (ii) the Ramachandran plot of the refined structure showing 91.1% amino acid residues in the most favoured region (green crosses)

5.4.1.5 Analysis of the interaction of bovine TB vaccine and toll-like receptors

TLRs interact with the pathogen/vaccine antigen for initiating the innate immune response. The stable interaction of a vaccine with receptors of host immune cells activates the immune system. For determining the ability of the designed bovine TB vaccine in generating an immune response, molecular docking analysis was performed. First, the structures of the bovine TLRs were constructed using the I-TASSER server (Figure 5.13). The protein sequences used for structure construction were Q95LA9, Q9GL65 and B5T278 for TLR-2, TLR-4 and TLR-6.
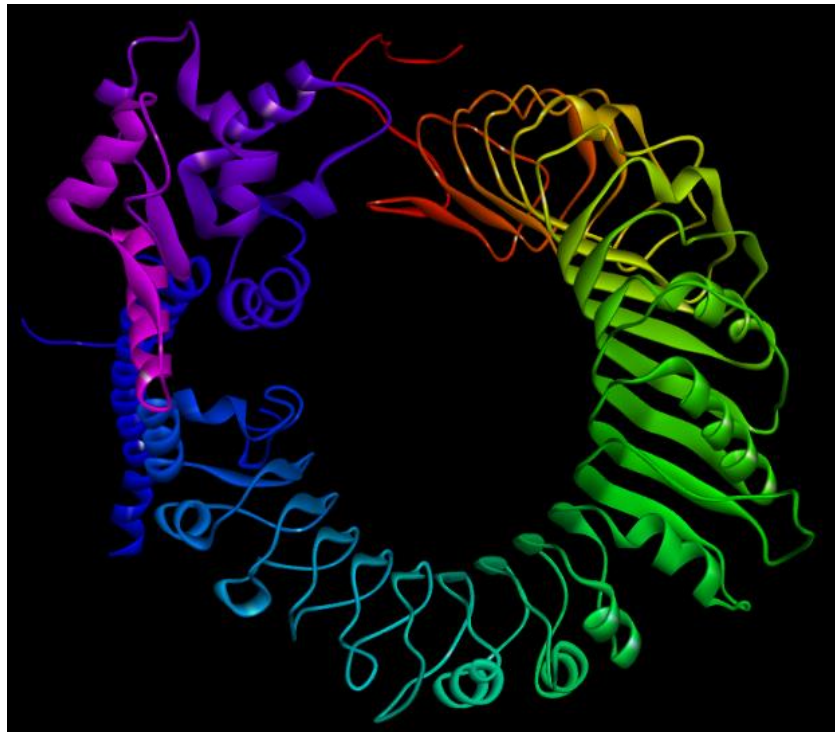
(i)
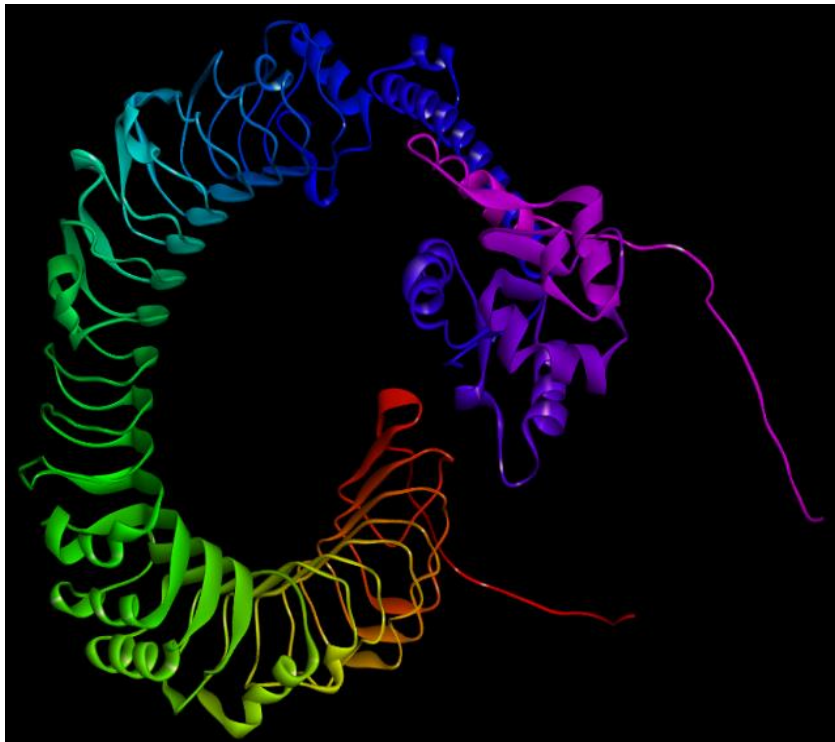
(ii)

(iii)

Figure 5. 6: Refined structural model of toll-like receptors of cattle. (i) TLR-2, (ii) TLR-4, and (iii) TLR-6. The 3D structure is in rainbow colours from violet to red (from the N-terminus to the C-terminus)

After the construction of the TLRs, PatchDock and FireDock were used to analyse the interactions between the bovine TB vaccine construct and the TLRs of cattle. The output of the PatchDock docking process displayed several docked transformations, which were further presented to FireDock for calculation of the global binding energy. PatchDock found 1455 docked transformations for TLR-2 and the bovine TB vaccine docked complex (Table 5.9). The lowest binding energy of -55.15 kcal/mol was predicted for the 433rd transformation of the docked complex. The global binding energies for the bovine TB vaccine docked with TLR-4 and TLR-6 were -61.78 kcal/mol and -53.89 kcal/mol, respectively. The docking results revealed that the bovine TB vaccine designed was adequately engaged in interaction with the selected TLRs and would accomplish the goal of initiating a strong immune response.

Table 5. 9: Molecular docking of bovine the TB vaccine and the toll-like receptors of the host (cattle). aVdW-attractive Van der Waals energy, rVdW-repulsive Van der Waals energy and HB-hydrogen bonding

Docked Complex	Total number of docked transformations	Transformation with lowest binding energy	Global binding energy (kcal/mol)	aVdW (kcal/mol)	rVdW (kcal/mol)	HB (kcal/mol)
Bovine TB vaccine-TLR2	1455	433rd	-55.15	-39.38	26.12	-4.1
Bovine TB vaccine-TLR4	1398	37th	-61.78	-39.51	22.19	-4.7
Bovine TB vaccine-TLR6	2911	12th	-53.89	-36.96	20.78	-5.21

This concludes the design of the epitope-based vaccine for bovine TB. Next section presents the identification of drug targets from selected conserved proteins. The aim is to select the most effective drug targets for designing improved drugs.

5.4.2 Identification of nine potential drug candidates from the 1163 conserved proteins within 11 strains of *Mycobacterium bovis*

The objective of this part of the study was to identify potential drug candidates for bovine TB. We performed some crucial steps to discover unique drug targets that are pivotal for the survival of *Mycobacterium bovis* and do not initiate hypersensitive reactions or side effects in the host. In our study, we took *Mycobacterium bovis* AF2122/97 as a reference organism for research and it has a total of 3988 protein sequences in its proteome. Through comparative proteomic analysis, we found 1163 conserved proteins within 11 strains of bovine TB bacteria. Using the 1163 conserved proteins, we performed a subtractive proteomics approach to identify drug targets that could further help investigate therapeutic drugs for the treatment of bovine TB. Table 5.10 shows the summary of the results of the subtractive proteomic analysis

Table 5. 10: Exploration of potential drug targets for bovine TB using the subtractive proteomic approach for *Mycobacterium bovis*

Steps	Result
Total proteins	3988
Conserved proteins	1163
Non-homologous proteins to cattle	815
Non-homologous host metabolic pathways	735
Proteins in unique metabolic pathways	386
Essential metabolic proteins	195
Virulent proteins	29
Highly expressed target proteins	13
Potential bovine TB drug target	9

The homologous proteins present in *Mycobacterium bovis* and the host are not considered to be effective drug targets. Thus, to exclude homologous proteins, we performed sequence similarity analysis using BLASTp. The protein sequence identities of 30% or more, i.e., bit scores more than 100 were eliminated from the study. Out of 1163 conserved proteins, we found 348 homologous protein sequences between the host and *Mycobacterium bovis.* These 348 proteins were excluded, and the 815 non-homologous proteins were considered for further examination. To enhance the drug targets' safety profile, we performed an analysis to identify proteins involved in metabolic pathways common in cattle and *Mycobacterium bovis*. Seven hundred and thirty-five non-homologous proteins were subsequently obtained that were not involved in metabolic pathways similar to the host's metabolic system.

For determining the involvement of these 735 proteins in unique metabolic pathways, we used KEGG **a**utomated **a**nnotation **s**erver (KAAS). In the sub-sequential analysis, we found that more than 50% of the non-homologous proteins were involved in distinctive pathways in *Mycobacterium bovis.* A total of 386 proteins were involved in the unique metabolic pathways of *Mycobacterium bovis*. The proteins retrieved were first classified as enzymatic or non-enzymatic with the help of enzyme classification (EC) numbers. Around 56.74% of the proteins were enzymatic and 43.26% were non-enzymatic. In the KEGG database, the pathways were broadly classified into seven categories: metabolism, genetic information processing, environmental information processing, cellular processes, organism system, diseases and drugs.

Out of 386 proteins, 80.05% (309 proteins) were involved in different metabolic pathways (Figure 5.14). For example, 60 proteins were involved in amino acid metabolism followed by 48 in carbohydrate metabolism, 34 in energy metabolism, 18 in lipid metabolism, 11 in nucleotide metabolism and the remaining proteins were involved in the biosynthesis of secondary metabolites, cofactors and vitamins, terpenoids and polyketides. Approximately 26 proteins (14%) were involved in genetic and environmental information processing, 14 in cellular processes (3%), four in the organism's system, three in diseases and four in drug categories (Figure 5.14).
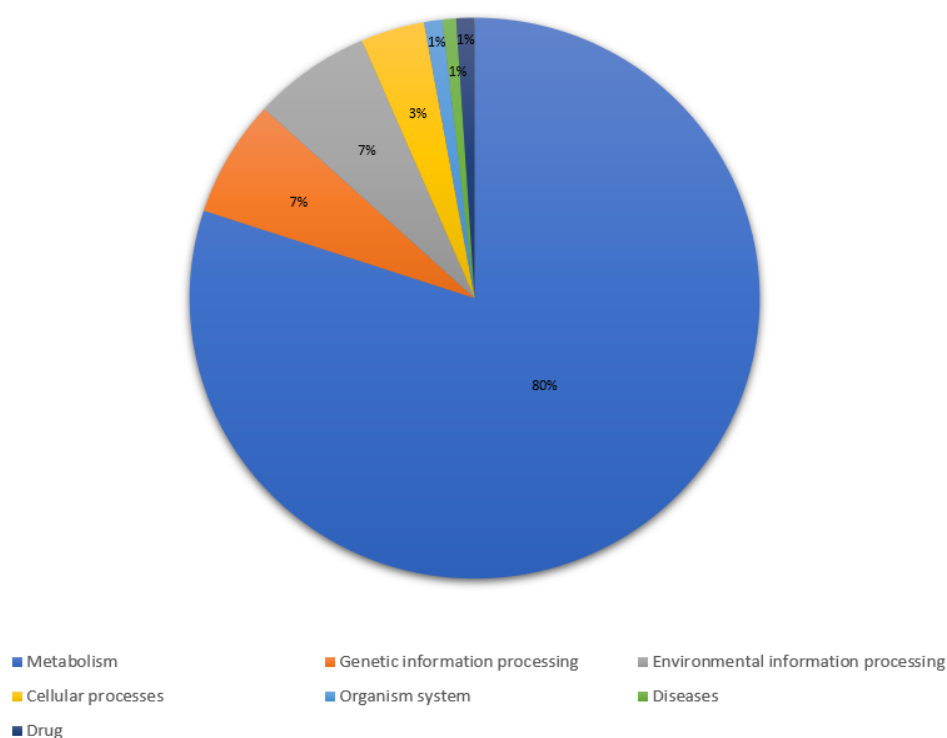
Figure 5. 7: Percentage distribution of non-homologous proteins in seven pathways: metabolism, genetic information processing, environmental information processing, cellular processes, organism system, diseases and drugs

Table 5.11 shows the classification of the 386 proteins based on the protein families involved in metabolism, genetic information processing, signalling mechanism and cellular processes.

Table 5. 11: Distribution of the 386 proteins found in unique pathways of *Mycobacterium bovis* in the protein families involved in the seven KEGG functional categories

Proteins involved in unique metabolic pathways	Number of proteins
Enzymes	219
Transporters	31
DNA repair and recombination proteins	16
Chromosome and associated proteins	11
Peptidases and inhibitors	10
Peptidoglycan biosynthesis and degradation proteins	9
Glycosyltransferases	9
Two-component system	9
Lipid biosynthesis proteins	9
Transcription factors	8
Ribosome biogenesis	6
Transfer RNA biogenesis	5
Protein kinases	5
Cytochrome P450	5
Prokaryotic defence system	5
DNA replication proteins	3
Chaperones and folding catalysts	3
Polyketide biosynthesis proteins	3
Proteasome	3
Cytoskeleton proteins	3
Secretion system	3
Transcription machinery	3
Bacterial toxins	3
Ribosome	2
Lipopolysaccharide biosynthesis proteins	1
Translation factors	1
Amino acid related enzymes	1

Identifying non-homologous proteins involved in the unique metabolic pathways is not the sole criterion for selecting potential drug targets for bovine TB. A protein might be involved in more than two or three metabolic pathways, but it could be a protein non-essential protein for *Mycobacterium*

bovis survival. Thus, it is crucial to target essential proteins that disrupt the normal functioning of bovine TB bacteria. Based on the criteria mentioned, above, we used the DEG database to select proteins that are important for the survival of *Mycobacterium bovis*. Proteins showing good sequence similarity, 100% query coverage and a bit score more than 100 with the laboratory validated essential proteins present in DEG were considered essential proteins. A total of 195 essential proteins were retrieved in this step.

Other crucial criteria for the identification of drug targets are virulence and pathogenicity. A non-homologous and essential protein involved in initiating the virulence mechanism in the host is considered a pivotal drug candidate. Targeting virulent proteins that help survive pathogenic bacteria inside the host would swiftly eliminate bovine TB disease. VirulentPred and MP3 tools were used for predicting the virulence of essential proteins. Twenty-nine proteins, commonly predicted as virulent by both methods, were selected for further analysis. In the next step, the codon adaptation index (CAI) was calculated. CAI indicates the translational efficiency and level of expression of a gene. CAI helps choose highly expressed proteins from the set of selected essential and virulent proteins. The value of CAI score ranges from 0 to 1 in the bacterial genome and proteins with a CAI score of more than 0.7 are likely to be highly expressed in the *Mycobacterium bovis* genome. Of the 29 virulent proteins, only 13 proteins were found to have a CAI score greater than 0.7. Table 5.12 provides information on the 13 proteins selected. The chosen proteins were further analysed for their physiochemical properties, structural homology and druggability properties.

Table 5. 12 : Results of subtractive proteomic approach for identification of non-homologous, essential, virulent and highly expressed proteins

Protein Accession	Protein Note	KEGG ID	DEG	Role in virulence	CAI
YP_009357346.1	PROBABLE PENICILLIN-BINDING PROTEIN pbpA	K05364	Yes	Yes	0.718
YP_009357462.1	Maltokinase mak	K16146	Yes	Yes	0.716
YP_009357464.1	SECRETED ANTIGEN 85-C FBPC (85C) (ANTIGEN 85 COMPLEX C) (AG58C) (MYCOLYL TRANSFERASE 85C) (FIBRONECTIN-BINDING PROTEIN C)	K18851	Yes	Yes	0.76
YP_009357822.1	PROBABLE UDP-N-ACETYLENOLPYRUVOYLGLUCOSAMINE REDUCTASE MURB (UDP-N-ACETYLMURAMATE DEHYDROGENASE)	K00075	Yes	Yes	0.734
YP_009358286.1	PROBABLE PHOSPHATE-TRANSPORT INTEGRAL MEMBRANE ABC TRANSPORTER PSTA1	K02038	Yes	Yes	0.701
YP_009358342.1	Two component sensor kinase mprb	K07653	Yes	Yes	0.711
YP_009358652.1	PROBABLE OLIGOPEPTIDE-TRANSPORT INTEGRAL MEMBRANE PROTEIN ABC TRANSPORTER OPPB	K02033	Yes	Yes	0.703

YP_009358831.1	POSSIBLE CONSERVED INTEGRAL MEMBRANE PROTEIN	K14339	Yes	Yes	0.722
YP_009358851.1	Possible invasion protein	K21474	Yes	Yes	0.747
YP_009359528.1	Probable penicillin-binding membrane protein pbpB	K03587	Yes	Yes	0.731
YP_009359962.1	POSSIBLE GLYOXALASE II (HYDROXYACYLGLUTATHIONE HYDROLASE) (GLX II)	K01069	Yes	Yes	0.776
YP_009360863.1	PROBABLE DICARBOXYLIC ACID TRANSPORT INTEGRAL MEMBRANE PROTEIN KGTP (DICARBOXYLATE TRANSPORTER)	K03761	Yes	Yes	0.709
YP_009361152.1	Probable two component transcriptional regulatory protein tcrx	K02483	Yes	Yes	0.749

ProtParam was used to predict the physiochemical properties of the selected proteins. Table 5.13 shows the physiochemical properties of the selected 13 proteins of *Mycobacterium bovis*. The range of amino acid length was from 224-679 residues and the molecular weights varied from 21-72 kDa. All 13 proteins had a molecular weight lower than 100 kDa or 100000 Da, making them possible drug targets. Out of the 13 proteins, eight proteins were stable with an instability index of lower than 40. The presence of dipeptides was the reason for the unstable nature of the remaining five proteins. The aliphatic index predicts the relative volume occupied by the aliphatic side chains in a protein. A value of aliphatic index greater than 30 indicates that the protein is thermodynamically stable. All 13 proteins had aliphatic index greater than 30, indicating the highly thermodynamically stable nature of the selected protein molecules. The GRAVY score predicts the hydropathy value, indicating the hydrophilic or hydrophobic nature of a protein molecule. Table 5.13 shows that six proteins had negative GRAVY scores indicating their hydrophilic nature. The remaining proteins had a GRAVY score below one, suggesting they are less hydrophobic molecules. Analysis of physiochemical properties showed that the selected 13 proteins have all the characteristics of potential drug targets.

Table 5. 13: Physiochemical properties of the 13 non-homologous, essential, virulent and highly expressed proteins of *Mycobacterium bovis*

Protein Accession	Amino acids	MW (Dalton or Da)	pI	Extinction	Instability index	Aliphatic index	GRAVY
YP_009357346.1	491	51575.6	8.59	36120	32.66	87.17	-0.085
YP_009357462.1	455	49889.11	5.06	69455	38.05	88.62	-0.185
YP_009357464.1	340	36771.27	5.92	85830	49.34	68.97	-0.253
YP_009357822.1	369	38521	6.4	25690	27.64	99.95	0.133
YP_009358286.1	304	32988.36	10.35	54430	41.3	129.51	0.868
YP_009358342.1	504	54431.72	9.67	43430	50.18	102.38	0.006

YP_009358652.1	325	35083.25	10.68	46410	27.19	120.34	0.517
YP_009358831.1	591	62694.38	10.95	112535	36.24	121.4	0.631
YP_009358851.1	241	24623.99	9.71	26930	31.48	81.87	0.033
YP_009359528.1	679	72537	9.64	49975	43.17	79.47	-0.346
YP_009359962.1	224	21497.63	6.14	10095	26.06	98.35	-0.021
YP_009360863.1	449	49578.72	9.37	81500	38.34	102.92	0.461
YP_009361152.1	234	25843.77	5.65	19940	52.93	104.49	-0.059

The selected proteins of *Mycobacterium bovis* were then investigated for structural homology. A template in the PDB database makes a protein a suitable target for drug discovery and development. For this, structural similarity analysis was performed using BLASTp. The proteins with a sequence identity and bit score of more than 30% and 100, respectively, were selected in the study (Table 5.14). Out of the 13 proteins, four proteins with no suitable template structure were eliminated from the study.

Table 5. 14: Results of template search analysis of the selected nine potential drug targets

Protein Accession	Query cover	Sequence similarity	PDB ID
YP_009357346.1	93	100	3LO7_A
YP_009357462.1	100	99.78	4O7O_A
YP_009357464.1	86	100	4MQM_A
YP_009357822.1	100	100	5JZX_A
YP_009358342.1	30	72.26	6BLK_A
YP_009358851.1	87	100	3PBI_A
YP_009359528.1	82	99.82	6KGH_A
YP_009359962.1	95	32.57	2XF4_A
YP_009361152.1	95	54.67	5ED4_A

Finally, the druggability potential of the nine shortlisted drug targets were predicted using DrugBank. Six target proteins were found to have significant hits in DrugBank with bit scores of more than 100

(Table 5.15). The six proteins (YP_009357346.1, YP_009357464.1, YP_009357822.1, YP_009358342.1, YP_009359528.1 and YP_009361152.1) with considerable similarity found in DrugBank, were termed druggable target proteins. The remaining three (YP_009357462.1, YP_009358851.1 and YP_009359962.1) proteins with no significant hits found were considered novel target proteins.

The two FDA approved drugs Cefepime and Cefiderocol were found to target the same proteins, having accession numbers YP_009357346.1 and YP_009359528.1. Cefepime and Cefiderocol have the same mode of action as beta-lactam antibiotics, such as penicillin. They disrupt the synthesis of peptidoglycan in the bacterial cell wall, thus, weakening the structural integrity of the pathogenic bacteria. Cefpodoxime was found to be vet-approved with the potential to target the YP_009359528.1 protein by inhibiting the synthesis of peptidoglycan, so deteriorating the bacterial cell wall. Cefpodoxime, flavin adenine dinucleotide approved compound, could target YP_009357822.1 protein, but the mechanism of compound action is not yet established. Experimental compounds were found for YP_009357464.1, YP_009358342.1 and YP_009361152.1 but no mechanism of action is known.

Table 5. 15: Identified drugs, FDA approval or experimental compounds for them and their action on the six druggable proteins from the DrugBank

Protein Accession	Potential drugs	Drug group	Action
YP_009357346.1	Cefepime, Cefiderocol	Approved	Inhibitor
YP_009357464.1	Diethyl phosphonate	Experimental	Unknown
YP_009357822.1	Flavin adenine dinucleotide	Approved	Unknown
YP_009358342.1	Phosphoaminophosphonic acid-adenylate ester	Experimental	Unknown
YP_009359528.1	Cefepime, Cefiderocol, Doripenem, Cefpodoxime	Approved	Inhibitor
YP_009361152.1	Adenosine-5'-RP-alpha-thio-triphosphate	Experimental	Unknown

In our study, we selected nine proteins as potential broad-spectrum drug targets (YP_009357346.1, YP_009357462.1, YP_009357464.1, YP_009357822.1, YP_009358342.1, YP_009358851.1, YP_009359528.1, YP_009359962.1 and YP_009361152.1) for bovine TB treatment that minimises the chances of drug resistance. YP_009357346.1 was a probable penicillin-binding protein having 491

amino acid residues in its sequence. This protein was found to be involved in peptidoglycan biosynthesis, thus, maintaining the cell wall and cellular structure of *Mycobacterium bovis*. The disruption of this protein by a drug would cause structural irregularity and loss of cell wall integrity, leading to the weakening of bacterial cells. Two potential drugs, Cefepime and Cefiderocol, were found to have the capability to disrupt the YP_009357346.1 protein. YP_009357464.1 was present in the cell wall of *Mycobacterium bovis*. This was involved in lipid metabolism for the biosynthesis of arabinogalactan. Arabinogalactan is an important cell wall constituent and provides strength to the cell wall of *Mycobacterium bovis*. Inhibition of the biosynthesis of arabinogalactan can make the bacteria fragile. DrugBank identified diethyl phosphonate as a potential experimental compound for targeting this protein.

YP_009357822.1, present in the cytoplasm, was involved in amino sugar and nucleotide sugar metabolism with a significant role in cell division of the bacterial cell. Based on its function, targeting this protein would interfere in the cell division of *Mycobacterium bovis.* DrugBank identified the flavin adenine dinucleotide compound with the potential to target the YP_009357822.1 protein, but the compound's action is unknown. YP_009358342.1, found in the cell membrane of *Mycobacterium bovis,* was involved in the environment information processing pathway. This two-component system present in mycobacteria plays a crucial role in the signal transduction mechanism required for continuing the bacterial infection. Attacking this protein by drugs would lead to halting the signalling mechanism; thus, containing the spread of bacterial infection. The phosphoaminophosphonic acid-adenylate ester experimental compound could target YP_009358342.1 protein, but the mechanism of the compound's action is not yet established.

YP_009359528.1, a probable penicillin-binding membrane protein, pbpB, present in the cell wall, is involved in peptidoglycan biosynthesis. Four potential FDA approved drugs were identified for this drug target: Cefepime, Cefiderocol, Doripenem and Cefpodoxime. The role of all four drugs is inhibition of the synthesis of peptidoglycans. Cefpodoxime was found to be a vet-approved drug. YP_009361152.1, a two-component system protein, was found in the cytoplasm of *Mycobacterium bovis*. This protein is involved in the environment information processing pathway. Adenosine-5'-RP-alpha-thio-triphosphate was found to have the potential for attacking YP_009361152.1 protein.

The three (YP_009357462.1, YP_009358851.1 and YP_009359962.1) proteins with no significant hits found were considered to be novel target proteins. The YP_009357462.1 protein was found to be involved in carbohydrate metabolism and present in the cytoplasm of *Mycobacterium bovis*. The primary function of this protein is to transfer phosphorus-containing groups in starch and sucrose metabolism. Targeting carbohydrate metabolism usually induces a stress response in the cell leading

to the death of bacteria. YP_009358851.1 was involved in peptidoglycan biosynthesis and degradation. An effective drug with an inhibitory mechanism would contribute to disrupting the synthesis of peptidoglycan. YP_009359962.1 present in the cytoplasm is involved in carbohydrate metabolism, specifically, pyruvate metabolism. This protein benefits the bacterial cell during stress conditions. Thus, YP_009359962.1 is a potential target for enhancing stress inside the bacterial cell. The nine proteins selected are highly conserved, unique to *Mycobacterium bovis*, and involved in essential and virulent pathways. These nine proteins can be used for drug discovery and development against *Mycobacterium bovis*.

5.5 Chapter Summary

Bovine TB is a chronic infectious disease caused by the pathogenic bacteria, *Mycobacterium bovis.* It has become of global concern over the last two decades. Bovine TB primarily affects cattle, but other domestic livestock, such as deer, goats and sheep are also affected. Bovine TB is more common in less-developed and developing countries. The disease can lead to a severe economic crisis because of significant losses of livestock, and the subsequent trade restrictions. Bovine TB is a zoonotic disease that can spread to humans by the inhalation of aerosols or the ingestion of unpasteurised milk. Thus, the zoonotic potential of bovine TB is raising health concerns for the public. Out of 10 million incidences in humans, in 2019, WHO estimated 0.14 million cases were zoonotic TB caused by *Mycobacterium bovis*, with 11,400 human deaths (WHO Global Tuberculosis Report, 2020; World Organization for Animal Health, 2019). Currently, there is no effective treatment available for bovine tuberculosis due to its infectious nature and the drug resistance of *Mycobacterium bovis*. The available treatment for bovine tuberculosis mainly depends on the health status of the infected animal species. Antibiotic therapy can only be given to the animal species living in captivity, as it is not reliable for herd or free-grazing animals. The BCG vaccine is another option available for treating disease, but it shows limited efficacy in cattle.

In our study, we attempted to provide a potential solution for bovine TB treatment by identifying therapeutic targets. With this aim, we proposed an effective epitope-based vaccine for bovine TB. In designing an effective epitope-based bovine TB vaccine that would elicit a robust immune response inside the host, we focussed on selecting the highly immunodominant epitopes from the conserved and surface-exposed antigenic proteins of *Mycobacterium bovis*. Our study incorporates different branches of bioinformatics, such as subtractive reverse vaccinology, immunoinformatics and structural vaccinology, to identify potential vaccine candidates and design an epitope-based bovine TB vaccine. Our study used 11 strains of *Mycobacterium bovis* to cover their diversity and to identify the conserved proteins among them. *Mycobacterium bovis* AF2122/97 was used as a reference proteome for comparative analysis. Using standalone BLAST, we found that out of 3988 proteins of *Mycobacterium*

bovis AF2122/97, 1163 proteins were discovered to be conserved among 11 strains of *Mycobacterium bovis*. After identifying conserved proteins, a reductionist reverse vaccinology process was performed to identify non-homologous, surface-exposed, antigenic and non-allergenic proteins having unique characteristics, such as signal peptides, membrane-spanning regions, lipoprotein signatures and adhesion probabilities. Out of 1163 conserved proteins, a total of nine conserved, membrane-spanning, antigenic and non-allergic proteins were selected from the reverse vaccinology approach. Next, extensive immunoinformatic analysis was performed and we shortlisted 26 epitopes (HTL epitopes-2, CTL epitopes-8 and B-cell epitopes-16) from the nine antigens of *Mycobacterium bovis*. These 26 epitopes were highly immunogenic, non-toxic and non-allergenic to the host.

The epitopes selected were used for designing the vaccine sequence. The sequence of the bovine TB vaccine was created using a simple strategy. A 50s ribosomal protein L7/L12 adjuvant was attached at N- terminal connected with the help of EAAAY linker followed by HTL, CTL and B-cell epitopes bound together, using the flexible linker, AAY. Assessment of antigenicity, allergenicity and the physiochemical properties of the bovine TB vaccine showed that the vaccine sequence designed had all the qualities of a potential vaccine that would generate an effective immune response inside the host. A model of the bovine TB vaccine was constructed using RaptorX. Structural analysis of the epitope-based bovine TB vaccine construct model highlighted its structural integrity with more than 91.1% amino acid residues in the most favourable regions. Molecular docking of bovine TB vaccine construct and TLRs (TLR-2,4 and 6) was then undertaken. The results of our docking analysis suggested stable and robust interactions between the TB vaccine construct and TLRs with binding energies of - 55.15 kcal/mol (bovine TB vaccine-TLR2), -61.78 kcal/mol (bovine TB vaccine-TLR4) and -53.89 kcal/mol (bovine TB vaccine-TLR6). The strong interaction between the bovine TB vaccine construct and the host TLRs showed that the designed vaccine can initiate a strong immune response inside the host.

Next, we performed subtractive proteomic analysis to identify potential drug targets that could further help to investigate therapeutics drugs for the treatment of bovine TB. We performed some crucial steps in discovering unique drug targets that are pivotal for the survival of *Mycobacterium bovis* and do not initiate hypersensitive reactions or side effects in the host. We took 1163 conserved proteins (identified in comparative proteomic analysis in bovine TB vaccine part) of *Mycobacterium bovis* AF2122/97 and undertook a subtractive proteomics approach to identify drug targets. First, the homologous protein sequences between *Mycobacterium bovis* and host (cattle) were excluded from the study. Thus, we considered 735 non-homologous proteins out of the 1163 proteins. In the subsequent analysis, we found that more than 50% of the non-homologous proteins were involved in distinctive pathways in *Mycobacterium bovis.* Specifically, a total of 386 proteins were found to be

involved in the unique metabolic pathways of *Mycobacterium bovis*. Around 56.74% of these proteins were enzymatic and 43.26% were non-enzymatic proteins.

An assessment of other crucial criteria, such as essentiality of target for survival, virulence, and determining the translation efficiency of the selected protein, to identify potential drug targets was then carried out. Out of 195 essential proteins predicted by DEG, 29 proteins were commonly predicted as virulent by the VirulentPred and MP3 tools. Of the 29 virulent proteins, only 13 proteins were found to have a CAI score greater than 0.7. The 13 proteins selected were further analysed for their physiochemical properties, structural homology and druggability properties. Analysis of their physiochemical properties showed that the 13 proteins selected all had characteristics of potential drug targets. Structural homology analysis showed that out of the 13 proteins, four proteins had no suitable template structure available in the PDB database and were excluded from the study. Finally, the druggability potential of the shortlisted nine drug targets was predicted using DrugBank. The six proteins (YP_009357346.1, YP_009357464.1, YP_009357822.1, YP_009358342.1, YP_009359528.1 and YP_009361152.1) with considerable similarity found in DrugBank were termed druggable target proteins. The remaining three (YP_009357462.1, YP_009358851.1 and YP_009359962.1) proteins with no significant hits found were considered novel target proteins. The nine highly potential drug targets identified in the study could facilitate the development of novel drugs against *Mycobacterium bovis.*

Thus, using an extensive and systematic approach, this chapter presented the development of an epitope-based vaccine that would be far more effective than the current BCG vaccine for bovine TB. It also found from a comprehensive investigation nine potential drug targets that could be far more effective than the currently used drugs in treating TB. Next Chapter provides a summary and conclusions from the thesis.

Chapter 6

Summary, Conclusions and Contributions

In this thesis, we conducted an in-depth investigation into pathogen-host interactions to gain deeper insights into the pathogenesis of infection, evolution of pathogen and their drug resistance mechanisms and use this understanding to provide potential solutions for effective vaccine and drug development for human and bovine tuberculosis. The central focus of our study was to acquire a deeper understanding of the pathogenesis of human and bovine TB, the interaction of TB bacteria with its host, the host defence mechanism, bacterial survival strategies in evading the host immune response, in-depth knowledge of the mechanisms of TB drug resistance, and challenges in developing drugs and vaccines for TB (Chapter 2). Thus, for reducing the burden of TB globally, we developed holistic strategies/frameworks for accomplishing the three main goals of this study. In this chapter, we give a brief overview of our research work and conclusions in section 6.1. Section 6.2 highlights the significant contributions of the study, and section 6.3 suggests the directions for future work.

6.1   Research summary and conclusions

Tuberculosis is an evolving and deadly disease caused by one of the world's most infectious bacteria, *Mycobacterium tuberculosis,* which poses a significant threat to global health. According to the World Health Organization (WHO), tuberculosis is a major threat with significant mortality and morbidity rates worldwide. Despite the advancements made in medical science, TB remained the cause of death of 1.4 million people in 2019 and was responsible for 10.0 million new cases worldwide (WHO Global Tuberculosis Report, 2020). Vaccine (BCG) and drug therapy (first-line TB drugs) are the two most essential countermeasures humans have developed against TB. The evolution of *Mycobacterium tuberculosis* has led to the emergence of drug-resistant TB strains that makes drugs ineffective. Protection of BCG is highly variable for the following reasons: differences in clinical assays, genetic variability in a sample population, different levels of protection against the clinical forms of tuberculosis, malnutrition and variability in *Mycobacterium tuberculosis* strains. The worldwide emergence of tuberculosis resistant strains and inefficiency of BCG in reducing the prevalence and emergence of the disease and protecting adults is threatening to make one of humankind's most important infectious diseases incurable.

Bovine tuberculosis, caused by *Mycobacterium bovis*, has also become a global concern over the last two decades. *Mycobacterium bovis* is considered naturally resistant to pyrazinamidase (first-line TB drug) (Nakajima et al., 2010). First-line human TB drugs for treating livestock are also ineffective and costly as the treatment requires six to nine months of daily medication. The BCG vaccine is available

for treating the disease, but it shows limited efficacy in cattle. There is no effective treatment available for bovine TB due to its infectious nature and the drug resistance of *Mycobacterium bovis*. The prevention of bovine TB is a long-term goal that can only be accomplished by developing a more effective vaccine than BCG and designing new drugs based on a deep understanding of host-pathogen interaction and mechanisms of drug resistance that can help eliminate the disease at its root.

Goal 1: The first goal of this research is to understand the impact of drug resistant mutations to gain an in-depth knowledge of the survival strategies of *Mycobacterium tuberculosis* and resulting drug resistance mechanisms (Chapter 3). For an extensive investigation of drug resistant mutations, we designed a holistic framework to unravel the survival strategies of drug resistant TB bacteria and an in-depth look into global mutation patterns to explore the global evolution of drug resistance. Therefore, we did a comprehensive and systematic in-depth analysis of drug resistance mechanisms from global mutation data for *Mycobacterium tuberculosis* reported over the last 30 years. We collected mutational data from 31,073 *Mycobacterium tuberculosis* isolates published in 149 studies over 30 years and found 821 non-synonymous drug resistant mutations in the four first-line drug targets (*katG*-202, *pncA*-273, *rpoB*-120 and *embCAB*-226). The coverage of a large number of strains (mutations) is novel in this study as well as the study of the mutational impact on targets and drugs. We investigated the mutation statistics by calculating single mutation frequency to understand the prevalence and diversity of mutations in first-line TB drug targets across the globe. S315T was highly prevalent (60.58%), followed by S450L (58.83%), M306V (29.19%) and Q10P (1.45%) among the INH, RIF, EMB and PZA resistant isolates, respectively. Our research focused on identifying the impact of different mutations on TB bacteria and drug binding, using detailed bioinformatics analysis to understand crucial changes at the molecular level of the target affecting its function, structural stability, sequence conservation and the influence of mutation position on drug binding affinity. This research introduced a new concept of ranking drug-resistant TB mutations into lethal, moderate, mild and neutral to uncover the totality of the *Mycobacterium tuberculosis* strategy for drug resistance. Out of 821 non-synonymous mutations, we identified 340 'lethal,' 284 'moderate,' 185 'mild' and 12 'neutral' mutations. We observed that the frequently occurring drug resistant mutations had mild to moderate impacts on drug binding with reduced drug binding energies, changes in enzymatic activity and low steric hindrance caused by structural changes. For example, *katG* S315T mutation, *rpoB* S450L mutation and *embB* M306V mutation were widespread across the globe but had only a mild impact on their respective proteins.

Conclusion: The comprehensive analysis of drug resistant mutations for determining the prevalence and ranking of mutations based on their effect on *Mycobacterium tuberculosis* has provided crucial insights and improved our knowledge of the survival strategies used by this bacterium to maintain its

fitness. We found that *Mycobacterium tuberculosis* explores specific drug resistance strategies that balances the harmful impact of drug-resistant mutations on itself and drug resistance by following evolutionary pathways. We discovered that most frequent first-line drug resistant mutations had followed a generic pattern over the last three decades, suggesting that *Mycobacterium tuberculosis* may also follow the same general pattern across the globe in the future. The method developed in this study can also help in studying future mutations, and there is the scope for introducing new steps in the method for improvement. Further, our method can be used to examine the nature and impact of mutations in other infectious diseases.

Goal 2: The second goal of this research is to develop an epitope-based *in-silico* human TB vaccine for greatly improved efficacy (Chapter 4). The low efficacy of BCG and drug-resistant *Mycobacterium tuberculosis* strains require developing a more effective and safer vaccine. This research developed a conceptual framework that uses different bioinformatics approaches, such as comparative proteome analysis, reverse vaccinology, immunoinformatics, and structural vaccinology, to identify potential vaccine candidates and construct an epitope-based TB vaccine. We performed a comparative proteomic analysis among the 159 strains of *Mycobacterium tuberculosis* to cover the diversity and identify conserved proteins among those strains. We found that out of 3906 proteins of H37Rv, 1982 proteins were conserved. Out of 1982 conserved proteins, 24 antigenic proteins with unique characteristics, such as signal peptides, membrane-spanning regions, lipoprotein signatures, adhesion probabilities and non-allergic proteins, were selected from the reverse vaccinology approach. From 24 shortlisted proteins, an extensive immunoinformatics analysis provided highly immunogenic, non-toxic and non-allergenic 27 epitopes (CTL epitopes-14, HTL epitopes-5 and B-cell epitopes-8) required for three-dimensional structure construction of TB vaccine construct based on a new concept introduced in this research. Population coverage analysis showed 99.16% of the world's population is covered by the predicted CTL and HTL epitopes which strengthens the confidence in our strategy used to develop a universal TB vaccine. This extensive study analysing strains reported in the last 30 years identified the broadest coverage of effective epitopes to date. Finally, evaluating the interaction between the pathogen and host TLR and the immune response profile of our constructed TB vaccine helped in predicting the vaccine's efficacy in generating a strong and specific humoral and cell-mediated immune response. The constructed epitope-based TB vaccine had a strong interaction inside the host, thus activating the macrophages, further leading to the production of B-cells, T-cells and cytokines and generating efficient cell-mediated and humoral immune responses.

Conclusion: The generation of a strong and swift immune response that can prevent disease progression and transmission is a prerequisite for a TB vaccine. An in-depth understanding of the interaction of TB bacteria with its host, the host defence mechanism, bacterial survival strategies in

evading the host immune response helped develop an *in-silico* epitope-based TB vaccine using computational vaccinology. The crucial features of the epitope-based TB vaccine constructed in this research include sequence conservancy, antigenicity, exclusion of self-peptides and multiple allelic interactions. The epitope-based TB vaccine is expected to be highly effective, safer and provide optimum protection against tuberculosis infection.

Goal 3: The third goal of this research is to develop an epitope-based vaccine and identify drug targets for the treatment of bovine TB (Chapter 5). Currently, no effective treatment is available for treating bovine tuberculosis, and animal slaughtering is usually undertaken to reduce the burden of bovine tuberculosis in the environment. In our study, we used 11 strains of *Mycobacterium bovis* to cover their diversity and identify the conserved proteins among them. 1163 proteins were discovered to be conserved among 11 strains of *Mycobacterium bovis*. For designing the bovine TB vaccine, we used the holistic framework developed in chapter 4. Out of 1163 conserved proteins, a total of nine conserved, membrane-spanning, antigenic and non-allergic proteins were selected from the reverse vaccinology approach. Twenty-six epitopes (HTL epitopes-2, CTL epitopes-8 and B-cell epitopes-16) that were highly immunogenic, non-toxic and non-allergenic were shortlisted after extensive immunoinformatic analysis. The shortlisted epitopes were used for designing the bovine TB vaccine sequence. The strong interaction between the bovine TB vaccine construct, and the host TLRs showed that the designed vaccine could initiate a robust immune response inside the host (cattle). Next, for identifying the potential anti-bovine TB drug targets, we developed a conceptual method and tested the framework using bioinformatics tools to provide a solution for bovine TB treatment. The emergence of antimicrobial resistance in bacteria has reduced the efficacy of antibiotics in treating the disease. A novel subtractive proteomic approach was developed to identify bovine TB drug targets. Using the 1163 conserved proteins, we performed a subtractive proteomics approach to identify drug targets that could further help investigate therapeutic drugs for the treatment of bovine TB. This approach helped identify nine drug targets that are conserved, essential, antigenic and have unique metabolic pathways in *Mycobacterium bovis*. Finally, the druggability potential of the shortlisted nine drug targets was predicted using DrugBank. The six proteins were termed druggable target proteins, and the remaining three proteins with no significant hits were considered novel target proteins.

Conclusion: We proposed nine therapeutic drug targets and an epitope-based vaccine for treating bovine TB in this research. The *in-silico* epitope-based bovine TB vaccine developed in this research would be far more effective than the current BCG vaccine for bovine TB. The nine highly potential drug targets identified in the study could facilitate the development of novel drugs against *Mycobacterium bovis.* The holistic strategy developed in this research for identifying drug targets and vaccine candidates is generic and can be used for other zoonotic infectious diseases.

6.2 Contributions

The first significant contribution of the study is the development of a conceptual framework for in-depth understanding of the impact of drug resistant mutations of *Mycobacterium tuberculosis* and investigate drug resistance mechanisms and bacterial survival strategy. Specific contributions include:

1) To our knowledge, this is the first study using systematic sorting and comprehensive *in silico* analysis of 821 non-synonymous mutations in first-line drug targets, based on five crucial factors - sequence conservation, distribution of mutations into three sites, function, structure stability and drug binding, to probe into drug resistance mechanisms and *Mycobacterium tuberculosis* strategies for survival.

2) Development of a reliable catalogue of drug-resistant mutations that can be used as a reference standard for validating mutations identified in the genome of new drug-resistant TB strains when available.

3) The more profound insight into *Mycobacterium tuberculosis* drug resistance mechanisms and survival strategies found in this study can significantly contribute to eradicating TB globally. This research can aid in developing novel diagnostic tools that can help in the early diagnosis of drug resistance TB, reducing the transmission rate and planning proper effective treatment for the patients. It can help develop effective and personalised treatment plans, develop new drugs and repurpose existing drugs for the frequent mutations worldwide.

4) The ranking of mutations into four different categories can assist in developing inhibitors for a specific mutation or group of mutations and help develop personalized treatment plans for TB patients.

The second significant contribution is the development of an *in-silico* epitope-based TB vaccine based on a deep understanding of host-pathogen interaction using computational vaccinology to provide a broad spectrum of protection against many *Mycobacterium tuberculosis* strains and drug resistance. The developed TB vaccine can significantly contribute to eliminating TB globally or drastically minimise its prevalence. Our research has addressed several challenges of TB vaccine development that include:

1) Expensive, time-consuming and arduous experimental testing in developing a vaccine.

2) Safety concerns while culturing the pathogen in a laboratory.

3) Identification of surface-exposed, secreted, and adhesin proteins; extract conserved epitopes in highly variable or drug-resistant *Mycobacterium tuberculosis*; identification of immunodominant epitopes for inducing potent humoral and cell-mediated immune responses.

4) Elimination of cross-reactive epitopes to avoid autoimmunity in the host.

The third significant contribution is the development of an epitope-based bovine TB vaccine to provide a broad spectrum of protection against many *Mycobacterium bovis* strains. The developed bovine TB vaccine can help achieve the long-term goal of prevention of bovine TB. The last contribution is the nine potential drug targets identified using new subtractive proteomic approach to answer pathogenicity and drug resistance against bovine TB. Our research can reduce the burden of significant loss of livestock and economic crisis due to bovine tuberculosis.

6.3 Future directions

The current research recommends the following future work:

1) An atlas created in this research for drug resistant mutations in the targets of first-line TB drugs can be updated when a new mutation is identified, or the genome of new drug-resistant TB strains is available.

2) The holistic frameworks developed in this research can study the nature and impact of mutations in other infectious diseases. Further, there is scope for introducing new steps in the method for improvement.

3) The conceptual method developed in this research to study mutations can be used to train new computational models to predict positions in proteins with a higher tendency to acquire new mutations and their consequences.

4) A laboratory trial in suitable *in vitro* and *in vivo* models is recommended for the proposed human TB and bovine TB vaccine candidates to validate our prediction's safety, efficacy, and immunogenicity.

5) The nine highly potential bovine TB drug targets identified in the study facilitates the design and development of novel drugs against *Mycobacterium bovis.*

We believe that the conceptual methods developed in this research and the results obtained by testing these methods would help accomplish the goal of reducing the burden of human and bovine tuberculosis globally.

Appendix A

A. 149 publications selected for studying the impact of drug resistance on *Mycobacterium tuberculosis*

Table A1: Published articles selected for study (149 studies on drug resistance with information on the country-of-origin isolate, drug used in a specific study, clinical isolates used in the research and resistant isolates for specific drug studied). Table A.1 includes:

1- **PMID**: PubMed ID of the selected study
2- **Author**: List of authors performed the selected study
3- **PY** (**p**ublication **y**ear): Year in which the selected study was published
4- **YSC** (**y**ear of **s**ample **c**ollection): Final year of the sample collection; NS- Not Specified
5- **Country**: the origin of the sample used in the study; Unknown- the origin of the sample not specified in the study
6- **Drug studied**: Drug used in the specific study; INH- Isoniazid; RIF- Rifampicin; PZA- Pyrazinamide; EMB- Ethambutol
7- **CIUS** (**c**linical **i**solates **u**sed in the **s**tudy): Number of clinical isolates used in the study for predicting first-line TB drug resistance
8- **RITG (r**esistant **i**solates in **t**arget **g**ene): number of isolates showing resistance to first-line TB drugs in 149 studies

Table A1: 149 publications selected for the study

PMID	Autor	PY	YSC	Country	Drug studied	CIUS	RITG
7840574	Williams D L et al.	1994	NS	Unknown	RIF	122	*rpoB*-110
8027320	Kapur V et al.	1994	1993	United States	RIF	128	*rpoB*-121
8585728	Rouse D A et al.	1995	NS	United States	INH	26	*katG*-20
8878604	Marttila H J et al.	1996	1989-95	Finland	INH	54	*katG*-13
9003625	Kim B J et al.	1997	NS	Korea	RIF	58	*rpoB*-32
9055989	Scorpio A et al.	1997	1992-95	United States	PZA	38	*pncA*-33
9056006	Sreevatsan S et al.	1997	NS	United States	PZA	118	*pncA*-67
9210694	Haas W H et al.	1997	1980-88	South Africa	INH	212	*katG*-124
9257740	Sreevatsan S et al.	1997	NS	Unknown	EMB	118	*embCAB*-69

9431913	Matsiota-Bernard P et al.	1998	1995-97	Greece	RIF	17	rpoB-17
9692180	Hirano K et al.	1997	NS	Unknown	PZA	168	pncA-135
10068603	Cingolani A et al.	1999	1994-97	Itlay	RIF, INH	71	KatG-17, rpoB- 8
10074552	Pozzi G et al.	1999	NS	Itlay	RIF	37	rpoB-37
10390239	Marttila H J et al.	1999	1994-97	Russia	PZA	44	pncA-36
10428945	Lee A S et al.	1999	1994-96	Singapore	INH	160	katG-160
10449496	Portugal I et al.	1999	1999	Portugal	RIF	99	rpoB-45
10471589	Mestdagh M et al.	1999	NS	Unknown	PZA	65	pncA-23
10565894	Yuen LK et al.	1999	1990-97	Australia	RIF	33	rpoB-33
10639358	Ramaswamy S V et al.	2000	NS	United States	EMB	75	embCAB-38
10645449	Escalante P et al.	1998	1995-96	Peru	RIF, INH, PZA, EMB	29	katG- 19, rpoB- 23, embB- 5, pncA- 8
10681313	Cheng S J et al.	2000	1990-92	Canada	PZA	57	pncA-53
10813147	Hou L et al.	2000	1996-98	China	PZA	65	rpoB, katG-35
10882091	Brown T J et al.	2000	1999	Turkey	RIF, PZA	10	rpoB, pncA-11
10921994	Valim A R et al.	2000	1996-98	Brazil	RIF	100	rpoB-38
11179917	Harris K A Jr et al.	2000	1995-96	India	RIF, INH	63	rpoB, katG-32
11474030	Mani C et al.	2001	NS	India	RIF	50	rpoB-44
11641519	Lee K W et al.	2001	NS	Korea	PZA	95	pncA-92
11796356	Siddiqi N et al.	2002	1995-98	India	RIF, INH	126	rpoB- 94, katG- 74
11854934	Lee H Y et al.	2002	NS	Korea	EMB	26	embB-21
12654653	Ramaswamy S V et al.	2003	NS	United States	INH	124	katG-38
12870687	El Baghdadi J et al.	2003	1996-2001	Morocco	RIF	122	rpoB-37
14598972	Agdamag D M et al.	2003	1999-2001	Philippines	RIF	164	rpoB-50
14638486	Morlock G P et al.	2003	NS	Unknown	INH	41	katG-15
15195248	Post F A et al.	2004	1999-2003	South Africa	RIF, INH, PZA, EMB	13	katG- 6, rpoB- 14, pncA- 5, embB- 1
15616332	Rodrigues Vde F et al.	2005	1998-2003	Brazil	PZA	59	pncA-40
15728936	Hillemann D et al.	2005	2001	Germany	RIF, INH	113	rpoB, katG-103
15917515	Parsons L M et al.	2005	1998-2004	United States	INH, EMB	157	katG, embB-157
16081898	Hillemann D et al.	2005	2001	Germany	RIF, INH	143	rpoB, katG-103
16189082	Lavender C et al.	2005	2001-03	Australia	INH	645	katG-52
16672384	Park H et al.	2006	NS	Korea	RIF, INH	243	rpoB, katG-119
16789833	Gagneux S et al.	2006	1990-99	United States	INH	2081	katG-152

16870753	Hazbón M H et al.	2006	1990-2002	Unknown	INH	1011	*katG*-402
16972132	Srivastava S et al.	2006	2004	India	EMB	23	*embCB*-23
17360809	Chan R C et al.	2007	1994-2004	Hongkong	RIF, INH, PZA, EMB	250	*katG*-8, *pncA*- 5, *rpoB*- 17, *embB*- 15
17539290	Khadka D K et al.	2007	NS	Thailand	INH	29	*katG*-29
17596354	Sekiguchi J et al.	2007	2007	Unknown	PZA	258	*pncA*-30
18272712	Strauss O J et al.	2008	2001-04	South Africa	RIF	429	*rpoB*-429
18294243	Prammananan T et al.	2008	2003-05	Thailand	RIF	267	*rpoB*-154
18508939	Abe C et al.	2008	NS	Japan	INH	96	*katG*-30
18573039	Perdigão J et al.	2008	2003	Portugal	RIF, INH, PZA	116	*rpoB, katG, pncA*-58
18701663	Akpaka P E et al.	2008	2006-07	Caribbean	RIF, INH	81	*rpoB*- 23, *katG*- 9
18753350	Mphahlele et al.	2008	2001-02	South Africa	PZA	71	*pncA*-42
19010731	Srivastava S et al.	2009	NS	India	EMB	44	*embCB*-44
19090721	Doustdar F et al.	2008	2005-06	Iran	INH	73	*katG*-48
19486070	Boonaiam S et al.	2010	2005-06	Thailand	INH	170	*katG*-160
19494067	Huang W L et al.	2009	2007-08	Taiwan	RIF, INH	272	*rpoB*-231, *katG*- 198
19520715	Hauck Y et al.	2009	NS	France	RIF	144	*rpoB*-119
19547874	Zenteno-Cuevas R et al.	2009	2007	Mexico	RIF, INH	25	*rpoB, katG*-20
19673965	Garza-González E et al.	2010	NS	Mexico	RIF	75	*rpoB*-33
19715569	Miotto P et al.	2009	2008	Itlay	RIF, INH	108	*rpoB, katG*-46
19800845	Plinke C et al.	2009	2002	Uzbekistan	EMB	197	*embB*-39
19832709	Ando H et al.	2010	2002	Japan	PZA	36	*pncA*-19
19900109	Pandey S et al.	2009	2000-06	New Zealand	PZA	26	*pncA*-11
20427375	Plinke C et al.	2010	2007	Germany	EMB	152	*embCAB*-34
20708497	Li G L et al.	2010	2007-08	China	RIF, INH, EMB	822	*rpoB*- 116, *katG*- 169, *embB*- 34
20727143	Jonmalung et al.	2010	2005-07	Thailand	PZA	150	*pncA*-52
21396209	Imperiale B R et al.	2011	2004-09	Argentina	RIF, INH	198	*katG*- 138, *rpoB*- 85
21554227	Bostanabad S Z et al.	2011	2007-08	Iran	INH	163	*katG*-42
21562102	Huang W L et al.	2011	2008-09	China	EMB	234	*emb*B-91
21911575	Ali A et al.	2011	2004-09	Pakistan	RIF, INH	50	*rpoB, katG*-50
22325117	Jin J et al.	2011	2006-09	China	RIF, INH	237	*rpoB, katG*-149
22325147	Tessema B et al.	2012	2009	Ethiopia	INH, RIF, EMB	260	*katG*- 35, *rpoB*- 15, *embB*- 8
22507192	Greif G et al.	2012	2000-05	Uruguay	INH	45	*katG*-20

22535987	Chakravorty S et al.	2012	NS	Korea	RIF	589	*rpoB*, *katG*-44
22646308	Feuerriegel S et al.	2012	2003-04	Sierra Leone	RIF, INH, PZA, EMB	97	*katG*- 32, *rpoB*- 16, *pncA*- 10, *embB*- 15
22708343	Madania A et al.	2012	2007-10	Syria	RIF, INH	160	*rpoB*- 69, *katG*- 72
22747769	Yadav R et al.	2012	2008-10	India	RIF, INH	61	*rpoB*- 29, *katG*- 35
22825123	Stoffels K et al.	2012	1994-2008	Belgium	PZA	138	*pncA*-60
22863574	Rahim Z et al.	2012	2001-10	Bangladesh	RIF, INH	218	*rpoB*, *katG*-218
22972833	Daum L T et al.	2012	2011	South Africa	RIF, INH, PZA	26	*rpoB*, *pncA*, *katG*-21
22984115	Sun G et al.	2012	NS	China	RIF, INH, PZA, EMB	7	*rpoB*, *katG*, *pncA*, *embB*-6
23146281	Poudel A et al.	2013	2007-10	Nepal	RIF, INH	109	*rpoB*, *katG*-13
23321280	Cuevas-Córdoba B et al.	2013	2007-10	Mexico	PZA	127	*pncA*-42
23453008	Escalante P et al.	2013	1997-2001	United States	INH	79	*katG*28
23467605	Maschmann Rde A et al.	2013	2009-11	Brazil	RIF, INH	68	*rpoB*, *katG*-32
23539241	Machado D et al.	2013	2009-11	Portugal	INH	17	*katG*-1
23561273	Jnawali H N et al.	2013	2009-10	Korea	RIF, INH, PZA, EMB	192	*rpoB*-159, *katG*- 67, *pncA*-107, *embB*- 127
23770140	Bhuju S et al.	2013	2003-05	Brazil	PZA	97	*pncA*-62
24326341	Yu X L et al.	2014	2009-10	China	RIF, INH	664	*rpoB*, *katG*-67
24670703	Huang Z K et al.	2014	2008-11	China	RIF, INH	208	*rpoB*- 9, *KatG*- 15
24855126	Jagielski T et al.	2014	2004	Poland	INH	50	*katG*-50
24867972	Aono A et al.	2014	NS	Japan	PZA	83	*pncA*-49
24884632	Tekwu E M et al.	2014	2010-11	Cameroon	RIF, INH, EMB	725	*rpoB*, *katG*, *embB*-63
25008819	Gupta A	2014	2008-10	India	RIF, INH	101	*katG*, *rpoB*-52
25093512	de Freitas F A et al.	2014	1995-2003	Brazil	RIF, INH	99	*rpoB*, *katG*-99
25182646	Cui Z et al.	2014	NS	China	EMB	767	*embB*-180
25186245	Bhembe N L et al.	2014	2012-13	South Africa	INH	190	*pncA*-140
25336456	Miotto P et al.	2014	NS	Unknown	PZA	1950	*pncA*-843
25427352	Htike Min P K et al.	2014	NS	Thailand	RIF	39	*rpoB*-36
25574916	Akhmetova A et al.	2015	2011	Kazakhstan	PZA	77	*pncA*-41
25605360	Zhao L L et al.	2015	2008-10	China	EMB	139	*embCAB*-79
25673793	Witney A A et al.	2015	2008-14	United Kingdom	RIF, INH, PZA, EMB	16	*rpoB*, *katG*, *pncA*, *embCAB*-16
25977398	Martinez E et al.	2015	2007-14	Australia	RIF, INH, PZA, EMB	15	*pncA*, *rpoB*, *katG*, *embB*-15

26033726	Brossier F et al.	2015	2009-14	France	RIF	131	*rpoB*-71
26117709	Thirumurugan R et al.	2015	2011-13	India	INH	127	*katG*-90
26124153	Cuevas-Córdoba B et al.	2015	2007-14	Mexico	EMB	110	*embB*-61
26369965	Rueda J et al.	2015	2000-13	Colombia	INH	64	*katG*-57
26786485	Qazi O et al.	2014	NS	Pakistan	RIF	1080	*rpoB*-24
26787207	Rosales-Klintz S et al	2012	1994-2009	Unknown	RIF, INH	120	*rpoB, katG*-117
26792466	Georghiou S B et al.	2016	2014	India	RIF, INH	79	*katG*-22, *rpoB*- 12
27096759	Ahmad S et al.	2016	2006-10	Kuwait	RIF, INH, EMB	70	*rpoB, katG, embB*-24
27530852	Aung H L et al.	2016	NS	Myanmar	RIF, INH, PZA, EMB	14	*rpoB*- 14, *katG*- 14, *pncA*- 6, *embB*- 8
27671062	Li Y et al.	2016	2014	China	EMB	280	*embB*-62
28053915	Desikan P et al.	2016	2012-13	India	RIF, INH	720	*rpoB, katG*-269
28404672	Georghiou S B et al.	2017	2012-13	Unknown	RIF, INH	451	*katG*- 203, *rpoB*- 158
28438132	Chen J et al.	2017	2014-15	China	RIF, INH	186	*rpoB, katG*-51
28627432	Shah Y et al.	2017	2008-13	Nepal	RIF, INH	601	*rpoB, katG*-47
28697808	Sengstake S et al.	2017	2012-13	Georgia	PZA	67	*pncA*-33
28716625	Sakhaee F et al	2017	2013-16	Iran	RIF, INH, EMB	395	*katG*- 24, *rpoB*- 21, *embB*- 32
28780247	Tam K K et al.	2017	2015-16	Hongkong	RIF, INH	187	*rpoB*-18, *katG*- 25
28839308	Bashir A et al.	2017	2015-16	Pakistan	INH	163	*katG*-79
28947708	Naidoo C C et al.	2017	NS	South Africa	RIF, INH, PZA, EMB	10	*katG, rpoB, pncA, embCAB*-10
29084750	Sun Q et al.	2017	NS	China	EMB	125	*embB*- 68, *embA*- 15, *embC*- 2
29110640	Pang Y et al.	2017	2014-16	China	PZA	133	*pncA*-83
29171560	Ahmad B et al.	2017	2015-16	Pakistan	RIF, INH, EMB	794	*rpoB*- 56, *embB*- 31, *pncA*- 16
29310751	Bainomugisa A et al.	2018	2012-15	Papua New Guinea	RIF, INH, PZA, EMB	100	*rpoB, katG, pncA, embB*-100
29486710	Oudghiri A et al.	2018	2010-12	Morocco	RIF, INH	703	*rpoB, katG*-90
29523326	Giri A et al.	2018	2011-15	India	EMB	360	*embCAB*-29
29559115	Bwalya P et al.	2018	2013-15	Zambia	PZA	131	*pncA*-32
29610587	Rezaei F et al.	2017	2010-14	Iran	EMB	20	*emb*-19
29847567	Kigozi E et al.	2018	NS	Uganda	RIF, INH	97	*katG*- 50, *rpoB*- 45
30061294	Jagielski T et al.	2018	2004-13	Poland	RIF	115	*rpoB*-65
30082293	Kandler J L et al.	2018	NS	United States	INH	52	*katG*-52

30130640	Roa M B et al.	2018	2009-11	Philippines	RIF, INH, PZA, EMB	10	*katG*-7, *rpoB*-6, *pncA*-2, *embCAB*- 6
30157763	Xu Y et al.	2018	2017-18	China	RIF, INH, PZA, EMB	18	*pncA, rpoB, katG, embB*- 18
30455227	Andres S et al.	2019	2016-17	Germany	PZA, EMB	85	*pncA*-49, *embCAB*- 42
30513056	Franco-Sotomayor G et al.	2018	2006-12	Ecuador	RIF, INH	152	*katG*- 28, *rpoB*- 35
30583880	Daum L T et al.	2019	2016	Ukraine	PZA	98	*pncA*-91
30881063	Khosravi A D et al.	2019	2016-17	Iran	EMB	307	*embB*-10
31086215	Luo D et al.	2019	2014-16	China	RIF, INH	1063	*rpoB, katG*- 157
141*	Marttila H A et al.	1998	1993-95	Russia	RIF, INH	27	katG- 26, rpoB- 22
142*	Ying X et al.	2011	2003-06	China	INH	100	*katG*-50
143*	Chiu Y c et al.	2011	2007-09	Taiwan	PZA	66	*pncA*-36
144*	Ramasubban et al.	2015	NS	India	RIF, INH	354	rpoB, katG- 18
145*	Zhang et al.	2018	2016	China	RIF, INH	671	rpoB-178, katG- 177
146*	Yoon J H et al.	2012	2005-08	Korea	RIF, INH, EMB	80	*rpoB*-41, *katG*- 52, *embB*- 45
147*	Xu Y et al.	2015	2009	China	EMB	1048	*embCAB*-109
148*	Jaiswal I et al.	2017	NS	India	INH	70	*katG*-50
149*	Bakula Z et al.	2013	2004	Poland	EMB	50	*embB*-17

'*'- PMID not known

141- A Ser315Thr substitution in *KatG* is predominant in genetically heterogeneous multidrug-resistant *Mycobacterium tuberculosis* isolates originating from the St. Petersburg area in Russia

142- An epidemiological study of resistant tuberculosis in Chongqing, China

143- Characteristics of *pncA* mutations in multidrug-resistant tuberculosis in Taiwan

144- Detection of novel and reported mutations in the *rpoB*, *katG* and *inhA* genes in multidrug-resistant tuberculosis isolates- A hospital-based study

145- GeneChip analysis of resistant *Mycobacterium tuberculosis* with previously treated tuberculosis in Changchun

146- Molecular characterization of drug-resistant and -susceptible *Mycobacterium tuberculosis* isolated from patients with tuberculosis in Korea

147- Mutations Found in *embCAB*, *embR*, and *ubiA* Genes of Ethambutol-Sensitive and -Resistant *Mycobacterium tuberculosis* Clinical Isolates from China

148- Mutations in *katG* and *inhA* genes of isoniazid-resistant and -sensitive clinical isolates of *Mycobacterium tuberculosis* from cases of pulmonary tuberculosis and their association with minimum inhibitory concentration of isoniazid

149- Mutations in the *embB* Gene and Their Association with Ethambutol Resistance in Multidrug-Resistant *Mycobacterium tuberculosis* Clinical Isolates from Poland

Appendix B

B. Atlas of first-line TB drug resistant mutations and single mutation frequency

Appendix B includes six tables. Tables B.1, B.2, B.3, B.4, B.5 and B.6 show the atlas of first-line drug resistant mutations in *Mycobacterium tuberculosis* and single mutation frequency for each mutation. Table B.1, B.2, B.3, B.4, B.5 and B.6 include:

1- Codon number or mutation position within a drug target

2- Amino acid substitution

3- Phenotypically resistant clinical isolates sequenced for a specific target or position

4- Mutation count

5- single amino acid mutation frequencies

Table B.1: Single mutation frequency for catalase-peroxidase (*katG*)

Table B.2: Single mutation frequency for Pyrazinamidase (*pncA*)

Table B.3: Single mutation frequency for β-subunit of RNA polymerase (*rpoB*)

Table B.4: Single mutation frequency for arabinosyltransferase A (*embA*)

Table B.5: Single mutation frequency for arabinosyltransferase B (*embB*)

Table B.6: Single mutation frequency for arabinosyltransferase C (*embC*)

Table B1: Single mutation frequency for catalase-peroxidase (*katG*)

Codon number	Amino acid Substitution	INH Resistant isolate	Mutation count	Single mutation frequency
315	S/T	5667	3433	60.5788
463	R/L	2400	330	13.75
315	S/N	5667	167	2.94689
315	S/R	5667	41	0.72349
315	S/I	5667	34	0.59996
337	Y/C	2884	20	0.69348

138	N/D	2088	20	0.95785
264	A/T	2158	19	0.88044
345	K/T	2884	18	0.62413
394	T/A	2326	18	0.77386
104	R/Q	2187	18	0.82305
106	A/V	2187	18	0.82305
107	W/R	2187	18	0.82305
109	A/V	2187	18	0.82305
229	Y/F	2123	18	0.84786
143	K/T	2088	18	0.86207
573	D/N	2088	18	0.86207
728	W/C	2088	18	0.86207
735	D/N	2088	18	0.86207
316	G/S	2957	16	0.54109
315	S/G	5667	9	0.15881
275	T/A	2257	8	0.35445
155	Y/C	2088	8	0.38314
280	P/H	2257	7	0.31015
85	T/P	2187	7	0.32007
251	T/M	2158	7	0.32437
121	G/C	2088	7	0.33525
162	A/T	2088	7	0.33525
607	E/K	2088	7	0.33525
322	T/A	3189	5	0.15679
326	T/M	2928	5	0.17077
141	L/F	2088	5	0.23946
315	S/D	5667	4	0.07058
317	I/L	2928	4	0.13661
300	W/G	2883	4	0.13874
309	G/V	2870	4	0.13937
110	A/V	2187	4	0.1829
328	W/C	3192	3	0.09398
305	G/A	2870	3	0.10453
309	G/S	2870	3	0.10453
311	D/F	2870	3	0.10453
406	D/A	2400	3	0.125
98	Y/S	2187	3	0.13717
172	A/V	2088	3	0.14368
328	W/R	3192	2	0.06266
322	T/N	3189	2	0.06272
295	Q/P	3008	2	0.06649
321	W/S	2928	2	0.06831
329	D/G	2884	2	0.06935
336	L/R	2884	2	0.06935
300	W/C	2883	2	0.06937
309	G/A	2870	2	0.06969
311	D/G	2870	2	0.06969
311	D/E	2870	2	0.06969

397	W/Y	2400	2	0.08333
409	A/D	2400	2	0.08333
428	G/R	2400	2	0.08333
463	R/H	2400	2	0.08333
248	I/M	2303	2	0.08684
270	H/A	2257	2	0.08861
271	T/V	2257	2	0.08861
275	T/P	2257	2	0.08861
279	G/D	2257	2	0.08861
101	L/P	2193	2	0.0912
94	D/N	2187	2	0.09145
261	E/K	2158	2	0.09268
261	E/Q	2158	2	0.09268
234	G/R	2123	2	0.09421
218	N/K	2103	2	0.0951
587	L/P	2101	2	0.09519
138	N/H	2088	2	0.09579
138	N/S	2088	2	0.09579
155	Y/S	2088	2	0.09579
172	A/T	2088	2	0.09579
315	S/A	5667	1	0.01765
315	S/L	5667	1	0.01765
328	W/L	3192	1	0.03133
328	W/S	3192	1	0.03133
322	T/M	3189	1	0.03136
316	G/D	2957	1	0.03382
303	S/W	2949	1	0.03391
317	I/V	2928	1	0.03415
318	E/V	2928	1	0.03415
318	E/G	2928	1	0.03415
321	W/R	2928	1	0.03415
321	W/L	2928	1	0.03415
321	W/G	2928	1	0.03415
324	T/P	2928	1	0.03415
308	T/P	2899	1	0.03449
312	A/V	2892	1	0.03458
312	A/R	2892	1	0.03458
312	A/G	2892	1	0.03458
314	T/N	2892	1	0.03458
299	G/C	2889	1	0.03461
331	S/C	2884	1	0.03467
335	I/T	2884	1	0.03467
335	I/V	2884	1	0.03467
336	L/P	2884	1	0.03467
341	W/S	2884	1	0.03467
344	T/P	2884	1	0.03467
350	A/S	2884	1	0.03467
357	D/H	2884	1	0.03467

357	D/N	2884	1	0.03467
379	A/V	2884	1	0.03467
380	T/I	2884	1	0.03467
384	L/R	2884	1	0.03467
385	R/P	2884	1	0.03467
300	W/R	2883	1	0.03469
302	S/R	2870	1	0.03484
307	G/R	2870	1	0.03484
309	G/C	2870	1	0.03484
309	G/F	2870	1	0.03484
311	D/Y	2870	1	0.03484
424	A/G	2560	1	0.03906
408	F/L	2400	1	0.04167
415	L/P	2400	1	0.04167
442	V/G	2400	1	0.04167
446	S/R	2400	1	0.04167
449	L/F	2400	1	0.04167
454	E/R	2400	1	0.04167
463	R/W	2400	1	0.04167
471	Q/R	2400	1	0.04167
485	G/V	2400	1	0.04167
490	G/C	2400	1	0.04167
490	G/D	2400	1	0.04167
496	R/L	2400	1	0.04167
498	R/H	2400	1	0.04167
491	G/C	2394	1	0.04177
388	P/S	2346	1	0.04263
388	P/L	2346	1	0.04263
274	K/R	2321	1	0.04308
505	W/S	2301	1	0.04346
11	T/A	2261	1	0.04423
12	T/P	2261	1	0.04423
35	N/D	2261	1	0.04423
61	A/T	2261	1	0.04423
63	D/E	2261	1	0.04423
68	V/G	2261	1	0.04423
74	D/Y	2261	1	0.04423
74	D/G	2261	1	0.04423
269	G/T	2257	1	0.04431
280	P/S	2257	1	0.04431
285	G/R	2257	1	0.04431
285	G/C	2257	1	0.04431
285	G/D	2257	1	0.04431
289	E/D	2257	1	0.04431
291	A/V	2257	1	0.04431
291	A/T	2257	1	0.04431
291	A/P	2257	1	0.04431
117	D/A	2239	1	0.04466

257	M/I	2195	1	0.04556
84	M/I	2187	1	0.04572
91	W/G	2187	1	0.04572
91	W/R	2187	1	0.04572
93	A/T	2187	1	0.04572
94	D/G	2187	1	0.04572
105	M/I	2187	1	0.04572
108	H/D	2187	1	0.04572
108	H/Q	2187	1	0.04572
118	G/A	2187	1	0.04572
515	R/C	2182	1	0.04583
525	Q/P	2175	1	0.04598
529	N/D	2162	1	0.04625
241	P/S	2158	1	0.04634
262	T/R	2158	1	0.04634
230	V/A	2123	1	0.0471
232	P/R	2123	1	0.0471
232	P/A	2123	1	0.0471
232	P/S	2123	1	0.0471
234	G/E	2123	1	0.0471
236	N/T	2123	1	0.0471
191	W/R	2103	1	0.04755
194	D/Y	2103	1	0.04755
195	E/K	2103	1	0.04755
206	G/D	2103	1	0.04755
217	E/G	2103	1	0.04755
700	S/P	2101	1	0.0476
121	G/V	2088	1	0.04789
125	G/V	2088	1	0.04789
125	G/C	2088	1	0.04789
127	Q/P	2088	1	0.04789
128	R/Q	2088	1	0.04789
128	R/P	2088	1	0.04789
131	P/R	2088	1	0.04789
131	P/Q	2088	1	0.04789
138	N/T	2088	1	0.04789
142	D/N	2088	1	0.04789
146	R/W	2088	1	0.04789
161	W/Q	2088	1	0.04789
161	W/R	2088	1	0.04789
169	G/S	2088	1	0.04789
176	M/I	2088	1	0.04789
186	G/V	2088	1	0.04789
573	D/G	2088	1	0.04789
611	L/R	2088	1	0.04789
629	G/S	2088	1	0.04789
636	A/E	2088	1	0.04789
652	S/A	2088	1	0.04789

Codon number	Amin acid substitution	PZA resistant isolate	Mutation count	Single mutation frequency
653	L/P	2088	1	0.04789
705	R/L	2088	1	0.04789
717	Q/P	2088	1	0.04789
735	D/A	2088	1	0.04789

Table B.2: Single mutation frequency for pyrazinamidase (*pncA*)

Codon number	Amin acid substitution	PZA resistant isolate	Mutation count	Single mutation frequency
10	Q/P	2608	38	1.457055
68	W/R	2608	27	1.035276
103	Y/D	2608	27	1.035276
141	Q/P	2608	25	0.958589
12	D/A	2608	23	0.881902
133	I/T	2608	23	0.881902
68	W/G	2608	22	0.843558
57	H/R	2608	21	0.805215
140	R/S	2608	21	0.805215
57	H/D	2608	20	0.766871
125	V/G	2608	20	0.766871
172	L/P	2608	20	0.766871
7	V/G	2608	18	0.690184
159	L/R	2608	17	0.65184
85	L/P	2608	16	0.613497
120	L/P	2608	16	0.613497
132	G/D	2608	16	0.613497
8	D/G	2608	15	0.575153
58	F/L	2608	15	0.575153
76	T/P	2608	15	0.575153
96	K/E	2608	15	0.575153
97	G/S	2608	15	0.575153
135	T/P	2608	15	0.575153
47	T/A	2608	14	0.53681
51	H/Q	2608	14	0.53681
14	C/R	2608	13	0.498466
54	P/L	2608	13	0.498466
10	Q/R	2608	12	0.460123
94	F/L	2608	12	0.460123
139	V/A	2608	12	0.460123
139	V/G	2608	11	0.421779
175	M/I	2608	11	0.421779
71	H/D	2608	10	0.383436
142	T/M	2608	10	0.383436
1	M/T	2608	9	0.345092
12	D/G	2608	9	0.345092
35	L/R	2608	9	0.345092
49	D/A	2608	9	0.345092
62	P/L	2608	9	0.345092

175	M/V	2608	9	0.345092
31	I/S	2608	8	0.306748
46	A/V	2608	8	0.306748
51	H/P	2608	8	0.306748
68	W/C	2608	8	0.306748
103	Y/H	2608	8	0.306748
155	V/G	2608	8	0.306748
8	D/E	2608	7	0.268405
8	D/N	2608	7	0.268405
12	D/E	2608	7	0.268405
51	H/R	2608	7	0.268405
94	F/C	2608	7	0.268405
97	G/D	2608	7	0.268405
102	A/V	2608	7	0.268405
130	V/G	2608	7	0.268405
132	G/S	2608	7	0.268405
139	V/L	2608	7	0.268405
142	T/K	2608	7	0.268405
146	A/T	2608	7	0.268405
63	D/G	2608	6	0.230061
72	C/R	2608	6	0.230061
82	H/R	2608	6	0.230061
134	A/V	2608	6	0.230061
140	R/P	2608	6	0.230061
171	A/E	2608	6	0.230061
177	T/P	2608	6	0.230061
43	H/P	2608	5	0.191718
49	D/G	2608	5	0.191718
51	H/Y	2608	5	0.191718
54	P/Q	2608	5	0.191718
57	H/Y	2608	5	0.191718
67	S/P	2608	5	0.191718
71	H/Y	2608	5	0.191718
94	F/S	2608	5	0.191718
104	S/R	2608	5	0.191718
136	D/Y	2608	5	0.191718
137	H/P	2608	5	0.191718
142	T/A	2608	5	0.191718
154	R/G	2608	5	0.191718
172	L/R	2608	5	0.191718
3	A/E	2608	4	0.153374
7	V/F	2608	4	0.153374
17	G/D	2608	4	0.153374
31	I/T	2608	4	0.153374
57	H/P	2608	4	0.153374
59	S/P	2608	4	0.153374
63	D/A	2608	4	0.153374
69	P/R	2608	4	0.153374

76	T/I	2608	4	0.153374
85	L/R	2608	4	0.153374
96	K/T	2608	4	0.153374
96	K/N	2608	4	0.153374
128	V/G	2608	4	0.153374
138	C/R	2608	4	0.153374
139	V/M	2608	4	0.153374
151	L/S	2608	4	0.153374
155	V/A	2608	4	0.153374
162	G/D	2608	4	0.153374
165	A/T	2608	4	0.153374
3	A/P	2608	3	0.115031
4	L/S	2608	3	0.115031
6	I/T	2608	3	0.115031
8	D/Y	2608	3	0.115031
9	V/G	2608	3	0.115031
10	Q/K	2608	3	0.115031
13	F/L	2608	3	0.115031
19	L/P	2608	3	0.115031
27	L/P	2608	3	0.115031
34	Y/D	2608	3	0.115031
68	W/L	2608	3	0.115031
69	P/L	2608	3	0.115031
96	K/Q	2608	3	0.115031
120	L/R	2608	3	0.115031
121	R/P	2608	3	0.115031
135	T/N	2608	3	0.115031
137	H/R	2608	3	0.115031
146	A/V	2608	3	0.115031
155	V/M	2608	3	0.115031
160	T/P	2608	3	0.115031
164	S/P	2608	3	0.115031
4	L/W	2608	2	0.076687
13	F/S	2608	2	0.076687
14	C/W	2608	2	0.076687
14	C/Y	2608	2	0.076687
26	A/G	2608	2	0.076687
34	Y/S	2608	2	0.076687
35	L/P	2608	2	0.076687
45	V/A	2608	2	0.076687
46	A/E	2608	2	0.076687
47	T/P	2608	2	0.076687
49	D/N	2608	2	0.076687
54	P/S	2608	2	0.076687
54	P/T	2608	2	0.076687
71	H/R	2608	2	0.076687
71	H/T	2608	2	0.076687
72	C/Y	2608	2	0.076687

72	C/W	2608	2	0.076687
78	G/D	2608	2	0.076687
82	H/D	2608	2	0.076687
87	T/M	2608	2	0.076687
94	F/P	2608	2	0.076687
96	K/R	2608	2	0.076687
97	G/A	2608	2	0.076687
100	T/A	2608	2	0.076687
100	T/P	2608	2	0.076687
103	Y/S	2608	2	0.076687
108	G/R	2608	2	0.076687
114	T/A	2608	2	0.076687
116	L/V	2608	2	0.076687
125	V/F	2608	2	0.076687
130	V/A	2608	2	0.076687
132	G/A	2608	2	0.076687
132	G/C	2608	2	0.076687
132	G/R	2608	2	0.076687
136	D/N	2608	2	0.076687
138	C/Y	2608	2	0.076687
148	R/S	2608	2	0.076687
160	T/A	2608	2	0.076687
161	A/P	2608	2	0.076687
171	A/V	2608	2	0.076687
172	L/A	2608	2	0.076687
180	V/G	2608	2	0.076687
180	V/F	2608	2	0.076687
182	L/S	2608	2	0.076687
1	M/I	2608	1	0.038344
3	A/S	2608	1	0.038344
3	A/Q	2608	1	0.038344
5	I/T	2608	1	0.038344
5	I/S	2608	1	0.038344
6	I/S	2608	1	0.038344
6	I/L	2608	1	0.038344
7	V/A	2608	1	0.038344
7	V/I	2608	1	0.038344
7	V/D	2608	1	0.038344
8	D/A	2608	1	0.038344
8	D/H	2608	1	0.038344
9	V/A	2608	1	0.038344
9	V/S	2608	1	0.038344
10	Q/H	2608	1	0.038344
12	D/H	2608	1	0.038344
12	D/N	2608	1	0.038344
13	F/V	2608	1	0.038344
17	G/S	2608	1	0.038344
18	S/P	2608	1	0.038344

19	L/R	2608	1	0.038344
21	V/G	2608	1	0.038344
23	G/A	2608	1	0.038344
23	G/V	2608	1	0.038344
25	A/E	2608	1	0.038344
27	L/R	2608	1	0.038344
28	A/V	2608	1	0.038344
28	A/D	2608	1	0.038344
41	Y/H	2608	1	0.038344
44	V/G	2608	1	0.038344
45	V/G	2608	1	0.038344
46	A/S	2608	1	0.038344
47	T/S	2608	1	0.038344
49	D/H	2608	1	0.038344
49	D/V	2608	1	0.038344
51	H/N	2608	1	0.038344
51	H/D	2608	1	0.038344
53	D/A	2608	1	0.038344
54	P/R	2608	1	0.038344
57	H/Q	2608	1	0.038344
57	H/L	2608	1	0.038344
59	S/F	2608	1	0.038344
62	P/H	2608	1	0.038344
62	P/T	2608	1	0.038344
62	P/Q	2608	1	0.038344
64	Y/D	2608	1	0.038344
66	S/P	2608	1	0.038344
68	W/D	2608	1	0.038344
68	W/S	2608	1	0.038344
69	P/A	2608	1	0.038344
71	H/Q	2608	1	0.038344
71	H/P	2608	1	0.038344
71	H/N	2608	1	0.038344
71	H/E	2608	1	0.038344
73	V/F	2608	1	0.038344
76	T/A	2608	1	0.038344
79	A/T	2608	1	0.038344
79	A/G	2608	1	0.038344
80	D/E	2608	1	0.038344
80	D/N	2608	1	0.038344
81	F/S	2608	1	0.038344
82	H/L	2608	1	0.038344
83	P/R	2608	1	0.038344
93	V/M	2608	1	0.038344
99	Y/D	2608	1	0.038344
102	A/T	2608	1	0.038344
103	Y/C	2608	1	0.038344
105	G/R	2608	1	0.038344

112	N/Y	2608	1	0.038344
114	T/P	2608	1	0.038344
116	L/R	2608	1	0.038344
118	N/T	2608	1	0.038344
119	W/L	2608	1	0.038344
119	W/G	2608	1	0.038344
119	W/R	2608	1	0.038344
119	W/C	2608	1	0.038344
125	V/D	2608	1	0.038344
125	V/L	2608	1	0.038344
130	V/L	2608	1	0.038344
131	V/F	2608	1	0.038344
134	A/S	2608	1	0.038344
135	T/A	2608	1	0.038344
136	D/H	2608	1	0.038344
136	D/G	2608	1	0.038344
137	H/D	2608	1	0.038344
138	C/T	2608	1	0.038344
138	C/W	2608	1	0.038344
138	C/S	2608	1	0.038344
140	R/H	2608	1	0.038344
142	T/P	2608	1	0.038344
143	A/P	2608	1	0.038344
143	A/G	2608	1	0.038344
143	A/T	2608	1	0.038344
146	A/P	2608	1	0.038344
146	A/E	2608	1	0.038344
148	R/C	2608	1	0.038344
154	R/T	2608	1	0.038344
155	V/L	2608	1	0.038344
157	V/G	2608	1	0.038344
159	L/P	2608	1	0.038344
160	T/K	2608	1	0.038344
162	G/A	2608	1	0.038344
163	V/A	2608	1	0.038344
168	T/P	2608	1	0.038344
168	T/N	2608	1	0.038344
171	A/T	2608	1	0.038344
171	A/P	2608	1	0.038344
175	M/T	2608	1	0.038344
175	M/R	2608	1	0.038344
184	C/Y	2608	1	0.038344

Table B.3: Single mutation frequency for β-subunit of RNA polymerase (*rpoB*)

Codon number	Amin acid substitution	RIF resistant isolate	Mutation count	Single mutation frequency
450	S/L	5143	2916	56.69843
445	H/Y	5143	412	8.010889
435	D/V	5143	334	6.494264
445	H/D	5143	233	4.53043
445	H/R	5143	160	3.111025
435	D/Y	5143	128	2.48882
452	L/P	5143	109	2.119386
450	S/W	5143	107	2.080498
445	H/L	5143	96	1.866615
441	S/L	5143	84	1.633288
430	L/P	5025	79	1.572139
435	D/G	5143	67	1.302742
445	H/N	5143	45	0.874976
445	H/C	5143	28	0.544429
432	Q/K	5025	27	0.537313
432	Q/P	5025	24	0.477612
445	H/P	5143	24	0.466654
432	Q/L	5025	19	0.378109
450	S/F	5143	17	0.330546
434	M/I	5143	16	0.311102
413	N/H	798	10	1.253133
435	D/E	5143	10	0.194439
450	S/Q	5143	9	0.174995
445	H/S	5143	8	0.155551
451	A/D	5143	8	0.155551
430	L/R	5025	7	0.139303
437	N/T	5143	7	0.136107
448	R/L	5143	7	0.136107
424	F/V	1899	6	0.315956
435	D/F	5143	6	0.116663
441	S/F	5143	6	0.116663
424	F /L	1899	5	0.263296
428	S/R	3978	5	0.125691
429	Q/H	3978	5	0.125691
435	D/H	5143	5	0.09722
441	S/Q	5143	5	0.09722
447	R/P	5143	5	0.09722
454	P/S	5095	5	0.098135
430	L/K	5025	4	0.079602
434	M/V	5143	4	0.077776
435	D/A	5143	4	0.077776
437	N/Y	5143	4	0.077776
437	N/H	5143	4	0.077776
438	N/K	5143	4	0.077776
444	T/P	5143	4	0.077776

445	H/Q	5143	4	0.077776
431	S/T	5025	3	0.059701
432	Q/H	5025	3	0.059701
437	N/I	5143	3	0.058332
441	S/N	5143	3	0.058332
441	S/W	5143	3	0.058332
445	H/T	5143	3	0.058332
445	H/G	5143	3	0.058332
454	P/H	5095	3	0.058881
491	I/F	981	3	0.30581
170	V/F	798	2	0.250627
426	G/D	2982	2	0.067069
428	S/T	3180	2	0.062893
431	S/C	5025	2	0.039801
431	S/I	5025	2	0.039801
431	S/R	5025	2	0.039801
431	S/M	5025	2	0.039801
432	Q/E	5025	2	0.039801
435	D/T	5143	2	0.038888
436	Q/L	5143	2	0.038888
437	N/D	5143	2	0.038888
442	G/W	5143	2	0.038888
445	H/A	5143	2	0.038888
446	K/N	5143	2	0.038888
446	K/Q	5143	2	0.038888
448	R/Q	5143	2	0.038888
450	S/A	5143	2	0.038888
452	L/E	5143	2	0.038888
452	L/V	5143	2	0.038888
455	G/D	5095	2	0.039254
460	E/G	1091	2	0.183318
480	I/V	1091	2	0.183318
483	P/L	981	2	0.203874
427	T/P	3133	1	0.031918
427	T/S	3133	1	0.031918
428	S/Q	3180	1	0.031447
428	S/G	3180	1	0.031447
428	S/I	3180	1	0.031447
430	L/V	5025	1	0.0199
430	L/M	5025	1	0.0199
431	S/N	5025	1	0.0199
431	S/G	5025	1	0.0199
433	F/L	5025	1	0.0199
433	F/V	5025	1	0.0199
434	M/T	5143	1	0.019444
435	D/N	5143	1	0.019444
435	D/P	5143	1	0.019444
435	D/K	5143	1	0.019444

436	Q/P	5143	1	0.019444
437	N/S	5143	1	0.019444
439	P/S	5143	1	0.019444
440	L/M	5143	1	0.019444
440	L/P	5143	1	0.019444
441	S/P	5143	1	0.019444
445	H/E	5143	1	0.019444
446	K/R	5143	1	0.019444
446	K/E	5143	1	0.019444
447	R/H	5143	1	0.019444
448	R/P	5143	1	0.019444
448	R/G	5143	1	0.019444
450	S/Y	5143	1	0.019444
450	S/G	5143	1	0.019444
450	S/C	5143	1	0.019444
452	L/R	5143	1	0.019444
452	L/M	5143	1	0.019444
453	G/A	5095	1	0.019627
453	G/W	5095	1	0.019627
453	G/V	5095	1	0.019627
457	L/R	1091	1	0.091659
481	E/G	1091	1	0.091659
482	T/P	1091	1	0.091659
487	N/S	981	1	0.101937
488	I/V	981	1	0.101937
493	S/L	821	1	0.121803
507	E/G	821	1	0.121803

Table B.4: Single mutation frequency for arabinosyltransferase A (*embA*)

Codon number	Amin acid substitution	EMB resistant isolate	Mutation count	Single mutation frequency
4	D/N	71	2	2.816901
5	G/S	71	2	2.816901
913	P/S	6	2	33.33333
5	G/V	71	1	1.408451
105	L/V	130	1	0.769231
122	V/G	130	1	0.769231
125	V/G	130	1	0.769231
200	G/S	115	1	0.869565
201	A/T	115	1	0.869565
206	V/M	115	1	0.869565
331	A/T	115	1	0.869565
343	V/L	115	1	0.869565
350	G/D	115	1	0.869565
380	R/P	115	1	0.869565
468	V/A	6	1	16.66667
554	G/D	6	1	16.66667

576	A/T	6	1	16.66667
639	P/S	6	1	16.66667
769	P/T	6	1	16.66667
838	P/L	6	1	16.66667

Table B.5: Single mutation frequency for arabinosyltransferase B (*embB*)

Codon number	Amin acid substitution	EMB resistant isolate	Mutation count	Single mutation frequency
306	M/V	1765	515	29.17847
306	M/I	1765	318	18.017
406	G/A	1592	75	4.711055
497	Q/R	1247	72	5.773857
306	M/L	1765	65	3.68272
406	G/D	1592	54	3.39196
497	Q/K	1247	30	2.405774
354	D/A	1395	24	1.72043
406	G/S	1592	22	1.38191
378	E/A	1390	15	1.079137
319	Y/C	1522	14	0.919842
497	Q/P	1247	11	0.882117
334	Y/H	1477	10	0.677048
328	D/G	1556	9	0.578406
281	A/V	1222	8	0.654664
296	N/H	1222	7	0.572831
406	G/P	1592	6	0.376884
328	D/Y	1556	6	0.385604
319	Y/S	1522	6	0.394218
330	F/L	1477	6	0.406229
406	G/C	1592	5	0.31407
246	G/R	1132	5	0.441696
504	E/D	1041	5	0.480307
360	V/A	1390	4	0.28777
404	P/S	1390	4	0.28777
299	D/E	1267	4	0.315706
507	R/K	1041	4	0.384246
1024	D/N	927	4	0.431499
319	Y/D	1522	3	0.197109
354	D/N	1395	3	0.215054
358	G/V	1390	3	0.215827
368	E/Q	1390	3	0.215827
397	P/T	1390	3	0.215827
402	L/V	1390	3	0.215827
448	G/V	1199	3	0.250209
460	R/C	1199	3	0.250209
347	S/C	1583	2	0.126342
328	D/H	1556	2	0.128535
356	A/F	1390	2	0.143885

359	L/I	1390	2	0.143885
368	E/D	1390	2	0.143885
374	G/V	1390	2	0.143885
377	V/G	1390	2	0.143885
380	S/R	1390	2	0.143885
384	Y/N	1390	2	0.143885
412	S/P	1281	2	0.156128
431	A/T	1281	2	0.156128
482	M/I	1202	2	0.166389
50	V/A	927	2	0.21575
74	L/R	927	2	0.21575
565	S/G	927	2	0.21575
1002	H/R	927	2	0.21575
306	M/F	1765	1	0.056657
306	M/T	1765	1	0.056657
406	G/K	1592	1	0.062814
406	G/R	1592	1	0.062814
347	S/T	1583	1	0.063171
345	D/G	1583	1	0.063171
347	S/I	1583	1	0.063171
328	D/V	1556	1	0.064267
309	V/A	1522	1	0.065703
309	V/G	1522	1	0.065703
310	A/R	1522	1	0.065703
311	D/R	1522	1	0.065703
311	D/F	1522	1	0.065703
311	D/G	1522	1	0.065703
311	D/H	1522	1	0.065703
312	H/R	1522	1	0.065703
315	Y/L	1522	1	0.065703
316	M/I	1522	1	0.065703
317	S/F	1522	1	0.065703
317	S/T	1522	1	0.065703
318	N/H	1522	1	0.065703
318	N/S	1522	1	0.065703
318	N/K	1522	1	0.065703
319	Y/N	1522	1	0.065703
320	F/L	1522	1	0.065703
322	W/C	1522	1	0.065703
322	W/R	1522	1	0.065703
330	F/I	1477	1	0.067705
330	F/V	1477	1	0.067705
331	G/R	1477	1	0.067705
332	W/R	1477	1	0.067705
354	D/T	1395	1	0.071685
356	A/S	1390	1	0.071942
356	A/V	1390	1	0.071942
357	A/S	1390	1	0.071942

357	A/T	1390	1	0.071942
357	A/V	1390	1	0.071942
360	V/M	1390	1	0.071942
366	S/L	1390	1	0.071942
366	S/P	1390	1	0.071942
367	R/P	1390	1	0.071942
368	E/A	1390	1	0.071942
369	V/L	1390	1	0.071942
369	V/A	1390	1	0.071942
370	L/R	1390	1	0.071942
371	P/R	1390	1	0.071942
375	P/A	1390	1	0.071942
377	V/M	1390	1	0.071942
377	V/E	1390	1	0.071942
378	E/K	1390	1	0.071942
379	A/T	1390	1	0.071942
379	A/D	1390	1	0.071942
380	S/N	1390	1	0.071942
380	S/G	1390	1	0.071942
380	S/D	1390	1	0.071942
388	A/G	1390	1	0.071942
393	T/A	1390	1	0.071942
395	W/R	1390	1	0.071942
395	W/C	1390	1	0.071942
397	P/R	1390	1	0.071942
397	P/Q	1390	1	0.071942
398	F/H	1390	1	0.071942
398	F/Y	1390	1	0.071942
399	N/T	1390	1	0.071942
399	N/I	1390	1	0.071942
399	N/D	1390	1	0.071942
399	N/H	1390	1	0.071942
400	N/P	1390	1	0.071942
400	N/K	1390	1	0.071942
401	G/S	1390	1	0.071942
404	P/A	1390	1	0.071942
405	E/D	1390	1	0.071942
405	E/P	1390	1	0.071942
412	S/L	1281	1	0.078064
423	M/T	1281	1	0.078064
426	S/N	1281	1	0.078064
430	P/L	1281	1	0.078064
435	V/G	1281	1	0.078064
436	V/G	1281	1	0.078064
437	T/A	1281	1	0.078064
298	S/A	1267	1	0.078927
298	S/W	1267	1	0.078927
304	L/V	1267	1	0.078927

497	Q/H	1247	1	0.080192
497	Q/F	1247	1	0.080192
281	A/S	1222	1	0.081833
282	V/G	1222	1	0.081833
288	L/V	1222	1	0.081833
293	I/T	1222	1	0.081833
296	N/I	1222	1	0.081833
296	N/K	1222	1	0.081833
297	S/A	1222	1	0.081833
465	I/D	1202	1	0.083195
469	R/P	1202	1	0.083195
471	R/P	1202	1	0.083195
446	P/H	1199	1	0.083403
450	I/M	1199	1	0.083403
452	V/L	1199	1	0.083403
454	A/T	1199	1	0.083403
459	G/A	1199	1	0.083403
460	R/L	1199	1	0.083403
461	P/S	1199	1	0.083403
239	L/P	1132	1	0.088339
240	D/H	1132	1	0.088339
257	R/W	1132	1	0.088339
507	R/G	1041	1	0.096061
505	A/V	1024	1	0.097656
13	N/S	927	1	0.107875
128	V/G	927	1	0.107875
557	M/I	927	1	0.107875
602	V/A	927	1	0.107875
624	N/D	927	1	0.107875
642	T/A	927	1	0.107875
643	T/I	927	1	0.107875
745	G/D	927	1	0.107875
1000	M/R	927	1	0.107875

Table B.6: Single mutation frequency for arabinosyltransferase C (*embC*)

Codon number	Amino acid substitution	EMB resistant isolate	Mutation count	Single mutation frequency
270	T/I	69	31	44.92754
981	V/L	69	30	43.47826
738	R/Q	69	5	7.246377
213	S/C	69	2	2.898551
251	L/R	69	2	2.898551
254	A/G	69	2	2.898551
297	I/L	69	2	2.898551
394	N/D	69	2	2.898551
150	P/S	69	1	1.449275
244	A/T	69	1	1.449275

247	A/P	69	1	1.449275
272	G/S	69	1	1.449275
285	H/Y	69	1	1.449275
287	V/F	69	1	1.449275
288	G/W	69	1	1.449275
288	G/V	69	1	1.449275
296	Y/H	69	1	1.449275
296	Y/S	69	1	1.449275
297	I/T	69	1	1.449275
300	M/R	69	1	1.449275
302	R/G	69	1	1.449275
303	V/G	69	1	1.449275
305	E/D	69	1	1.449275
307	A/T	69	1	1.449275
308	G/D	40	1	2.5
309	Y/N	40	1	2.5
310	M/K	40	1	2.5
325	G/S	40	1	2.5
326	W/R	40	1	2.5
327	Y/N	40	1	2.5
329	D/E	69	1	1.449275
378	A/V	69	1	1.449275
406	I/L	69	1	1.449275
426	A/T	69	1	1.449275
451	V/I	69	1	1.449275
707	P/L	69	1	1.449275
725	Q/R	69	1	1.449275
987	V/A	69	1	1.449275

Appendix C

C. Number of amino acid substitutions in the mutated position in the first-line drug targets

Appendix C includes six tables. Tables C.1, C.2, C.3, C.4, C.5 and C.6 show the number of amino acid substitutions (hotspot mutations) occurring at each mutated position of the first-line TB drug target in *Mycobacterium tuberculosis*. Table C.1, C.2, C.3, C.4, C.5 and C.6 include:

1- Codon number or mutation position within a drug target

2- Amino acid

3- Amino acid substitution occurring at a particular site

4- Number of different types of substitution

Table C1: Hotspot mutation count for catalase-peroxidase (*katG*)

Table C.2: Hotspot mutation count for pyrazinamidase (*pncA*)

Table C.3: hotspot mutation count for β-subunit of RNA polymerase (*rpoB*)

Table C.4: Hotspot mutation count for arabinosyltransferase A (*embA*)

Table C.5: Hotspot mutation count for arabinosyltransferase B (*embB*)

Table C.6: Hotspot mutation count for arabinosyltransferase C (*embC*)

Table C.1: Hotspot mutation count for catalase-peroxidase (*katG*)

katG Codon number	Amino acid at that position	Amino acid substitution	Number of different types of substitution
315	S	R, T, N, G, A, D, L, I,	8
309	G	C, S, V, A, F	5
138	N	H, D, S, T	4
311	D	Y, F, G, E	4
321	W	R, L, G, S	4
328	W	L, R, S, C	4
232	P	R, A, S	3

285	G	R, C, D	3
291	A	V, T, P	3
300	W	C, R, G	3
312	A	V, R, G	3
322	T	A, N, M	3
463	R	W, H, L	3
74	D	Y, G	2
91	W	G, R	2
94	D	N, G	2
108	H	D, Q	2
121	G	C, V	2
125	G	V, C	2
128	R	Q, P	2
131	P	R, Q	2
155	Y	S, C	2
161	W	Q, R	2
172	A	T, V	2
234	G	E, R	2
261	E	K, Q	2
275	T	A, P	2
280	P	S, H	2
316	G	S, D	2
317	I	L, V	2
318	E	V, G	2
335	I	T, V	2
336	L	R, P	2
357	D	H, N	2
388	P	S, L	2
490	G	C, D	2
573	D	G, N	2
735	D	N, A	2
11	T	A	1
12	T	P	1
35	N	D	1
61	A	T	1
63	D	E	1
68	V	G	1
84	M	I	1
85	T	P	1
93	A	T	1
98	Y	S	1
101	L	P	1
104	R	Q	1
105	M	I	1
106	A	V	1
107	W	R	1
109	A	V	1
110	A	V	1

117	D	A	1
118	G	A	1
127	Q	P	1
141	L	F	1
142	D	N	1
143	K	T	1
146	R	W	1
162	A	T	1
169	G	S	1
176	M	I	1
186	G	V	1
191	W	R	1
194	D	Y	1
195	E	K	1
206	G	D	1
217	E	G	1
218	N	K	1
229	Y	F	1
230	V	A	1
236	N	T	1
241	P	S	1
248	I	M	1
251	T	M	1
257	M	I	1
262	T	R	1
264	A	T	1
269	G	T	1
270	H	A	1
271	T	V	1
274	K	R	1
279	G	D	1
289	E	D	1
295	Q	P	1
299	G	C	1
302	S	R	1
303	S	W	1
305	G	A	1
307	G	R	1
308	T	P	1
314	T	N	1
324	T	P	1
326	T	M	1
329	D	G	1
331	S	C	1
337	Y	C	1
341	W	S	1
344	T	P	1
345	K	T	1

350	A	S	1
379	A	V	1
380	T	I	1
384	L	R	1
385	R	P	1
394	T	A	1
397	W	Y	1
406	D	A	1
408	F	L	1
409	A	D	1
415	L	P	1
424	A	G	1
428	G	R	1
442	V	G	1
446	S	R	1
449	L	F	1
454	E	R	1
471	Q	R	1
485	G	V	1
491	G	C	1
496	R	L	1
498	R	H	1
505	W	S	1
515	R	C	1
525	Q	P	1
529	N	D	1
587	L	P	1
607	E	K	1
611	L	R	1
629	G	S	1
636	A	E	1
652	S	A	1
653	L	P	1
700	S	P	1
705	R	L	1
717	Q	P	1
728	W	C	1

Table C.2: Hotspot mutation count for pyrazinamidase (*pncA*)

pncA Codon number	Amino acid at that position	Amino acid substitution	Number of different types of substitution
71	H	Q, Y, P, D, N, R, T, E	8
8	D	E, A, G, H, N, Y	6
51	H	N, Y, P, R, D, Q	6
57	H	D, Y, P, Q, R, L	6
68	W	R, D, C, G, S, L,	6

7	V	G, A, F, I, D	5
12	D	E, H, G, A, N	5
49	D	A, N, H, V, G	5
54	P	Q, S, R, T, L	5
96	K	R, E, Q, T, N	5
132	G	A, C, R, D, S	5
138	C	R, T, W, S, Y	5
3	A	E, S, Q, P	4
10	Q	H, P, K, R	4
62	P	H, T, Q, L	4
94	F	C, S, L, P	4
103	Y	D, H, S, C	4
119	W	L, G, R, C	4
125	V	D, F, L, G	4
136	D	Y, N, H, G	4
139	V	M, L, G, A	4
142	T	P, A, K, M	4
146	A	P, E, T, V	4
155	V	M, G, L, A	4
171	A	V, T, P, E	4
175	M	T, R, I, V	4
6	I	S, T, L	3
9	V	A, G, S	3
13	F	V, L, S	3
14	C	R, W, Y	3
46	A	E, S, V	3
47	T	A, P, S	3
69	P	A, R, L	3
72	C	R, Y, W	3
76	T	P, A, I	3
82	H	D, R, L	3
97	G	A, S, D	3
130	V	L, A, G	3
135	T	A, N, P	3
137	H	P, D, R	3
140	R	P, H, S	3
143	A	P, G, T,	3
160	T	K, A, P	3
172	L	P, R, A	3
1	M	I, T	2
4	L	W, S	2
5	I	T, S	2
17	G	S, D	2
19	L	P, R	2
23	G	A, V	2
27	L	R, P	2
28	A	V, D	2
31	I	T, S	2

34	Y	D, S	2
35	L	R, P	2
45	V	G, A	2
59	S	F, P	2
63	D	A, G	2
79	A	T, G	2
80	D	E, N	2
85	L	R, P	2
100	T	A, P	2
102	A	V, T	2
114	T	A, P	2
116	L	V, R	2
120	L	R, P	2
134	A	S, V	2
148	R	C, S	2
154	R	T, G	2
159	L	P, R	2
162	G	A, D	2
168	T	P, N	2
180	V	G, F	2
18	S	P	1
21	V	G	1
25	A	E	1
26	A	G	1
41	Y	H	1
43	H	P	1
44	V	G	1
53	D	A	1
58	F	L	1
64	Y	D	1
66	S	P	1
67	S	P	1
73	V	F	1
78	G	D	1
81	F	S	1
83	P	R	1
87	T	M	1
93	V	M	1
99	Y	D	1
104	S	R	1
105	G	R	1
108	G	R	1
112	N	Y	1
118	N	T	1
121	R	P	1
128	V	G	1
131	V	F	1
133	I	T	1

141	Q	P	1
151	L	S	1
157	V	G	1
161	A	P	1
163	V	A	1
164	S	P	1
165	A	T	1
177	T	P	1
182	L	S	1
184	C	Y	1

Table C.3: Hotspot mutation count for β-subunit of RNA polymerase (*rpoB*)

rpoB Codon Number	Amino acid present at that position	Amino acid substitution	Number of different types of substitution
445	H	Y, D, R, L, N, C, P, S, Q, T, G, A, E	13
435	D	V, Y, G, E, F, H, A, T, N, P, K	11
450	S	L, W, F, Q, A, Y, G, C	8
431	S	T, C, I, R, M, N, G	7
437	N	T, Y, H, I, D, S	6
441	S	L, F, Q, N, W, P	6
428	S	R, T, Q, G, I	5
430	L	P, R, K, V, M	5
432	Q	K, P, L, H, E	5
452	L	P, E, V, R, M	5
446	K	N, Q, R, E	4
448	R	L, Q, P, G	4
434	M	I, V, T	3
453	G	A, W, V	3
424	F	V, L	2
427	T	P, S	2
433	F	L, V	2
436	Q	L, P	2
440	L	M, P	2
447	R	P, H	2
454	P	S, H	2
170	V	F	1
413	N	H	1
426	G	D	1
429	Q	H	1
438	N	K	1
439	P	S	1
442	G	W	1
444	T	P	1
451	A	D	1
455	G	D	1
457	L	R	1

460	E	G	1
480	I	V	1
481	E	G	1
482	T	P	1
483	P	L	1
487	N	S	1
488	I	V	1
491	I	F	1
493	S	L	1
507	E	G	1

Table C.4: Hotspot mutation count for arabinosyltransferase A (*embA*)

embA Codon number	Amino acid at that position	Amino acid substitution	Number of types of different substitution
5	G	S, V	2
4	D	N	1
105	L	V	1
122	V	G	1
125	V	G	1
200	G	S	1
201	A	T	1
206	V	M	1
331	A	T	1
343	V	L	1
350	G	D	1
380	R	P	1
468	V	A	1
554	G	D	1
576	A	T	1
639	P	S	1
769	P	T	1
838	V	L	1
913	P	S	1

Table C.5: Hotspot mutation count for arabinosyltransferase B (*embB*)

embB Codon number	Amino acid at that position	Amino acid substitution	Number of different types of substitution
406	G	P, S, C, K, R, D, A	7
306	M	L, V, F, I, T	5
497	Q	P, H, F, R, K	5
311	D	R, F, G, H	4
319	Y	N, D, C, S	4
328	D	Y, G, H, V	4
380	S	R, N, G, D	4
399	N	T, I, D, H	4

296	N	H, I, K	3
318	N	H, S, K	3
330	F	I, L, V	3
347	S	T, C, I	3
354	D	N, T, A	3
356	A	F, S, V	3
357	A	S, T, V	3
368	E	Q, D, A	3
377	V	M, E, G	3
397	P	T, R, Q	3
281	A	S, V	2
298	S	A, W	2
309	V	A, G	2
317	S	F, T	2
322	W	C, R	2
360	V	A, M	2
366	S	L, P	2
369	V	L, A	2
378	E	A, K	2
379	A	T, D	2
395	W	R, C	2
398	F	H, Y	2
400	N	P, K	2
404	P	A, S	2
405	E	D, P	2
412	S	P	2
460	R	C, L	2
507	R	G, K	2
13	N	S	1
50	V	A	1
74	L	R	1
128	R	G	1
239	L	P	1
240	D	H	1
246	G	R	1
257	R	W	1
282	V	G	1
288	L	V	1
293	I	T	1
297	S	A	1
299	D	E	1
304	L	V	1
310	A	R	1
312	H	R	1
315	Y	L	1
316	M	I	1
320	F	L	1
331	G	R	1

332	W	R	1
334	Y	H	1
345	D	G	1
358	G	V	1
359	L	I	1
367	R	P	1
370	L	R	1
371	P	R	1
374	G	V	1
375	P	A	1
384	Y	N	1
388	A	G	1
393	T	A	1
401	G	S	1
402	L	V	1
423	M	T	1
426	S	N	1
430	P	L	1
431	A	T	1
435	V	G	1
436	V	G	1
437	T	A	1
446	P	H	1
448	G	V	1
450	I	M	1
452	V	L	1
454	A	T	1
459	G	A	1
461	P	S	1
465	I	D	1
469	R	P	1
471	R	P	1
482	M	I	1
504	E	D	1
505	A	V	1
557	M	I	1
565	S	G	1
602	V	A	1
624	N	D	1
642	T	A	1
643	T	I	1
745	G	D	1
1000	M	R	1
1002	H	R	1
1024	D	N	1

Table C.6: Hotspot mutation count Arabinosyltransferase C (*embC*)

embC Codon number	Amino acid	Substitution	Number of different types of substitution
288	G	W, V	2
296	Y	H, S	2
297	I	L, T	2
150	P	S	1
213	S	C	1
244	A	T	1
247	A	P	1
251	L	R	1
254	A	G	1
270	T	I	1
272	G	S	1
285	H	Y	1
287	V	F	1
300	M	R	1
302	R	G	1
303	V	G	1
305	E	D	1
307	A	T	1
308	G	D	1
309	Y	N	1
310	M	K	1
325	G	S	1
326	W	R	1
327	Y	N	1
329	D	E	1
378	A	V	1
394	N	D	1
406	I	L	1
426	A	T	1
451	V	I	1
707	P	L	1
725	Q	R	1
738	R	Q	1
981	V	L	1
987	V	A	1

Appendix D

D. Conservation score, domain region and functional site of mutations in the first-line TB drug targets

Appendix D includes six tables. Tables D.1, D.2, D.3, D.4, D.5 and D.6 summarise sequence conservation prediction using ConSurf. ConSurf grades conservation score from a scale of 1-9 where 1-3 are variable, 4-6 are average, and 7-9 represent highly conserved amino acid residue in a protein. The tables also provide information on the presence of the mutation in the protein domain region and distribution of mutated sites in site-1, 2 and 3 for each first-line TB drug target in *Mycobacterium tuberculosis.* Table D.1, D.2, D.3, D.4, D.5 and D.6 include:

1- Codon number or mutation position within a drug target
2- Amino acid
3- Consurf score
4- Presence of mutation in the domain region
5- Distribution of mutant site: site-1, 2 and 3

Table D1: Conservation score, domain region and functional site of mutations for catalase-peroxidase (*katG*)

Table D.2: Conservation score, domain region and functional site of mutations for pyrazinamidase (*pncA*)

Table D.3: Conservation score, domain region and functional site of mutations for β-subunit of RNA polymerase (*rpoB*)

Table D.4: Conservation score, domain region and functional site of mutations for arabinosyltransferase A (*embA*)

Table D.5: Conservation score, domain region and functional site of mutations for arabinosyltransferase B (*embB*)

Table D.6: Conservation score, domain region and functional site of mutations for arabinosyltransferase C (*embC*)

Table D.1: Conservation score, domain region and functional site of mutations for Catalase-peroxidase (*katG*)

Codon number	Amino acid	ConSurf Score	Domain	Mutant site
11	T	1	X	site-3
12	T	1	X	site-3
35	N	9	X	site-3
61	A	5	X	site-3
63	D	9	X	site-3
68	V	4	X	site-3
74	D	1	X	site-3
84	M	6	√	site-3
85	T	5	√	site-3
91	W	8	√	site-3
93	A	9	√	site-3
94	D	9	√	site-3
98	Y	8	√	site-3
101	L	4	√	site-1
104	R	9	√	site-1
105	M	9	√	site-2
106	A	6	√	site-3
107	W	8	√	site-1
108	H	9	√	site-2
109	A	9	√	site-3
110	A	9	√	site-3
117	D	9	√	site-3
118	G	8	√	site-3
121	G	8	√	site-3
125	G	8	√	site-3
127	Q	9	√	site-3
128	R	9	√	site-3
131	P	8	√	site-3
138	N	9	√	site-3
141	L	8	√	site-3
142	D	9	√	site-3
143	K	6	√	site-3
146	R	9	√	site-3
155	Y	4	√	site-3
161	W	8	√	site-3
162	A	9	√	site-3
169	G	8	√	site-3
172	A	9	√	site-2
176	M	9	√	site-2
186	G	8	√	site-3
191	W	8	√	site-3
194	D	6	√	site-3
195	E	5	√	site-3
206	G	3	√	site-3

217	E	6	√	site-3
218	N	6	√	site-3
229	Y	8	√	site-1
230	V	9	√	site-1
232	P	8	√	site-1
234	G	8	√	site-2
236	N	6	√	site-3
241	P	8	√	site-3
248	I	6	√	site-1
251	T	9	√	site-2
257	M	9	√	site-2
261	E	9	√	site-3
262	T	9	√	site-2
264	A	9	√	site-3
269	G	8	√	site-1
270	H	9	√	site-1
271	T	6	√	site-3
274	K	9	√	site-1
275	T	9	√	site-1
279	G	8	√	site-3
280	P	5	√	site-3
285	G	8	√	site-3
289	E	9	√	site-3
291	A	9	√	site-3
295	Q	6	√	site-3
299	G	8	√	site-3
300	W	8	√	site-3
302	S	6	√	site-3
303	S	3	√	site-3
305	G	8	√	site-3
307	G	8	√	site-3
308	T	1	√	site-3
309	G	8	√	site-3
311	D	9	√	site-3
312	A	9	√	site-3
314	T	5	√	site-1
315	S	1	√	site-1
316	G	8	√	site-2
317	I	4	√	site-1
318	E	9	√	site-2
321	W	8	√	site-1
322	T	9	√	site-3
324	T	9	√	site-3
326	T	9	√	site-3
328	W	8	√	site-2
329	D	5	√	site-3
331	S	1	√	site-3
335	I	1	√	site-3

336	L	8	√	site-3
337	Y	8	√	site-3
341	W	8	√	site-3
344	T	9	√	site-3
345	K	9	√	site-3
350	A	9	√	site-2
357	D	9	√	site-3
379	A	2	√	site-3
380	T	6	√	site-1
384	L	1	√	site-2
385	R	9	√	site-3
388	P	8	√	site-3
394	T	9	√	site-3
397	W	8	√	site-3
406	D	1	√	site-3
408	F	4	√	site-1
409	A	9	√	site-3
415	L	8	√	site-2
424	A	3	√	site-3
428	G	8	√	site-3
442	V	5	√	site-3
446	S	1	√	site-3
449	L	8	√	site-3
454	E	2	√	site-3
463	R	1	√	site-3
471	Q	9	√	site-3
485	G	1	√	site-3
490	G	8	√	site-3
491	G	8	√	site-3
496	R	9	√	site-3
498	R	9	√	site-3
505	W	8	√	site-3
515	R	1	√	site-3
525	Q	9	√	site-3
529	N	9	√	site-3
573	D	9	√	site-3
587	L	8	√	site-3
607	E	9	√	site-3
611	L	1	√	site-3
629	G	8	√	site-3
636	A	6	√	site-3
652	S	1	√	site-3
653	L	8	√	site-3
700	S	6	√	site-3
705	R	9	√	site-3
717	Q	1	X	site-3
728	W	8	X	site-3
735	D	9	X	site-3

Table D.2: Conservation score, domain region and functional site of mutations for pyrazinamidase (*pncA*)

Codon number	Amino acid	ConSurf Score	Domain	Mutant site
1	M	9	X	site-3
3	A	9	√	site-3
4	L	1	√	site-3
5	I	9	√	site-3
6	I	9	√	site-3
7	V	9	√	site-3
8	D	9	√	site-1
9	V	9	√	site-3
10	Q	9	√	site-3
12	D	9	√	site-3
13	F	8	√	site-1
14	C	7	√	site-3
17	G	7	√	site-3
18	S	9	√	site-3
19	L	8	√	site-2
21	V	9	√	site-2
23	G	7	√	site-3
25	A	9	√	site-3
26	A	9	√	site-3
27	L	8	√	site-3
28	A	9	√	site-3
31	I	9	√	site-3
34	Y	8	√	site-3
35	L	8	√	site-3
41	Y	8	√	site-3
43	H	9	√	site-3
44	V	9	√	site-3
45	V	9	√	site-3
46	A	9	√	site-3
47	T	9	√	site-2
49	D	9	√	site-1
51	H	9	√	site-1
53	D	9	√	site-3
54	P	8	√	site-2
57	H	1	√	site-1
58	F	8	√	site-2
59	S	9	√	site-3
62	P	8	√	site-3
63	D	9	√	site-3
64	Y	8	√	site-3
66	S	9	√	site-3
67	S	9	√	site-3

68	W	7	√	site-2
69	P	8	√	site-3
71	H	9	√	site-1
72	C	7	√	site-2
73	V	9	√	site-3
76	T	9	√	site-3
78	G	7	√	site-3
79	A	9	√	site-3
80	D	9	√	site-3
81	F	8	√	site-3
82	H	9	√	site-3
83	P	8	√	site-3
85	L	8	√	site-3
87	T	1	√	site-3
93	V	9	√	site-3
94	F	8	√	site-3
96	K	9	√	site-2
97	G	7	√	site-2
99	Y	8	√	site-3
100	T	9	√	site-3
102	A	9	√	site-1
103	Y	8	√	site-3
104	S	9	√	site-3
105	G	7	√	site-3
108	G	7	√	site-3
112	N	9	√	site-3
114	T	9	√	site-3
116	L	8	√	site-3
118	N	9	√	site-3
119	W	7	√	site-3
120	L	8	√	site-3
121	R	9	√	site-3
125	V	9	√	site-3
128	V	9	√	site-3
130	V	9	√	site-3
131	V	9	√	site-3
132	G	1	√	site-3
133	I	9	√	site-2
134	A	9	√	site-3
135	T	9	√	site-3
136	D	9	√	site-3
137	H	9	√	site-3
138	C	7	√	site-3
139	V	9	√	site-3
140	R	9	√	site-3
141	Q	9	√	site-3
142	T	1	√	site-3
143	A	9	√	site-3

146	A	9	√	site-3
148	R	9	√	site-3
151	L	8	√	site-3
154	R	1	√	site-3
155	V	1	√	site-3
157	V	1	√	site-3
159	L	1	√	site-3
160	T	1	√	site-3
161	A	1	√	site-3
162	G	7	√	site-3
163	V	1	√	site-3
164	S	1	√	site-3
165	A	1	√	site-3
168	T	1	√	site-3
171	A	1	√	site-3
172	L	1	√	site-3
175	M	1	√	site-3
177	T	1	√	site-3
180	V	1	√	site-3
182	L	8	√	site-3
184	C	1	√	site-3

Table D.3: Conservation score, domain region and functional site of mutations for β-subunit of RNA polymerase (*rpoB*)

Codon number	Amino acid	ConSurf Score	Domain	Mutant site
170	V	9	√	site-3
413	N	9	√	site-3
424	F	9	X	site-3
426	G	9	X	site-3
427	T	9	X	site-3
428	S	9	X	site-3
429	Q	9	X	site-3
430	L	9	X	site-3
431	S	9	X	site-3
432	Q	9	√	site-1
433	F	9	√	site-3
434	M	9	√	site-3
435	D	1	√	site-3
436	Q	9	√	site-3
437	N	9	√	site-3
438	N	9	√	site-3
439	P	9	√	site-3
440	L	9	√	site-3
441	S	9	√	site-3
442	G	9	√	site-3
444	T	9	√	site-3

445	H	9	√	site-2
446	K	9	√	site-3
447	R	9	√	site-1
448	R	9	√	site-3
450	S	1	√	site-3
451	A	9	√	site-3
452	L	1	√	site-3
453	G	9	√	site-3
454	P	9	√	site-3
455	G	9	√	site-3
457	L	9	√	site-3
460	E	9	√	site-3
480	I	9	√	site-3
481	E	9	√	site-3
482	T	9	√	site-3
483	P	9	√	site-1
487	N	1	√	site-1
488	I	9	√	site-3
491	I	9	√	site-1
493	S	9	√	site-3
507	E	9	X	site-3

Table D.4: Conservation score, domain region and functional site of mutations for arabinosyltransferase A (*embA*)

Codon number	Amino Acid	ConSurf Score	Domain	Mutant site
4	D	1	X	site-3
5	G	3	X	site-3
105	L	8	√	site-3
122	V	5	√	site-3
125	V	2	√	site-3
200	G	5	√	site-3
201	A	1	√	site-3
206	V	6	√	site-3
331	A	6	√	site-3
343	V	6	√	site-3
350	G	5	√	site-3
380	R	9	√	site-3
468	V	1	√	site-3
554	G	7	√	site-3
576	A	4	√	site-3
639	P	8	√	site-3
769	P	6	√	site-3
839	V	9	√	site-3
913	P	4	√	site-3

Table D.5: Conservation score, domain region and functional site of mutations for arabinosyltransferase B (*embB*)

Codon number	Amino acid	ConSurf Score	Domain	Mutant site
13	N	2	X	site-3
50	V	1	√	site-3
74	L	1	√	site-3
128	R	8	√	site-3
239	L	4	√	site-3
240	D	8	√	site-3
246	G	5	√	site-3
257	R	8	√	site-3
281	A	5	√	site-3
282	V	3	√	site-3
288	L	4	√	site-3
293	I	6	√	site-3
296	N	1	√	site-3
297	S	6	√	site-3
298	S	9	√	site-3
299	D	8	√	site-3
304	L	6	√	site-3
306	M	1	√	site-3
309	V	5	√	site-3
310	A	9	√	site-3
311	D	2	√	site-3
312	H	4	√	site-3
315	Y	8	√	site-3
316	M	6	√	site-3
317	S	3	√	site-3
318	N	9	√	site-3
319	Y	8	√	site-3
320	F	5	√	site-3
322	W	4	√	site-3
328	D	4	√	site-3
330	F	8	√	site-3
331	G	4	√	site-3
332	W	7	√	site-3
334	Y	3	√	site-3
345	D	4	√	site-3
347	S	5	√	site-3
354	D	4	√	site-3
356	A	3	√	site-3
357	A	9	√	site-3
358	G	2	√	site-3
359	L	5	√	site-3
360	V	5	√	site-3
366	S	9	√	site-3
367	R	8	√	site-3

368	E	2	√	site-3
369	V	8	√	site-3
370	L	8	√	site-3
371	P	8	√	site-3
374	G	7	√	site-3
375	P	4	√	site-3
377	V	8	√	site-3
378	E	1	√	site-3
379	A	4	√	site-3
380	S	6	√	site-3
384	Y	1	√	site-3
388	A	4	√	site-3
393	T	5	√	site-3
395	W	7	√	site-3
397	P	8	√	site-3
398	F	8	√	site-3
399	N	9	√	site-3
400	N	9	√	site-3
401	G	7	√	site-3
402	L	8	√	site-3
404	P	8	√	site-3
405	E	9	√	site-3
406	G	3	√	site-3
412	S	4	√	site-3
423	M	5	√	site-3
426	S	5	√	site-3
430	P	8	√	site-3
431	A	9	√	site-3
435	V	3	√	site-3
436	V	5	√	site-3
437	T	5	√	site-3
446	P	8	√	site-3
448	G	7	√	site-3
450	I	9	√	site-3
452	V	5	√	site-3
454	A	5	√	site-3
459	G	4	√	site-3
460	R	8	√	site-3
461	P	4	√	site-3
465	I	6	√	site-3
469	R	5	√	site-3
471	R	4	√	site-3
482	M	6	√	site-3
497	Q	9	√	site-3
504	E	9	√	site-3
505	A	6	√	site-3
507	R	8	√	site-3
557	M	4	√	site-3

Codon number	Amino acid	ConSurf Score	Domain	Mutant site
565	S	1	√	site-3
602	V	4	√	site-3
624	N	9	√	site-3
642	T	1	√	site-3
643	T	1	√	site-3
745	G	4	√	site-3
1000	M	5	√	site-3
1002	H	4	√	site-3
1024	D	4	√	site-3

Table D.6: Conservation score, domain region and functional site of mutations arabinosyltransferase C (*embC*)

Codon number	Amino acid	ConSurf Score	Domain	Mutant site
150	P	3	√	site-3
213	S	5	√	site-3
244	A	5	√	site-3
247	A	5	√	site-3
251	L	8	√	site-3
254	A	4	√	site-3
270	T	1	√	site-3
272	G	1	√	site-3
285	H	9	√	site-3
287	V	6	√	site-3
288	G	7	√	site-3
296	Y	8	√	site-3
297	I	5	√	site-3
300	M	5	√	site-3
302	R	8	√	site-3
303	V	8	√	site-3
305	E	4	√	site-3
307	A	5	√	site-3
308	G	7	√	site-3
309	Y	8	√	site-3
310	M	5	√	site-3
325	G	4	√	site-3
326	W	7	√	site-3
327	Y	8	√	site-3
329	D	5	√	site-3
378	A	2	√	site-3
394	N	9	√	site-3
406	I	6	√	site-3
426	A	9	√	site-3
451	V	5	√	site-3
707	P	8	√	site-3
725	Q	2	√	site-3
738	R	6	√	site-3

981	V	6	√	site-3
987	V	5	√	site-3

Appendix E

E. Results of prediction of functional change and structural stability change in *Mycobacterium tuberculosis* drug resistant mutations

Appendix E includes six tables. Tables E.1, E.2, E.3, E.4, E.5 and E. 6 summarise the prediction of functional change by Polyphen-2, PROVEAN and SIFT and structure stability change by I-MUTANT 3.0 and mCSM for each first-line TB drug target in *Mycobacterium tuberculosis.* Table E.1, E.2, E.3, E.4, E.5 and E.6 include:

1- Drug resistant mutation
2- PolyPhen-2: B- Benign, PSD- Possibly Damaging, PD- Probably Damaging
3- PROVEAN: N-Neutral, D- Deleterious
4- SIFT: T-Tolerated/Neutral, D- Deleterious (Affect Function)
5- I-MUTANT 3.0: Increase- structure stability increases, Decrease- Structure stability decreases
6- mCSM: ST-Stabilizing, HST- Highly Stabilizing, DT- Destabilizing, HDT- Highly Destabilizing

Table E.1: Functional change and structural stability change for catalase-peroxidase (*katG*)

Table E.2: Functional change and structural stability change for pyrazinamidase (*pncA*)

Table E.3: Functional change and structural stability change for β-subunit of RNA polymerase (*rpoB*)

Table E.4: Functional change and structural stability change for arabinosyltransferase A (*embA*)

Table E.5: Functional change and structural stability change for arabinosyltransferase B (*embB*)

Table E.6: Functional change and structural stability change for arabinosyltransferase C (*embC*)

Table E.1: Functional change and structural stability change for catalase-peroxidase (*katG*)

MUTATION	PolyPhen-2	PROVEAN	SIFT	I-MUTANT 3.0	mCSM
T11A	B	N	T	Decrease	-
T12P	B	N	T	Increase	-
N35D	PD	D	D	Increase	DT
A61T	B	N	T	Decrease	DT

D63E	B	D	T	Increase	DT
V68G	PD	D	D	Decrease	HDT
D74Y	B	D	T	Increase	DT
D74G	B	D	T	Increase	DT
M84I	PSD	N	D	Decrease	DT
T85P	PSD	D	D	Decrease	DT
W91G	PD	D	D	Decrease	HDT
W91R	PD	D	D	Decrease	DT
A93T	PD	D	D	Decrease	DT
D94N	PD	D	D	Decrease	DT
D94G	PD	D	D	Decrease	DT
Y98S	PD	D	D	Decrease	HDT
L101P	PD	D	D	Decrease	DT
R104Q	PD	D	D	Decrease	DT
M105I	PD	D	D	Decrease	DT
A106V	PD	D	D	Decrease	DT
W107R	PD	D	D	Decrease	HDT
H108D	PD	D	D	Decrease	DT
H108Q	PD	D	D	Decrease	DT
A109V	PSD	N	D	Decrease	DT
A110V	PD	D	D	Decrease	DT
D117A	PD	D	D	Increase	DT
G118A	PD	D	D	Increase	DT
G121C	PD	D	D	Decrease	DT
G121V	PD	D	D	Decrease	DT
G125V	PD	D	D	Increase	DT
G125C	PD	D	D	Decrease	DT
Q127P	PD	D	D	Decrease	ST
R128Q	PD	D	D	Decrease	DT
R128P	PD	D	D	Decrease	DT
P131R	PD	D	D	Decrease	DT
P131Q	PD	D	D	Decrease	DT
N138H	PD	D	D	Decrease	DT
N138D	PD	D	D	Decrease	HDT
N138S	PD	D	D	Decrease	HDT
N138T	PD	D	D	Decrease	DT
L141F	PD	N	D	Decrease	DT
D142N	PD	D	D	Decrease	DT
K143T	PD	D	D	Decrease	DT
R146W	PD	D	D	Decrease	DT
Y155S	PD	D	D	Decrease	HDT
Y155C	PD	D	D	Increase	DT
W161Q	PD	D	D	Decrease	HDT
W161R	PD	D	D	Decrease	HDT
A162T	PD	D	D	Decrease	DT
G169S	PD	D	D	Decrease	DT
A172T	PD	D	D	Decrease	HDT
A172V	PD	D	D	Increase	DT

M176I	PD	D	D	Decrease	DT
G186V	PD	D	D	Decrease	DT
W191R	PD	D	D	Decrease	DT
D194Y	PD	D	D	Increase	ST
E195K	B	D	D	Decrease	DT
G206D	B	D	N	Decrease	HDT
E217G	PSD	D	D	Decrease	DT
N218K	B	D	D	Decrease	DT
Y229F	PD	D	D	Decrease	DT
V230A	PD	D	D	Decrease	HDT
P232R	PD	D	D	Decrease	DT
P232A	PD	D	D	Decrease	DT
P232S	PD	D	D	Decrease	DT
G234E	PD	D	D	Decrease	HDT
G234R	PD	D	D	Decrease	DT
N236T	B	D	D	Increase	ST
P241S	PD	D	D	Decrease	HDT
I248M	PD	D	N	Decrease	DT
T251M	PD	D	D	Decrease	DT
M257I	PD	D	D	Decrease	DT
E261K	PD	D	D	Decrease	DT
E261Q	PD	D	D	Decrease	DT
T262R	PD	D	D	Decrease	DT
A264T	PD	D	D	Decrease	DT
G269T	PD	D	D	Decrease	ST
H270A	PD	D	D	Decrease	DT
T271V	PD	D	D	Increase	DT
K274R	PD	D	N	Decrease	DT
T275A	PD	D	D	Decrease	DT
T275P	B	D	N	Decrease	DT
G279D	B	D	N	Decrease	DT
P280S	B	D	N	Decrease	DT
P280H	B	D	D	Decrease	DT
G285R	PD	D	D	Increase	DT
G285C	PD	D	D	Decrease	DT
G285D	PD	D	D	Decrease	DT
E289D	B	D	N	Decrease	DT
A291V	PD	D	D	Increase	DT
A291T	PD	D	D	Decrease	DT
A291P	PD	D	D	Increase	DT
Q295P	PSD	D	D	Decrease	ST
G299C	PD	D	D	Decrease	DT
W300C	PD	D	D	Decrease	DT
W300R	PD	D	D	Decrease	DT
W300G	PD	D	D	Decrease	HDT
S302R	PD	D	D	Increase	DT
S303W	PD	D	D	Increase	DT
G305A	B	D	D	Decrease	DT

G307R	PD	D	D	Decrease	DT
T308P	B	N	N	Decrease	DT
G309C	PD	D	D	Decrease	DT
G309S	PD	D	D	Decrease	DT
G309V	PD	D	N	Decrease	ST
G309A	B	D	D	Decrease	DT
G309F	B	D	D	Decrease	DT
D311Y	PSD	D	D	Increase	DT
D311F	PD	D	D	Decrease	DT
D311G	PSD	D	D	Decrease	DT
D311E	B	D	D	Increase	DT
A312V	PSD	N	D	Increase	DT
A312R	PD	D	D	Decrease	DT
A312G	PSD	D	D	Decrease	DT
T314N	PD	D	N	Decrease	DT
S315R	PD	D	D	Increase	DT
S315T	PD	D	D	Increase	DT
S315N	PD	D	D	Increase	DT
S315G	PD	D	D	Decrease	DT
S315A	B	D	D	Decrease	DT
S315D	PD	D	D	Increase	ST
S315L	PSD	D	D	Increase	DT
S315I	PSD	D	D	Increase	DT
G316S	PD	D	D	Decrease	DT
G316D	PD	D	D	Decrease	HDT
I317L	B	N	N	Decrease	DT
I317V	B	N	D	Decrease	DT
E318V	PD	D	D	Increase	ST
E318G	PD	D	D	Decrease	DT
W321R	PD	D	D	Decrease	DT
W321L	PD	D	D	Decrease	HDT
W321G	PD	D	D	Decrease	HDT
W321S	PD	D	D	Decrease	HDT
T322A	PD	D	D	Decrease	DT
T322N	PSD	D	D	Decrease	DT
T322M	PD	D	D	Decrease	DT
T324P	PSD	D	D	Decrease	DT
T326M	PD	D	D	Decrease	ST
W328L	PD	D	D	Decrease	DT
W328R	PD	D	D	Decrease	HDT
W328S	PD	D	D	Decrease	HDT
W328C	PD	D	D	Decrease	DT
D329G	B	D	N	Decrease	DT
S331C	PD	D	D	Decrease	DT
I335T	B	N	N	Decrease	HDT
I335V	B	N	N	Decrease	DT
L336R	PD	D	D	Decrease	HDT
L336P	PD	D	D	Decrease	DT

Y337C	B	N	N	Decrease	DT
W341S	PD	D	D	Decrease	HDT
T344P	PSD	D	D	Decrease	DT
K345T	B	D	D	Decrease	DT
A350S	PD	D	D	Decrease	DT
D357H	B	D	D	Increase	ST
D357N	B	N	N	Increase	ST
A379V	B	N	N	Increase	DT
T380I	PD	D	D	Decrease	DT
L384R	PD	D	D	Decrease	HDT
R385P	PD	D	D	Decrease	DT
P388S	PD	D	D	Decrease	DT
P388L	PD	D	D	Decrease	DT
T394A	B	N	N	Decrease	DT
W397Y	B	N	N	Decrease	HDT
D406A	B	D	N	Increase	DT
F408L	PD	D	D	Decrease	DT
A409D	PD	D	D	Decrease	DT
L415P	PD	D	D	Decrease	DT
A424G	B	N	D	Decrease	DT
G428R	PD	D	D	Increase	DT
V442G	PD	D	D	Decrease	DT
S446R	B	N	N	Increase	DT
L449F	B	D	N	Decrease	DT
E454R	PSD	D	D	Decrease	ST
R463W	PD	N	D	Decrease	DT
R463H	PSD	N	D	Decrease	DT
R463L	B	N	N	Decrease	DT
Q471R	B	D	D	Decrease	DT
G485V	PSD	D	D	Decrease	DT
G490C	PD	D	D	Decrease	DT
G490D	PD	D	D	Decrease	DT
G491C	PD	D	D	Decrease	DT
R496L	PD	D	D	Decrease	DT
R498H	PD	D	D	Decrease	DT
W505S	PD	D	D	Decrease	HDT
R515C	PSD	N	D	Decrease	DT
Q525P	PSD	D	D	Decrease	ST
N529D	B	D	N	Decrease	DT
D573G	PD	D	D	Increase	DT
D573N	PD	D	D	Increase	DT
L587P	PSD	N	N	Decrease	DT
E607K	PD	D	D	Decrease	ST
L611R	PD	D	D	Decrease	HDT
G629S	PD	D	D	Decrease	DT
A636E	PD	D	D	Decrease	HDT
S652A	B	N	N	Decrease	DT
L653P	PD	D	D	Decrease	DT

S700P	PD	D	D	Increase	DT
R705L	PD	D	D	Decrease	DT
Q717P	B	D	D	Decrease	ST
W728C	PD	D	D	Decrease	DT
D735N	PD	D	D	Decrease	DT
D735A	PD	D	D	Decrease	DT

Table E.2: Functional change and structural stability change for pyrazinamidase (*pncA*)

MUTATION	PolyPhen-2	PROVEAN	SIFT	I-MUTANT 3.0	mCSM
M1I	PD	N	D	Decrease	DT
M1T	PD	N	D	Decrease	DT
A3E	PD	D	N	Decrease	HDT
A3S	PD	D	N	Decrease	HDT
A3Q	PD	D	N	Decrease	DT
A3P	PD	D	N	Decrease	DT
L4W	PD	D	D	Decrease	HDT
L4S	PD	D	D	Decrease	HDT
I5T	PD	D	D	Decrease	HDT
I5S	PD	D	D	Decrease	HDT
I6S	PD	D	D	Decrease	HDT
I6T	PD	D	D	Decrease	HDT
I6L	B	N	N	Decrease	DT
V7G	PD	D	D	Decrease	HDT
V7A	PD	D	D	Decrease	HDT
V7F	PD	D	D	Decrease	DT
V7I	PD	D	D	Decrease	DT
V7D	B	N	D	Decrease	HDT
D8E	PD	D	D	Increase	DT
D8A	PD	D	D	Decrease	DT
D8G	PD	D	D	Decrease	DT
D8H	PD	D	D	Decrease	DT
D8N	PD	D	D	Decrease	DT
D8Y	PD	D	D	Decrease	DT
V9A	PD	D	D	Decrease	HDT
V9G	PSD	D	D	Decrease	HDT
V9S	B	N	N	Decrease	HDT
Q10H	PD	D	D	Decrease	DT
Q10P	PD	D	D	Decrease	DT
Q10K	PD	D	D	Decrease	DT
Q10R	PD	D	D	Decrease	DT
D12E	PD	D	D	Increase	DT
D12H	PD	D	D	Decrease	DT
D12G	PD	D	D	Decrease	ST
D12A	PD	D	D	Decrease	ST
D12N	PD	D	D	Decrease	DT

F13V	PD	D	D	Decrease	HDT
F13L	PD	D	D	Decrease	DT
F13S	PD	D	D	Decrease	HDT
C14R	PD	D	D	Decrease	DT
C14W	PD	D	D	Decrease	DT
C14Y	PD	D	N	Decrease	DT
G17S	PD	D	N	Decrease	DT
G17D	PD	D	N	Decrease	DT
S18P	PD	D	N	Increase	DT
L19P	PD	D	D	Decrease	DT
L19R	PD	D	D	Decrease	DT
V21G	PD	D	D	Decrease	DT
G23A	PD	D	N	Decrease	DT
G23V	PD	D	N	Decrease	DT
A25E	B	D	N	Decrease	DT
A26G	B	N	N	Decrease	DT
L27R	PSD	D	D	Decrease	DT
L27P	B	D	D	Decrease	DT
A28V	PD	D	N	Decrease	DT
A28D	PD	D	D	Decrease	HDT
I31T	PD	D	D	Decrease	HDT
I31S	PD	D	D	Decrease	HDT
Y34D	B	D	N	Decrease	DT
Y34S	PSD	D	N	Decrease	HDT
L35R	PD	D	D	Decrease	DT
L35P	PSD	D	D	Decrease	DT
Y41H	B	D	N	Increase	DT
H43Y	B	D	N	Increase	ST
V44G	PD	D	D	Decrease	HDT
V45G	PD	D	D	Decrease	HDT
V45A	PSD	D	D	Decrease	HDT
A46E	PD	D	N	Decrease	DT
A46S	PD	D	N	Decrease	DT
A46V	PSD	D	N	Increase	ST
T47A	PD	D	D	Decrease	DT
T47P	PSD	D	N	Decrease	DT
T47S	B	D	N	Decrease	HDT
D49A	PD	D	D	Decrease	DT
D49N	PD	D	D	Decrease	HDT
D49H	PD	D	D	Decrease	DT
D49V	PD	D	D	Increase	DT
D49G	PD	D	D	Decrease	DT
H51N	PD	D	D	Decrease	HDT
H51Y	PD	D	D	Increase	ST
H51P	PD	D	D	Increase	DT
H51R	PD	D	D	Increase	DT
H51D	PD	D	D	Increase	HDT
H51Q	PD	D	D	Decrease	DT

D53A	B	D	D	Decrease	DT
P54Q	PD	D	D	Decrease	DT
P54S	PD	D	D	Decrease	HDT
P54R	PD	D	D	Decrease	DT
P54T	PD	D	D	Decrease	DT
P54L	PD	D	D	Decrease	DT
H57D	PD	D	D	Decrease	DT
H57Y	PD	D	D	Increase	ST
H57P	PD	D	D	Increase	DT
H57Q	PD	D	D	Decrease	DT
H57R	PD	D	D	Decrease	DT
H57L	PD	D	D	Increase	DT
F58L	B	D	N	Decrease	DT
S59F	PSD	D	D	Decrease	DT
S59P	B	D	D	Decrease	DT
P62H	PD	D	N	Decrease	DT
P62T	PD	D	N	Decrease	DT
P62Q	PD	D	N	Decrease	DT
P62L	PSD	D	N	Decrease	DT
D63A	B	D	N	Decrease	DT
D63G	B	D	N	Decrease	DT
Y64D	PSD	D	N	Decrease	DT
S66P	PSD	N	N	Decrease	DT
S67P	B	D	N	Increase	DT
W68R	PD	D	D	Decrease	DT
W68D	PD	D	D	Decrease	HDT
W68C	PD	D	D	Decrease	DT
W68G	PD	D	D	Decrease	HDT
W68S	PD	D	D	Decrease	HDT
W68L	PD	D	D	Decrease	HDT
P69A	PD	D	N	Decrease	DT
P69R	PD	D	N	Decrease	DT
P69L	PD	D	N	Decrease	DT
H71Q	PD	D	D	Decrease	HDT
H71Y	PD	D	D	Increase	DT
H71P	PD	D	D	Increase	DT
H71D	PD	D	D	Decrease	HDT
H71N	PD	D	D	Decrease	HDT
H71R	PD	D	D	Decrease	DT
H71T	PD	D	D	Decrease	HDT
H71E	PD	D	D	Increase	HDT
C72R	PD	D	D	Decrease	DT
C72Y	PD	D	D	Decrease	DT
C72W	PD	D	D	Decrease	DT
V73F	PD	D	D	Decrease	DT
T76P	PD	D	D	Increase	DT
T76A	PSD	D	N	Decrease	DT
T76I	PSD	D	D	Increase	DT

G78D	PD	D	D	Decrease	HDT
A79T	PD	D	N	Decrease	DT
A79G	B	D	N	Decrease	DT
D80E	B	N	N	Increase	DT
D80N	PSD	D	N	Decrease	DT
F81S	PSD	D	D	Decrease	HDT
H82D	B	D	N	Decrease	DT
H82R	B	D	N	Decrease	DT
H82L	B	D	N	Increase	ST
P83R	PD	D	N	Decrease	ST
L85R	PD	D	D	Decrease	DT
L85P	PD	D	D	Decrease	DT
T87M	B	D	N	Increase	DT
V93M	PD	D	D	Decrease	DT
F94C	PD	D	D	Decrease	DT
F94S	PD	D	N	Decrease	HDT
F94L	PD	D	D	Decrease	DT
F94P	B	D	N	Decrease	DT
K96R	PD	D	D	Decrease	DT
K96E	PD	D	D	Decrease	HDT
K96Q	PD	D	D	Decrease	DT
K96T	PD	D	D	Decrease	DT
K96N	PD	D	D	Decrease	DT
G97A	PD	D	D	Decrease	DT
G97S	PD	D	D	Decrease	DT
G97D	PD	D	D	Decrease	HDT
Y99D	PD	D	D	Decrease	ST
T100A	B	N	N	Decrease	DT
T100P	B	N	N	Decrease	DT
A102V	B	D	N	Increase	DT
A102T	PSD	D	N	Decrease	DT
Y103D	PD	D	D	Decrease	ST
Y103H	PD	D	D	Decrease	ST
Y103S	PSD	D	D	Decrease	ST
Y103C	PSD	D	D	Decrease	ST
S104R	PD	D	D	Increase	DT
G105R	PD	D	D	Decrease	DT
G108R	PD	D	N	Decrease	DT
N112Y	PSD	D	N	Increase	DT
T114A	B	D	N	Decrease	DT
T114P	B	D	D	Decrease	DT
L116V	PD	D	D	Decrease	DT
L116R	PD	D	N	Decrease	DT
N118T	B	N	N	Decrease	DT
W119L	PD	D	N	Decrease	HDT
W119G	PD	D	N	Decrease	HDT
W119R	PSD	D	N	Decrease	HDT
W119C	B	D	N	Decrease	HDT

L120R	PD	D	D	Decrease	HDT
L120P	PD	D	D	Decrease	DT
R121W	PD	D	D	Decrease	DT
V125D	PD	D	D	Decrease	HDT
V125F	PD	D	N	Decrease	DT
V125L	B	N	N	Decrease	DT
V125G	PD	D	D	Decrease	HDT
V128G	PD	D	D	Decrease	HDT
V130L	PD	D	N	Decrease	DT
V130A	PD	D	D	Decrease	HDT
V130G	PD	D	D	Decrease	HDT
V131F	PD	D	N	Decrease	DT
G132A	PD	D	D	Decrease	DT
G132C	PD	D	D	Decrease	DT
G132R	PD	D	D	Decrease	DT
G132D	PD	D	D	Decrease	HDT
G132S	PD	D	D	Decrease	DT
I133T	PD	D	D	Decrease	HDT
A134S	PD	D	D	Decrease	DT
A134V	PD	D	D	Decrease	DT
T135A	PD	D	N	Decrease	DT
T135N	PD	D	N	Increase	DT
T135P	B	D	D	Decrease	ST
D136Y	PD	D	D	Increase	DT
D136N	PD	D	D	Increase	ST
D136H	PD	D	D	Decrease	DT
D136G	PD	D	D	Increase	ST
H137P	PSD	D	N	Increase	ST
H137D	PD	D	N	Decrease	ST
H137R	PSD	D	N	Decrease	DT
C138R	PD	D	D	Increase	DT
C138T	PD	D	D	Increase	ST
C138W	PD	D	D	Increase	DT
C138S	PD	D	D	Decrease	DT
C138Y	PD	D	D	Decrease	DT
V139M	PD	D	D	Decrease	DT
V139L	PD	D	D	Decrease	DT
V139G	PD	D	D	Decrease	HDT
V139A	PD	D	D	Decrease	DT
R140P	PSD	D	D	Decrease	DT
R140H	PD	D	D	Decrease	DT
R140S	B	D	D	Decrease	DT
Q141P	PSD	D	N	Decrease	ST
T142P	PD	D	D	Decrease	DT
T142A	PD	D	D	Decrease	DT
T142K	PD	D	D	Decrease	DT
T142M	PD	D	D	Decrease	HDT
A143P	PD	D	N	Decrease	DT

A143G	PD	D	N	Decrease	DT
A143T	PD	D	N	Decrease	HDT
A146P	PD	D	D	Increase	DT
A146E	PD	D	D	Decrease	DT
A146T	PD	D	D	Decrease	DT
A146V	PD	D	D	Decrease	DT
R148C	PD	D	D	Decrease	DT
R148S	B	N	N	Decrease	DT
L151S	PSD	D	N	Decrease	HDT
R154T	B	D	N	Decrease	DT
R154G	B	D	N	Decrease	DT
V155M	PD	D	D	Decrease	DT
V155G	PD	D	D	Decrease	HDT
V155L	B	D	N	Decrease	DT
V155A	PD	D	D	Decrease	HDT
V157G	B	D	N	Decrease	DT
L159P	PD	D	N	Decrease	DT
L159R	PD	D	N	Decrease	DT
T160K	PD	D	N	Decrease	DT
T160A	B	D	N	Decrease	DT
T160P	PD	D	N	Decrease	DT
A161P	B	D	N	Increase	DT
G162A	B	D	N	Decrease	DT
G162D	PD	D	N	Decrease	DT
V163A	B	D	N	Decrease	DT
S164P	PSD	D	N	Increase	DT
A165T	B	N	N	Decrease	DT
T168P	PD	D	N	Decrease	ST
T168N	PSD	D	N	Increase	DT
A171V	PD	D	N	Decrease	DT
A171T	PD	D	N	Decrease	DT
A171P	PD	D	N	Increase	DT
A171E	PD	D	N	Decrease	HDT
L172P	PD	D	N	Decrease	DT
L172R	PSD	D	N	Decrease	DT
L172A	PSD	D	N	Decrease	HDT
M175T	PD	D	D	Decrease	DT
M175R	PD	D	D	Decrease	DT
M175I	PSD	N	N	Decrease	DT
M175V	PSD	N	N	Decrease	DT
T177P	B	N	N	Decrease	DT
V180G	PD	D	D	Decrease	HDT
V180F	PSD	D	N	Decrease	DT
L182S	PD	D	D	Decrease	HDT
C184Y	B	N	D	Decrease	DT

Table E.3: Functional change and structural stability change for β-subunit of RNA polymerase (*rpoB*)

MUTATION	PolyPhen-2	PROVEAN	SIFT	I-MUTANT 3.0	mCSM
V170F	B	N	D	Decrease	DT
N413H	PD	D	D	Decrease	DT
F424V	PD	D	D	Decrease	DT
F 424L	PD	D	D	Decrease	DT
G426D	PD	D	D	Decrease	ST
T427P	PD	D	D	Decrease	DT
T427S	PD	D	N	Decrease	DT
S428R	PD	D	D	Increase	DT
S428Q	PD	D	D	Increase	DT
S428G	PD	D	N	Decrease	DT
S428T	PD	D	D	Decrease	DT
S428I	PD	D	D	Increase	DT
Q429H	PD	D	D	Decrease	DT
L430P	PD	D	D	Decrease	DT
L430V	PD	D	D	Decrease	DT
L430R	PD	D	D	Decrease	DT
L430M	PD	N	D	Decrease	DT
L430K	PD	D	D	Decrease	DT
S431C	PD	D	D	Decrease	DT
S431T	PD	D	D	Decrease	DT
S431I	PD	D	D	Increase	DT
S431N	PD	D	D	Increase	DT
S431R	PD	D	D	Increase	DT
S431G	PD	D	D	Decrease	DT
S431M	PD	D	D	Increase	DT
Q432L	B	D	D	Increase	ST
Q432K	PD	D	D	Increase	DT
Q432E	PD	D	D	Increase	DT
Q432P	PD	D	D	Decrease	DT
Q432H	PD	D	D	Decrease	DT
F433L	PSD	D	D	Decrease	DT
F433V	PD	D	D	Decrease	DT
M434V	PSD	D	D	Decrease	DT
M434T	B	N	N	Decrease	DT
M434I	B	N	N	Decrease	DT
D435Y	PD	D	D	Increase	DT
D435G	PD	D	D	Decrease	ST
D435V	PD	D	D	Increase	ST
D435N	PD	D	D	Decrease	DT
D435H	PD	D	D	Decrease	DT
D435E	PSD	D	D	Increase	DT
D435A	PD	D	D	Decrease	ST
D435P	PD	D	D	Decrease	ST
D435K	PD	D	D	Increase	DT

D435T	PD	D	D	Decrease	DT
D435F	PD	D	D	Increase	DT
Q436P	PD	D	D	Increase	ST
Q436L	PD	D	D	Increase	HST
N437Y	PD	D	D	Increase	ST
N437T	B	D	N	Increase	ST
N437S	PSD	D	D	Decrease	ST
N437I	PD	D	D	Increase	ST
N437H	PD	D	D	Decrease	ST
N437D	PD	D	N	Decrease	HDT
N438K	PD	D	D	Decrease	DT
P439S	PD	D	D	Decrease	HDT
L440M	PD	N	D	Decrease	DT
L440P	PD	D	D	Decrease	DT
S441L	PD	D	D	Increase	ST
S441F	PD	D	D	Increase	DT
S441Q	PD	D	D	Decrease	DT
S441P	PD	D	D	Increase	ST
S441N	PD	D	D	Increase	DT
S441W	PD	D	D	Increase	DT
G442W	PD	D	D	Decrease	DT
T444P	PD	D	D	Decrease	DT
H445D	PSD	D	D	Increase	DT
H445C	PD	D	D	Increase	DT
H445L	PD	D	D	Increase	DT
H445N	PD	D	D	Decrease	DT
H445Y	PD	D	D	Increase	ST
H445T	PD	D	D	Decrease	DT
H445S	PD	D	D	Decrease	DT
H445G	PD	D	D	Decrease	DT
H445A	PD	D	D	Decrease	DT
H445R	PD	D	D	Decrease	DT
H445E	PD	D	D	Increase	DT
H445P	PD	D	D	Increase	DT
H445Q	PD	D	D	Decrease	DT
K446N	PD	D	D	Decrease	HDT
K446R	PD	D	D	Decrease	DT
K446E	PD	D	D	Decrease	DT
K446Q	PD	D	D	Decrease	DT
R447H	PD	D	D	Decrease	HDT
R447P	PD	D	D	Decrease	DT
R448P	PD	D	D	Decrease	DT
R448Q	PD	D	D	Decrease	DT
R448G	PD	D	D	Decrease	DT
R448L	PD	D	D	Decrease	DT
S450Q	PD	D	D	Decrease	DT
S450W	PD	D	D	Increase	DT
S450L	PD	D	D	Increase	DT

S450A	PD	D	D	Decrease	DT
S450Y	PD	D	D	Increase	DT
S450F	PD	D	D	Increase	DT
S450G	PD	D	D	Decrease	DT
S450C	PD	D	D	Decrease	DT
A451D	PD	D	D	Increase	DT
L452P	PD	D	D	Decrease	DT
L452R	PD	D	D	Decrease	DT
L452E	PD	D	D	Decrease	DT
L452V	PD	D	D	Decrease	DT
L452M	PD	N	D	Decrease	DT
G453A	PD	D	D	Decrease	DT
G453W	PD	D	D	Decrease	DT
G453V	PD	D	D	Decrease	DT
P454S	PSD	D	D	Decrease	DT
P454H	PD	D	D	Decrease	DT
G455D	PD	D	D	Decrease	DT
L457R	PD	D	D	Decrease	DT
E460G	PD	D	D	Decrease	DT
I480V	B	N	N	Decrease	DT
E481G	PD	D	D	Decrease	ST
T482P	PD	D	D	Decrease	DT
P483L	PD	D	D	Decrease	DT
N487S	PD	D	N	Increase	DT
I488V	PD	N	D	Decrease	DT
I491F	PD	D	D	Decrease	DT
S493L	PD	D	D	Decrease	ST
E507G	PD	D	D	Decrease	DT

Table E.4: Functional change and structural stability change for arabinosyltransferase A (*embA*)

MUTATION	PolyPhen-2	PROVEAN	SIFT	I-MUTANT 3.0
D4N	B	N	N	Decrease
G5S	B	N	N	Decrease
G5V	PSD	N	D	Decrease
L105V	PSD	D	D	Decrease
V122G	PD	D	N	Decrease
V125G	PSD	N	N	Decrease
G200S	B	N	N	Decrease
A201T	B	N	N	Decrease
V206M	PD	N	D	Decrease
A331T	PD	D	N	Decrease
V343L	B	N	D	Decrease
G350D	B	N	D	Decrease
R380P	PD	D	D	Decrease
V468A	B	D	N	Decrease

G554D	PD	D	D	Decrease
A576T	PSD	N	N	Decrease
P639S	PD	D	D	Decrease
P769T	PD	D	N	Decrease
P838L	PD	D	D	Decrease
P913S	PSD	D	N	Decrease

Table E.5: Functional change and structural stability change for arabinosyltransferase B (*embB*)

MUTATION	PolyPhen-2	PROVEAN	SIFT	I-MUTANT 3.0
N13S	B	N	N	Decrease
V50A	B	N	N	Decrease
L74R	PSD	N	N	Decrease
R128G	PSD	D	D	Decrease
L239P	PD	D	D	Decrease
D240H	PD	D	D	Decrease
G246R	B	N	N	Increase
R257W	PSD	N	D	Decrease
A281S	B	N	N	Decrease
A281V	B	N	N	Increase
V282G	B	D	N	Decrease
L288V	B	N	N	Decrease
I293T	PSD	D	D	Decrease
N296H	PD	D	N	Decrease
N296I	PD	D	D	Increase
N296K	PD	D	N	Decrease
S297A	PSD	D	D	Decrease
S298A	PD	N	N	Decrease
S298W	PD	D	D	Increase
D299E	PD	D	D	Increase
L304V	PSD	N	D	Decrease
M306L	PD	D	D	Decrease
M306V	PD	D	N	Decrease
M306F	PD	D	D	Decrease
M306I	PSD	D	N	Decrease
M306T	PSD	D	N	Decrease
V309A	B	N	N	Decrease
V309G	PSD	D	D	Decrease
A310R	PSD	D	D	Decrease
D311R	B	D	N	Decrease
D311F	PD	D	D	Decrease
D311G	B	N	N	Decrease
D311H	B	D	N	Decrease
H312R	B	D	N	Decrease
Y315L	PD	D	D	Increase
M316I	PSD	D	N	Decrease

S317F	PSD	D	D	Increase
S317T	B	N	N	Decrease
N318H	PD	D	D	Decrease
N318S	PD	D	D	Decrease
N318K	PD	D	D	Decrease
Y319N	PD	D	D	Decrease
Y319D	PD	D	D	Decrease
Y319C	PD	D	D	Decrease
Y319S	PD	D	D	Decrease
F320L	PD	D	D	Decrease
W322C	PD	D	D	Decrease
W322R	PD	D	D	Decrease
D328G	PD	D	N	Decrease
D328V	PSD	N	N	Decrease
D328H	PD	N	N	Decrease
D328Y	PSD	N	N	Increase
F330I	PD	D	D	Decrease
F330L	PD	D	D	Decrease
F330V	PD	D	D	Decrease
G331R	PD	D	D	Decrease
W332R	PD	D	D	Decrease
Y334H	PSD	D	N	Decrease
D345G	PSD	N	N	Decrease
S347T	PSD	D	D	Increase
S347C	PD	D	N	Decrease
S347I	PSD	D	N	Increase
D354N	PSD	N	N	Decrease
D354T	PD	N	N	Decrease
D354A	B	N	N	Decrease
A356F	PSD	N	D	Decrease
A356S	B	N	N	Decrease
A356V	B	N	N	Increase
A357S	B	N	N	Decrease
A357T	PSD	N	N	Decrease
A357V	B	N	N	Increase
G358V	PD	D	D	Decrease
L359I	B	N	N	Decrease
V360A	B	N	N	Decrease
V360M	PSD	N	N	Decrease
S366L	PD	D	D	Decrease
S366P	PD	D	D	Increase
R367P	PD	D	D	Decrease
E368Q	PD	D	N	Decrease
E368D	PSD	D	N	Decrease
E368A	PSD	N	N	Decrease
V369L	PD	D	D	Decrease
V369A	PSD	D	D	Decrease
L370R	PD	D	D	Decrease

261

P371R	PD	D	N	Decrease
G374V	PD	D	D	Decrease
P375A	B	N	N	Decrease
V377M	PD	D	D	Decrease
V377E	PD	D	D	Decrease
V377G	PD	N	N	Decrease
E378A	B	N	N	Decrease
E378K	B	N	N	Decrease
A379T	B	N	N	Decrease
A379D	B	N	N	Decrease
S380R	PSD	D	N	Increase
S380N	PSD	N	N	Increase
S380G	PSD	N	N	Decrease
S380D	B	N	N	Increase
Y384N	B	N	N	Increase
A388G	B	N	N	Decrease
T393A	B	N	N	Decrease
W395R	PD	D	D	Decrease
W395C	PD	D	D	Decrease
P397T	PD	D	D	Decrease
P397R	PD	D	D	Decrease
P397Q	PD	D	D	Decrease
F398H	PD	D	D	Decrease
F398Y	B	N	N	Decrease
N399T	B	D	N	Increase
N399I	PSD	D	D	Increase
N399D	PD	D	D	Decrease
N399H	PD	D	D	Decrease
N400P	PD	D	D	Increase
N400K	PD	D	D	Decrease
G401S	PD	D	D	Decrease
L402V	PD	D	D	Decrease
P404A	PD	D	D	Decrease
P404S	PD	D	D	Decrease
E405D	PD	D	D	Decrease
E405P	PD	D	D	Increase
G406P	PD	N	N	Decrease
G406S	PSD	N	N	Decrease
G406C	B	N	N	Decrease
G406K	PSD	N	N	Decrease
G406R	PD	N	N	Decrease
G406D	B	N	N	Decrease
G406A	B	N	N	Decrease
S412L	B	N	N	Decrease
S412P	PSD	N	N	Increase
M423T	B	N	D	Decrease
S426N	B	N	N	Increase
P430L	PD	D	D	Decrease

MUTATION	PolyPhen-2	PROVEAN	SIFT	I-MUTANT 3.0
A431T	PSD	D	N	Decrease
V435G	B	D	D	Decrease
V436G	B	D	D	Decrease
T437A	B	N	N	Decrease
P446H	PD	D	D	Decrease
G448V	PD	D	D	Decrease
I450M	PD	N	N	Decrease
V452L	PSD	N	N	Decrease
A454T	PSD	D	N	Decrease
G459A	B	N	N	Decrease
R460C	PD	D	N	Decrease
R460L	PSD	D	N	Decrease
P461S	PSD	D	N	Decrease
I465D	PD	D	D	Decrease
R469P	PD	D	D	Decrease
R471P	B	D	N	Decrease
M482I	B	N	N	Decrease
Q497P	PD	D	D	Decrease
Q497H	PD	D	D	Decrease
Q497F	PD	D	D	Increase
Q497R	PD	D	D	Decrease
Q497K	PD	D	D	Increase
E504D	PD	D	N	Decrease
A505V	PD	D	D	Decrease
R507G	PSD	D	N	Decrease
R507K	B	D	N	Decrease
M557I	B	N	N	Decrease
S565G	B	N	N	Decrease
V602A	PSD	D	D	Decrease
N624D	PD	D	D	Decrease
T642A	B	N	N	Decrease
T643I	B	N	N	Decrease
G745D	PD	D	N	Decrease
M1000R	B	N	D	Decrease
H1002R	PD	D	N	Decrease
D1024N	B	N	N	Decrease

Table E.6: Functional change and structural stability change for arabinosyltransferase C (*embC*)

MUTATION	PolyPhen-2	PROVEAN	SIFT	I-MUTANT 3.0
P150S	PD	N	N	Decrease
S213C	PD	D	D	Decrease
A244T	B	N	N	Decrease
A247P	PD	D	D	Increase
L251R	PD	D	D	Decrease
A254G	B	N	N	Decrease

T270I	B	N	N	Decrease
G272S	B	N	N	Decrease
H285Y	B	D	N	Increase
V287F	B	N	D	Decrease
G288W	PD	D	D	Decrease
G288V	PD	D	D	Decrease
Y296H	PD	D	D	Decrease
Y296S	PD	D	D	Decrease
I297L	PD	D	N	Decrease
I297T	B	N	N	Decrease
M300R	PD	D	D	Decrease
R302G	PD	D	D	Decrease
V303G	PSD	D	D	Decrease
E305D	PD	N	N	Decrease
A307T	PSD	N	D	Decrease
G308D	PSD	D	D	Decrease
Y309N	PD	D	D	Decrease
M310K	PD	D	D	Decrease
G325S	PD	D	N	Decrease
W326R	PD	D	D	Decrease
Y327N	PD	D	D	Decrease
D329E	PD	N	N	Increase
A378V	PSD	N	N	Increase
N394D	PD	D	D	Decrease
I406L	PSD	N	N	Decrease
A426T	PD	D	D	Decrease
V451I	PD	N	N	Decrease
P707L	PD	D	D	Decrease
Q725R	B	N	N	Decrease
R738Q	B	N	N	Decrease
V981L	PD	D	D	Decrease
V987A	PD	N	N	Decrease

Appendix F

F. Lethal, moderate, mild and neutral mutations in first-line TB drug targets identified from comprehensive bioinformatics analysis

Appendix F includes six tables. Tables F.1, F.2, F.3 and F4 summarise the mutations with a lethal, moderate, mild and neutral impact on each first-line TB drug target in *Mycobacterium tuberculosis.* Table F.1, F.2, F.3 and F4 include:

Table F.1: Mutations with lethal impact on catalase-peroxidase (*katG*), pyrazinamidase (*pncA*)*,* β-subunit of RNA polymerase (*rpoB*)*,* arabinosyltransferase A (*embA*), arabinosyltransferase B (*embB*) and arabinosyltransferase C (*embC*)

Table F.2: Mutations with moderate impact on catalase-peroxidase (*katG*), pyrazinamidase (*pncA*)*,* β-subunit of RNA polymerase (*rpoB*)*,* arabinosyltransferase A (*embA*), arabinosyltransferase B (*embB*) and arabinosyltransferase C (*embC*)

Table F.3: Mutations with mild impact on catalase-peroxidase (*katG*), pyrazinamidase (*pncA*)*,* β-subunit of RNA polymerase (*rpoB*)*,* arabinosyltransferase A (*embA*), arabinosyltransferase B (*embB*) and arabinosyltransferase C (*embC*)

Table F.4: Mutations with a neutral impact on catalase-peroxidase (*katG*), pyrazinamidase (*pncA*) and arabinosyltransferase B (*embB*). No neutral mutations were found in the β-subunit of RNA polymerase (*rpoB*)*,* arabinosyltransferase A (*embA*), and arabinosyltransferase C (*embC*)

Table F.1: Mutations with lethal impact on catalase-peroxidase (*katG*), pyrazinamidase (*pncA*)*,* β-subunit of RNA polymerase (*rpoB*)*,* arabinosyltransferase A (*embA*), arabinosyltransferase B (*embB*) and arabinosyltransferase C (*embC*)

Catalase-peroxidase (*katG*)		Pyrazinamidase (*pncA*)		β-subunit of RNA polymerase (*rpoB*)		Arabinosyltransferase A (*embA*)		Arabinosyltransferase B (*embB*)		Arabinosyltransferase C (*embC*)	
Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site
R104Q	site-1	D8A	site-1	Q432P	site-1	L105V	site-3	V128G	site-3	L215R	site-3
W107R	site-1	D8G	site-1	Q432H	site-1	R380P	site-3	D240H	site-3	G288W	site-3
Y229F	site-1	D8H	site-1	R447H	site-1	G554D	site-3	A310R	site-3	G288V	site-3
V230A	site-1	D8N	site-1	R447P	site-1	P639S	site-3	N318H	site-3	Y269H	site-3
P232R	site-1	D8Y	site-1	P483L	site-1	P838L	site-3	N318S	site-3	Y269S	site-3

P232A	site-1	F13V	site-1	I491F	site-1			N318K	site-3	R320G	site-3
P232S	site-1	F13L	site-1	H445N	site-2			Y319N	site-3	V330G	site-3
G269T	site-1	F13S	site-1	H445T	site-2			Y319D	site-3	G380D	site-3
H270A	site-1	D49A	site-1	H445S	site-2			Y319C	site-3	Y390N	site-3
T275A	site-1	D49N	site-1	H445G	site-2			Y319S	site-3	W362R	site-3
W321R	site-1	D49H	site-1	H445A	site-2			F330I	site-3	Y372N	site-3
W321L	site-1	D49G	site-1	H445R	site-2			F330L	site-3	N349D	site-3
W321G	site-1	H51N	site-1	H445Q	site-2			F330V	site-3	A462T	site-3
W321S	site-1	H51Q	site-1	N413H	site-3			W332R	site-3	P770L	site-3
M105I	site-2	H71Q	site-1	F424V	site-3			S366L	site-3		
H108D	site-2	H71D	site-1	F 424L	site-3			R367P	site-3		
H108Q	site-2	H71N	site-1	G426D	site-3			V369L	site-3		
A172T	site-2	H71R	site-1	T427P	site-3			V369A	site-3		
M176I	site-2	H71T	site-1	S428T	site-3			L370R	site-3		
G234E	site-2	L19P	site-2	Q429H	site-3			G374V	site-3		
G234R	site-2	L19R	site-2	L430P	site-3			V377M	site-3		
T251M	site-2	V21G	site-2	L430V	site-3			V377E	site-3		
M257I	site-2	T47A	site-2	L430R	site-3			W395R	site-3		
T262R	site-2	P54Q	site-2	L430K	site-3			W395C	site-3		
G316S	site-2	P54S	site-2	S431C	site-3			P397T	site-3		
G316D	site-2	P54R	site-2	S431T	site-3			P397R	site-3		
E318G	site-2	P54T	site-2	S431G	site-3			P397Q	site-3		
W328L	site-2	P54L	site-2	F433V	site-3			F398H	site-3		
W328R	site-2	W68R	site-2	F433L	site-3			N399D	site-3		
W328S	site-2	W68D	site-2	M434I	site-3			N399H	site-3		
W328C	site-2	W68C	site-2	N437H	site-3			N400K	site-3		
A350S	site-2	W68G	site-2	N437S	site-3			G401S	site-3		
L415P	site-2	W68S	site-2	N438K	site-3			L402V	site-3		
W91G	site-3	W68L	site-2	P439S	site-3			P404A	site-3		
W91R	site-3	C72R	site-2	L440P	site-3			P404S	site-3		
A93T	site-3	C72Y	site-2	S441Q	site-3			E405D	site-3		
D94N	site-3	C72W	site-2	G442W	site-3			P430L	site-3		
D94G	site-3	K96R	site-2	T444P	site-3			P446H	site-3		
Y98S	site-3	K96E	site-2	K446N	site-3			G448V	site-3		
A110V	site-3	K96Q	site-2	K446R	site-3			Q497P	site-3		
G121C	site-3	K96T	site-2	K446E	site-3			Q497H	site-3		
G121V	site-3	K96N	site-2	K446Q	site-3			Q497R	site-3		
G125C	site-3	G97A	site-2	R448P	site-3			N624D	site-3		
Q127P	site-3	G97S	site-2	R448Q	site-3						
R128Q	site-3	G97D	site-2	R448G	site-3						
R128P	site-3	I133T	site-2	R448L	site-3						
P131R	site-3	I5T	site-3	G453A	site-3						
P131Q	site-3	I5S	site-3	G453W	site-3						
N138H	site-3	I6S	site-3	G453V	site-3						
N138D	site-3	I6T	site-3	P454H	site-3						
N138S	site-3	V7G	site-3	P454S	site-3						
N138T	site-3	V7A	site-3	G455D	site-3						
D142N	site-3	V7F	site-3	L457R	site-3						

R146W	site-3	V7I	site-3	E460G	site-3						
W161Q	site-3	V9A	site-3	E481G	site-3						
W161R	site-3	V9G	site-3	T482P	site-3						
A162T	site-3	Q10H	site-3	S493L	site-3						
G169S	site-3	Q10P	site-3	E507G	site-3						
G186V	site-3	Q10K	site-3								
W191R	site-3	Q10R	site-3								
P241S	site-3	D12H	site-3								
E261K	site-3	D12G	site-3								
E261Q	site-3	D12A	site-3								
A264T	site-3	D12N	site-3								
G285C	site-3	C14R	site-3								
G285D	site-3	C14W	site-3								
A291T	site-3	L27R	site-3								
G299C	site-3	A28D	site-3								
W300C	site-3	I31T	site-3								
W300R	site-3	I31S	site-3								
W300G	site-3	L35R	site-3								
G307R	site-3	L35P	site-3								
G309C	site-3	V44G	site-3								
G309S	site-3	V45G	site-3								
D311F	site-3	V45A	site-3								
D311G	site-3	S59F	site-3								
A312R	site-3	V73F	site-3								
A312G	site-3	G78D	site-3								
T322A	site-3	F81S	site-3								
T322M	site-3	L85R	site-3								
T322N	site-3	L85P	site-3								
T324P	site-3	V93M	site-3								
T326M	site-3	F94C	site-3								
L336R	site-3	F94L	site-3								
L336P	site-3	Y99D	site-3								
W341S	site-3	Y103D	site-3								
T344P	site-3	Y103H	site-3								
R385P	site-3	Y103S	site-3								
P388S	site-3	Y103C	site-3								
P388L	site-3	G105R	site-3								
A409D	site-3	L116V	site-3								
G490C	site-3	L120R	site-3								
G490D	site-3	L120P	site-3								
G491C	site-3	R121W	site-3								
R496L	site-3	V125D	site-3								
R498H	site-3	V125G	site-3								
W505S	site-3	V128G	site-3								
Q525P	site-3	V130A	site-3								
E607K	site-3	V130G	site-3								
G629S	site-3	A134S	site-3								
L653P	site-3	A134V	site-3								

R705L	site-3	D136H	site-3
W728C	site-3	C138S	site-3
D735N	site-3	C138Y	site-3
D735A	site-3	V139M	site-3
		V139L	site-3
		V139G	site-3
		V139A	site-3
		R140H	site-3
		R140P	site-3
		A146E	site-3
		A146T	site-3
		A146V	site-3
		R148C	site-3
		L182S	site-3

Table F.2: Mutations with moderate impact on catalase-peroxidase (*katG*), pyrazinamidase (*pncA*), β-subunit of RNA polymerase (*rpoB*), arabinosyltransferase A (*embA*), arabinosyltransferase B (*embB*) and arabinosyltransferase C (*embC*)

Catalase-peroxidase (*katG*)		Pyrazinamidase (*pncA*)		β-subunit of RNA polymerase (*rpoB*)		Arabinosyltransferase A (*embA*)		Arabinosyltransferase B (*embB*)		Arabinosyltransferase C (*embC*)	
Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site
L101P	site-1	D8E	site-1	Q432K	site-1	V122G	site-3	L239P	site-3	S213C	site-3
K274R	site-1	D49V	site-1	Q432E	site-1	G200S	site-3	R257W	site-3	H285Y	site-3
T275P	site-1	H51Y	site-1	Q432L	site-1	V206M	site-3	I293T	site-3	M300R	site-3
S315G	site-1	H51P	site-1	H445C	site-2	A331T	site-3	S297A	site-3	M310K	site-3
T380I	site-1	H51R	site-1	H445L	site-2	V343L	site-3	S298W	site-3	V981L	site-3
F408L	site-1	H51D	site-1	H445Y	site-2	G350D	site-3	S298A	site-3		
A172V	site-2	H57D	site-1	H445E	site-2	A576T	site-3	D299E	site-3		
E318V	site-2	H57Q	site-1	H445P	site-2	P769T	site-3	M306L	site-3		
L384R	site-2	H57R	site-1	H445D	site-2	P913S	site-3	M306F	site-3		
N35D	site-3	H71Y	site-1	V170F	site-3			V309G	site-3		
D63E	site-3	H71P	site-1	T427S	site-3			D311F	site-3		
V68G	site-3	H71E	site-1	S428R	site-3			Y315L	site-3		
T85P	site-3	A102T	site-1	S428Q	site-3			F320L	site-3		
A106V	site-3	A102V	site-1	S428I	site-3			W322C	site-3		
A109V	site-3	T47P	site-2	S428G	site-3			W322R	site-3		
D117A	site-3	T47S	site-2	L430M	site-3			G331R	site-3		
G118A	site-3	F58L	site-2	S431I	site-3			A357S	site-3		
G125V	site-3	M1I	site-3	S431N	site-3			A357T	site-3		
L141F	site-3	M1T	site-3	S431R	site-3			G358V	site-3		
K143T	site-3	A3E	site-3	S431M	site-3			S366P	site-3		
Y155S	site-3	A3S	site-3	M434T	site-3			P371R	site-3		
E217G	site-3	A3Q	site-3	M434V	site-3			V377G	site-3		

G279D	site-3	A3P	site-3	D435G	site-3			F398Y	site-3		
G285R	site-3	L4W	site-3	D435N	site-3			N399T	site-3		
E289D	site-3	L4S	site-3	D435H	site-3			N399I	site-3		
A291V	site-3	I6L	site-3	D435A	site-3			N400P	site-3		
A291P	site-3	V7D	site-3	D435P	site-3			E405P	site-3		
Q295P	site-3	V9S	site-3	D435T	site-3			A431T	site-3		
G305A	site-3	D12E	site-3	Q436P	site-3			I450M	site-3		
G309V	site-3	C14Y	site-3	Q436L	site-3			R460C	site-3		
G309A	site-3	G17S	site-3	N437Y	site-3			R460L	site-3		
G309F	site-3	G17D	site-3	N437I	site-3			I465D	site-3		
D311Y	site-3	S18P	site-3	N437T	site-3			R469P	site-3		
D311E	site-3	G23A	site-3	N437D	site-3			Q497F	site-3		
A312V	site-3	G23V	site-3	L440M	site-3			Q497K	site-3		
S331C	site-3	A25E	site-3	S441L	site-3			E504D	site-3		
Y337C	site-3	A26G	site-3	S441F	site-3			A505V	site-3		
K345T	site-3	L27P	site-3	S441P	site-3			R507G	site-3		
D357H	site-3	A28V	site-3	S441N	site-3			R507K	site-3		
T394A	site-3	Y34D	site-3	S441W	site-3			V602A	site-3		
W397Y	site-3	Y34S	site-3	S450Q	site-3						
G428R	site-3	Y41H	site-3	S450A	site-3						
V442G	site-3	H43Y	site-3	S450G	site-3						
L449F	site-3	A46E	site-3	S450C	site-3						
E454R	site-3	A46S	site-3	A451D	site-3						
Q471R	site-3	A46V	site-3	L452P	site-3						
G485V	site-3	D53A	site-3	L452R	site-3						
N529D	site-3	S59P	site-3	L452E	site-3						
D573G	site-3	P62H	site-3	L452V	site-3						
D573N	site-3	P62T	site-3	I480V	site-3						
L587P	site-3	P62Q	site-3	I488V	site-3						
L611R	site-3	P62L	site-3								
A636E	site-3	D63A	site-3								
		D63G	site-3								
		Y64D	site-3								
		S66P	site-3								
		S67P	site-3								
		P69A	site-3								
		P69R	site-3								
		P69L	site-3								
		T76A	site-3								
		T76P	site-3								
		T76I	site-3								
		A79T	site-3								
		A79G	site-3								
		D80N	site-3								
		H82D	site-3								
		H82R	site-3								
		H82L	site-3								
		P83R	site-3								

		F94S	site-3									
		F94P	site-3									
		T100A	site-3									
		T100P	site-3									
		S104R	site-3									
		G108R	site-3									
		N112Y	site-3									
		T114A	site-3									
		T114P	site-3									
		L116R	site-3									
		N118T	site-3									
		W119L	site-3									
		W119G	site-3									
		W119R	site-3									
		W119C	site-3									
		V125L	site-3									
		V125F	site-3									
		V130L	site-3									
		V131F	site-3									
		G132A	site-3									
		G132C	site-3									
		G132R	site-3									
		G132D	site-3									
		G132S	site-3									
		T135A	site-3									
		T135P	site-3									
		T135N	site-3									
		D136Y	site-3									
		D136N	site-3									
		D136G	site-3									
		H137D	site-3									
		H137R	site-3									
		H137P	site-3									
		C138R	site-3									
		C138T	site-3									
		C138W	site-3									
		R140S	site-3									
		Q141P	site-3									
		T142P	site-3									
		T142A	site-3									
		T142K	site-3									
		T142M	site-3									
		A143P	site-3									
		A143G	site-3									
		A143T	site-3									
		A146P	site-3									
		R148S	site-3									
		L151S	site-3									

		Mutation	Mutant site								
		V155M	site-3								
		V155G	site-3								
		V155A	site-3								
		G162A	site-3								
		G162D	site-3								
		M175T	site-3								
		M175R	site-3								
		V180G	site-3								

Table F.3: Mutations with mild impact on catalase-peroxidase (*katG*), pyrazinamidase (*pncA*), β-subunit of RNA polymerase (*rpoB*), arabinosyltransferase A (*embA*), arabinosyltransferase B (*embB*) and arabinosyltransferase C (*embC*)

Catalase-peroxidase (*katG*)		Pyrazinamidase (*pncA*)		β-subunit of RNA polymerase (*rpoB*)		Arabinosyltransferase A (*embA*)		Arabinosyltransferase B (*embB*)		Arabinosyltransferase C (*embC*)	
Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site
I248M	site-1	H57Y	site-1	N487S	site-1	D4N	site-3	N13S	site-3	P150S	site-3
T314N	site-1	H57P	site-1	D435Y	site-3	G5S	site-3	V50A	site-3	A244T	site-3
S315A	site-1	H57L	site-1	D435V	site-3	G5V	site-3	L74R	site-3	A247P	site-3
S315R	site-1	T87M	site-3	D435K	site-3	V125G	site-3	A281S	site-3	A254G	site-3
S315T	site-1	R154T	site-3	D435F	site-3	A201T	site-3	V282G	site-3	T270I	site-3
S315N	site-1	R154G	site-3	D435E	site-3	V468A	site-3	L288V	site-3	G272S	site-3
S315D	site-1	V155L	site-3	S450W	site-3			N296I	site-3	V287F	site-3
S315L	site-1	V157G	site-3	S450L	site-3			N296H	site-3	I297T	site-3
S315I	site-1	L159P	site-3	S450Y	site-3			N296K	site-3	I297L	site-3
I317L	site-1	L159R	site-3	S450F	site-3			L304V	site-3	E305D	site-3
I317V	site-1	T160K	site-3	L452M	site-3			M306V	site-3	A307T	site-3
T11A	site-3	T160A	site-3					M306I	site-3	G325S	site-3
A61T	site-3	T160P	site-3					M306T	site-3	D329E	site-3
D74Y	site-3	A161P	site-3					V309A	site-3	A378V	site-3
D74G	site-3	V163A	site-3					D311G	site-3	I406L	site-3
M84I	site-3	S164P	site-3					D311R	site-3	V451I	site-3
Y155C	site-3	A165T	site-3					D311H	site-3	Q725R	site-3
D194Y	site-3	T168P	site-3					H312R	site-3	R738Q	site-3
E195K	site-3	T168N	site-3					M316I	site-3	V987A	site-3
G206D	site-3	A171V	site-3					S317F	site-3		
N218K	site-3	A171T	site-3					S317T	site-3		
N236T	site-3	A171E	site-3					D328Y	site-3		
T271V	site-3	A171P	site-3					D328V	site-3		
P280S	site-3	L172P	site-3					D328H	site-3		
P280H	site-3	L172R	site-3					D328G	site-3		
S302R	site-3	L172A	site-3					Y334H	site-3		
S303W	site-3	M175I	site-3					D345G	site-3		
T308P	site-3	M175V	site-3					S347T	site-3		

								S347C	site-3		
D329G	site-3	T177P	site-3					S347C	site-3		
I335T	site-3	V180F	site-3					S347I	site-3		
I335V	site-3	C184Y	site-3					D354N	site-3		
D406A	site-3							D354T	site-3		
A424G	site-3							D354A	site-3		
R463W	site-3							A356F	site-3		
R463H	site-3							A356S	site-3		
R463L	site-3							L359I	site-3		
R515C	site-3							V360A	site-3		
S652A	site-3							V360M	site-3		
S700P	site-3							E368A	site-3		
Q717P	site-3							E368Q	site-3		
								E368D	site-3		
								P375A	site-3		
								E378A	site-3		
								E378K	site-3		
								A379T	site-3		
								A379D	site-3		
								S380R	site-3		
								S380N	site-3		
								S380G	site-3		
								A388G	site-3		
								T393A	site-3		
								G406P	site-3		
								G406S	site-3		
								G406C	site-3		
								G406K	site-3		
								G406R	site-3		
								G406D	site-3		
								G406A	site-3		
								S412P	site-3		
								S412L	site-3		
								M423T	site-3		
								V435G	site-3		
								V436G	site-3		
								T437A	site-3		
								V452L	site-3		
								A454T	site-3		
								G459A	site-3		
								P461S	site-3		
								R471P	site-3		
								M482I	site-3		
								M557I	site-3		
								S565G	site-3		
								T642A	site-3		
								T643I	site-3		
								G745D	site-3		
								M1000R	site-3		

								H1002R	site-3		
								D1024N	site-3		

Table F.4: Mutations with a neutral impact on catalase-peroxidase (*katG*), pyrazinamidase (*pncA*) and arabinosyltransferase B (*embB*). No neutral mutations were found in the β-subunit of RNA polymerase (*rpoB*), arabinosyltransferase A (*embA*), and arabinosyltransferase C (*embC*)

Catalase peroxidase (*katG*)		Pyrazinamidase (*pncA*)		Arabinosyltransferase B (*embB*)	
Mutation	Mutant site	Mutation	Mutant site	Mutation	Mutant site
T12P	site-3	D80E	site-3	G264R	site-3
D357N	site-3			A218V	site-3
A379V	site-3			A365V	site-3
S446R	site-3			A375V	site-3
				S308D	site-3
				Y348N	site-3
				S462N	site-3

Appendix G

G. List of 159 *Mycobacterium tuberculosis* strains selected for identification of human TB vaccine candidates

Table G.1 provides information on the 159 *Mycobacterium tuberculosis* strains selected for identification of vaccine candidates for human tuberculosis. The complete proteome sequence of 159 different *Mycobacterium tuberculosis* strains were downloaded from the NCBI Genome FTP site. Table G.1 includes:

1- Strain number
2- *Mycobacterium tuberculosis* isolate
3- Taxon ID
4- INSDC (**i**nternational **n**ucleotide **s**equence **d**atabase **c**ollaboration)
5- GenBank assembly accession

Table G.1: 159 strains of *Mycobacterium tuberculosis*

Strain	Strain Name	Taxon ID	Accession	Assembly
Strain1	Mycobacterium tuberculosis H37Rv complete genome	83332	AL123456.3	GCA_000195955.2
Strain2	Mycobacterium tuberculosis CDC1551	83331	AE000516.2	GCA_000008585.1
Strain3	Mycobacterium tuberculosis F11	336982	CP000717.1	GCA_000016925.1
Strain4	Mycobacterium tuberculosis KZN 1435	478434	CP001658.1	GCA_000023625.1
Strain5	Mycobacterium tuberculosis str. Haarlem	395095	CP001664.1	GCA_000153685.2
Strain6	Mycobacterium tuberculosis KZN 4207	478433	CP001662.1	GCA_000154585.2
Strain7	Mycobacterium tuberculosis KZN 605	478435	CP001976.1	GCA_000154605.2
Strain8	Mycobacterium tuberculosis W-148	659019	CP012090.1	GCA_000193185.2
Strain9	Mycobacterium tuberculosis CTRI-2	707235	CP002992.1	GCA_000224435.1
Strain10	Mycobacterium tuberculosis CCDC5180	443150	CP001642.1	GCA_000270365.1
Strain11	Mycobacterium tuberculosis H37Rv	83332	CP003248.2	GCA_000277735.2
Strain12	Mycobacterium tuberculosis 7199-99	1138877	HE663067.1	GCA_000331445.1
Strain13	Mycobacterium tuberculosis str. Erdman = ATCC 35801 DNA	652616	AP012340.1	GCA_000350205.1
Strain14	Mycobacterium tuberculosis str. Beijing/NITR203	1306400	CP005082.1	GCA_000364825.1
Strain15	Mycobacterium tuberculosis EAI5/NITR206	1310115	CP005387.1	GCA_000389945.1
Strain16	Mycobacterium tuberculosis CCDC5079	443149	CP002884.1	GCA_000400615.1
Strain17	Mycobacterium tuberculosis EAI5	1306414	CP006578.1	GCA_000422125.1
Strain18	Mycobacterium tuberculosis HKBS1	1010834	CP002871.1	GCA_000572125.1
Strain19	Mycobacterium tuberculosis BT2	1010835	CP002882.1	GCA_000572155.1
Strain20	Mycobacterium tuberculosis BT1	1010836	CP002883.1	GCA_000572175.1

Strain21	Mycobacterium tuberculosis CCDC5180	443150	CP002885.1	GCA_000572195.1
Strain22	Mycobacterium tuberculosis K	1249615	CP007803.1	GCA_000698475.1
Strain23	Mycobacterium tuberculosis Korean Strain KIT87190	1773	CP007809.1	GCA_000706665.1
Strain24	Mycobacterium tuberculosis ZMC13-264	1773	CP009100.1	GCA_000738445.1
Strain25	Mycobacterium tuberculosis ZMC13-88	1773	CP009101.1	GCA_000738475.1
Strain26	Mycobacterium tuberculosis 96075	1773	CP009426.1	GCA_000756525.1
Strain27	Mycobacterium tuberculosis strain 96121	1773	CP009427.1	GCA_000756545.1
Strain28	Mycobacterium tuberculosis 49-02	1427516	HG813240.1	GCA_000786505.1
Strain29	Mycobacterium tuberculosis H37RvSiena	1437856	CP007027.1	GCA_000827085.1
Strain30	Mycobacterium tuberculosis str. Kurono DNA	1445606	AP014573.1	GCA_000828995.1
Strain31	Mycobacterium tuberculosis H37Rv, TMC 102	83332	CP009480.1	GCA_000831245.1
Strain32	Mycobacterium tuberculosis strain SCAID 187.0	1773	CP012506.2	GCA_001275565.2
Strain33	Mycobacterium tuberculosis strain F28	1773	CP010330.1	GCA_001544705.1
Strain34	Mycobacterium tuberculosis strain 22115	1773	CP010337.1	GCA_001544955.1
Strain35	Mycobacterium tuberculosis strain 22103	1773	CP010339.1	GCA_001545015.1
Strain36	Mycobacterium tuberculosis strain SCAID 320.0	1773	CP016794.1	GCA_001702435.1
Strain37	Mycobacterium tuberculosis strain SCAID 252.0	1773	CP016888.1	GCA_001708265.1
Strain38	Mycobacterium tuberculosis strain Beijing	1773	CP011510.1	GCA_001750865.1
Strain39	Mycobacterium tuberculosis strain 1458	1773	CP013475.1	GCA_001855255.1
Strain40	Mycobacterium tuberculosis strain TB282	1773	CP017920.1	GCA_001870145.1
Strain41	Mycobacterium tuberculosis strain I0004241-1	1773	CP018303.1	GCA_001895765.1
Strain42	Mycobacterium tuberculosis strain M0018684-2	1773	CP018305.1	GCA_001895785.1
Strain43	Mycobacterium tuberculosis strain I0004000-1	1773	CP018302.1	GCA_001895805.1
Strain44	Mycobacterium tuberculosis strain I0002801-4	1773	CP018301.1	GCA_001895825.1
Strain45	Mycobacterium tuberculosis strain I0002353-6	1773	CP018300.1	GCA_001895845.1
Strain46	Mycobacterium tuberculosis strain M0002959-6	1773	CP018304.1	GCA_001895865.1
Strain47	Mycobacterium tuberculosis strain DK9897	1773	CP018778.1	GCA_001922485.1
Strain48	Mycobacterium tuberculosis strain MTB1	1773	CP020381.2	GCA_002072775.2
Strain49	Mycobacterium tuberculosis strain Beijing-like/35049	1773	CP017593.1	GCA_002116755.1
Strain50	Mycobacterium tuberculosis strain Beijing-like/36918	1773	CP017594.1	GCA_002116775.1
Strain51	Mycobacterium tuberculosis strain Beijing-like/38774	1773	CP017595.1	GCA_002116795.1
Strain52	Mycobacterium tuberculosis strain Beijing/391	1773	CP017596.1	GCA_002116815.1
Strain53	Mycobacterium tuberculosis strain Beijing-like/50148	1773	CP017597.1	GCA_002116835.1
Strain54	Mycobacterium tuberculosis strain Beijing-like/1104	1773	CP017598.1	GCA_002116855.1
Strain55	Mycobacterium tuberculosis strain MTB2	1773	CP022014.1	GCA_002208235.1
Strain56	Mycobacterium tuberculosis DNA, complete genome, strain: NCGM946K2	1773	AP017901.1	GCA_002356015.1
Strain57	Mycobacterium tuberculosis DNA, complete genome, strain: HN-024	1773	AP018033.1	GCA_002356255.1
Strain58	Mycobacterium tuberculosis DNA, complete genome, strain: HN-205	1773	AP018034.1	GCA_002357935.1
Strain59	Mycobacterium tuberculosis DNA, complete genome, strain: HN-321	1773	AP018035.1	GCA_002357955.1
Strain60	Mycobacterium tuberculosis DNA, complete genome, strain: HN-506	1773	AP018036.1	GCA_002357975.1

Strain61	Mycobacterium tuberculosis strain CSV4519	1773	CP023573.1	GCA_002446875.1
Strain62	Mycobacterium tuberculosis strain CSV4644	1773	CP023574.1	GCA_002446895.1
Strain63	Mycobacterium tuberculosis strain CSV5769 chromosome	1773	CP023575.1	GCA_002446915.1
Strain64	Mycobacterium tuberculosis strain CSV10399 chromosome	1773	CP023576.1	GCA_002446935.1
Strain65	Mycobacterium tuberculosis strain CSV11678	1773	CP023577.1	GCA_002446955.1
Strain66	Mycobacterium tuberculosis strain LE486 chromosome	1773	CP023578.1	GCA_002446975.1
Strain67	Mycobacterium tuberculosis strain LE492	1773	CP023579.1	GCA_002446995.1
Strain68	Mycobacterium tuberculosis strain LN180	1773	CP023580.1	GCA_002447015.1
Strain69	Mycobacterium tuberculosis strain LN2358	1773	CP023581.1	GCA_002447035.1
Strain70	Mycobacterium tuberculosis strain LN3756	1773	CP023582.1	GCA_002447055.1
Strain71	Mycobacterium tuberculosis strain MDRDM260	1773	CP023583.1	GCA_002447075.1
Strain72	Mycobacterium tuberculosis strain MDRDM627	1773	CP023584.1	GCA_002447095.1
Strain73	Mycobacterium tuberculosis strain MDRDM1098	1773	CP023585.1	GCA_002447115.1
Strain74	Mycobacterium tuberculosis strain MDRMA2491	1773	CP023586.1	GCA_002447135.1
Strain75	Mycobacterium tuberculosis strain ME1473	1773	CP023587.1	GCA_002447155.1
Strain76	Mycobacterium tuberculosis strain TBDM425	1773	CP023588.1	GCA_002447175.1
Strain77	Mycobacterium tuberculosis strain TBV5000	1773	CP023589.1	GCA_002447195.1
Strain78	Mycobacterium tuberculosis strain TBV5362	1773	CP023590.1	GCA_002447215.1
Strain79	Mycobacterium tuberculosis strain TBV5365	1773	CP023591.1	GCA_002447235.1
Strain80	Mycobacterium tuberculosis strain SLM036	1773	CP023592.1	GCA_002447255.1
Strain81	Mycobacterium tuberculosis strain SLM040	1773	CP023593.1	GCA_002447275.1
Strain82	Mycobacterium tuberculosis strain SLM056	1773	CP023594.1	GCA_002447295.1
Strain83	Mycobacterium tuberculosis strain SLM060	1773	CP023595.1	GCA_002447315.1
Strain84	Mycobacterium tuberculosis strain SLM063	1773	CP023596.1	GCA_002447335.1
Strain85	Mycobacterium tuberculosis strain SLM088	1773	CP023597.1	GCA_002447355.1
Strain86	Mycobacterium tuberculosis strain SLM100	1773	CP023598.1	GCA_002447375.1
Strain87	Mycobacterium tuberculosis strain CSV383	1773	CP023599.1	GCA_002447395.1
Strain88	Mycobacterium tuberculosis strain CSV3611	1773	CP023600.1	GCA_002447415.1
Strain89	Mycobacterium tuberculosis strain CSV9577	1773	CP023601.1	GCA_002447435.1
Strain90	Mycobacterium tuberculosis strain LE13	1773	CP023602.1	GCA_002447455.1
Strain91	Mycobacterium tuberculosis strain LE63	1773	CP023603.1	GCA_002447475.1
Strain92	Mycobacterium tuberculosis strain LE76	1773	CP023604.1	GCA_002447495.1
Strain93	Mycobacterium tuberculosis strain LE79	1773	CP023605.1	GCA_002447515.1
Strain94	Mycobacterium tuberculosis strain LE103	1773	CP023606.1	GCA_002447535.1
Strain95	Mycobacterium tuberculosis strain LE371	1773	CP023607.1	GCA_002447555.1
Strain96	Mycobacterium tuberculosis strain LE410	1773	CP023608.1	GCA_002447575.1
Strain97	Mycobacterium tuberculosis strain LN55	1773	CP023609.1	GCA_002447595.1
Strain98	Mycobacterium tuberculosis strain LN317	1773	CP023610.1	GCA_002447615.1
Strain99	Mycobacterium tuberculosis strain LN763	1773	CP023611.1	GCA_002447635.1
Strain100	Mycobacterium tuberculosis strain LN2978	1773	CP023612.1	GCA_002447655.1
Strain101	Mycobacterium tuberculosis strain LN3584	1773	CP023613.1	GCA_002447675.1

Strain102	Mycobacterium tuberculosis strain LN3588	1773	CP023614.1	GCA_002447695.1
Strain103	Mycobacterium tuberculosis strain LN3589	1773	CP023615.1	GCA_002447715.1
Strain104	Mycobacterium tuberculosis strain LN3668	1773	CP023616.1	GCA_002447735.1
Strain105	Mycobacterium tuberculosis strain LN3672	1773	CP023617.1	GCA_002447755.1
Strain106	Mycobacterium tuberculosis strain LN3695	1773	CP023618.1	GCA_002447775.1
Strain107	Mycobacterium tuberculosis strain LN1100	1773	CP023619.1	GCA_002447795.1
Strain108	Mycobacterium tuberculosis strain LN1856	1773	CP023620.1	GCA_002447815.1
Strain109	Mycobacterium tuberculosis strain LN2900	1773	CP023621.1	GCA_002447835.1
Strain110	Mycobacterium tuberculosis strain MDRDM827	1773	CP023622.1	GCA_002447855.1
Strain111	Mycobacterium tuberculosis strain MDRMA203	1773	CP023623.1	GCA_002447875.1
Strain112	Mycobacterium tuberculosis strain MDRMA701	1773	CP023624.1	GCA_002447895.1
Strain113	Mycobacterium tuberculosis strain MDRMA863	1773	CP023625.1	GCA_002447915.1
Strain114	Mycobacterium tuberculosis strain MDRMA1565	1773	CP023626.1	GCA_002447935.1
Strain115	Mycobacterium tuberculosis strain MDRMA2019	1773	CP023627.1	GCA_002447955.1
Strain116	Mycobacterium tuberculosis strain MDRMA2082	1773	CP023628.1	GCA_002447975.1
Strain117	Mycobacterium tuberculosis strain MDRMA2260	1773	CP023629.1	GCA_002447995.1
Strain118	Mycobacterium tuberculosis strain MDRMA2441	1773	CP023630.1	GCA_002448015.1
Strain119	Mycobacterium tuberculosis strain TBDM1506	1773	CP023631.1	GCA_002448035.1
Strain120	Mycobacterium tuberculosis strain TBDM2189	1773	CP023632.1	GCA_002448055.1
Strain121	Mycobacterium tuberculosis strain TBDM2444	1773	CP023633.1	GCA_002448075.1
Strain122	Mycobacterium tuberculosis strain TBDM2487	1773	CP023634.1	GCA_002448095.1
Strain123	Mycobacterium tuberculosis strain TBDM2489	1773	CP023635.1	GCA_002448115.1
Strain124	Mycobacterium tuberculosis strain TBDM2699	1773	CP023636.1	GCA_002448135.1
Strain125	Mycobacterium tuberculosis strain TBDM2717	1773	CP023637.1	GCA_002448155.1
Strain126	Mycobacterium tuberculosis strain TBV4766	1773	CP023638.1	GCA_002448175.1
Strain127	Mycobacterium tuberculosis strain TBV4768	1773	CP023639.1	GCA_002448195.1
Strain128	Mycobacterium tuberculosis strain TBV4952	1773	CP023640.1	GCA_002448215.1
Strain129	Mycobacterium tuberculosis strain GG-111-10	1773	CP025593.1	GCA_002886145.1
Strain130	Mycobacterium tuberculosis strain GG-5-10	1773	CP025594.1	GCA_002886165.1
Strain131	Mycobacterium tuberculosis strain GG-20-11	1773	CP025595.1	GCA_002886195.1
Strain132	Mycobacterium tuberculosis strain GG-90-10	1773	CP025601.1	GCA_002886335.1
Strain133	Mycobacterium tuberculosis strain GG-129-11	1773	CP025604.1	GCA_002886405.1
Strain134	Mycobacterium tuberculosis strain GG-134-11	1773	CP025605.1	GCA_002886505.1
Strain135	Mycobacterium tuberculosis strain GG-229-10	1773	CP025608.1	GCA_002886585.1
Strain136	Mycobacterium tuberculosis strain GG-37-11	1773	CP025598.1	GCA_002886685.1
Strain137	Mycobacterium tuberculosis strain GG-121-10	1773	CP025603.1	GCA_002886775.1
Strain138	Mycobacterium tuberculosis strain GG-186-10	1773	CP025607.1	GCA_002886865.1
Strain139	Mycobacterium tuberculosis strain GG-27-11	1773	CP025596.1	GCA_002886945.1
Strain140	Mycobacterium tuberculosis strain GG-36-11	1773	CP025597.1	GCA_002887065.1
Strain141	Mycobacterium tuberculosis strain GG-45-11	1773	CP025599.1	GCA_002887145.1
Strain142	Mycobacterium tuberculosis strain GG-109-10	1773	CP025602.1	GCA_002887255.1

Strain143	Mycobacterium tuberculosis strain GG-137-10	1773	CP025606.1	GCA_002887335.1
Strain144	Mycobacterium tuberculosis CCDC5079	443149	CP001641.1	GCA_000270345.1
Strain145	Mycobacterium tuberculosis RGTB327	1091500	CP003233.1	GCA_000277085.1
Strain146	Mycobacterium tuberculosis RGTB423	1091501	CP003234.1	GCA_000277105.1
Strain147	Mycobacterium tuberculosis str. Haarlem/NITR202	1304279	CP004886.1	GCA_000389905.1
Strain148	Mycobacterium tuberculosis CAS/NITR204	1310114	CP005386.1	GCA_000389925.1
Strain149	Mycobacterium tuberculosis strain Beijing-like	1773	CP010873.1	GCA_000954155.1
Strain150	Mycobacterium tuberculosis strain F1	1773	CP010329.1	GCA_001544675.1
Strain151	Mycobacterium tuberculosis strain 2242	1773	CP010335.1	GCA_001544895.1
Strain152	Mycobacterium tuberculosis strain 2279	1773	CP010336.1	GCA_001544935.1
Strain153	Mycobacterium tuberculosis strain 37004	1773	CP010338.1	GCA_001544985.1
Strain154	Mycobacterium tuberculosis strain 26105	1773	CP010340.1	GCA_001545055.1
Strain155	Mycobacterium tuberculosis strain Beijing2014PNGD	1773	CP022704.1	GCA_002895185.1
Strain156	Mycobacterium tuberculosis UT205	1097669	HE608151.1	GCA_000304555.1
Strain157	Mycobacterium tuberculosis 18b	1452723	CP007299.1	GCA_000835125.1
Strain158	Mycobacterium tuberculosis strain PR08	1773	CP010895.1	GCA_000934585.1
strain159	Mycobacterium tuberculosis strain PR10	1773	CP010968.1	GCA_001584745.1

Appendix H

H. Python scripts used for homology modelling

H.1: MODELLER v9.23 python script for template-target alignment- "catp-1sj2.py"

```python
from modeller import *
env = environ()
aln = alignment(env)
mdl = model(env, file='1sj2', model_segment=('FIRST:A','LAST:A'))
aln.append_model(mdl, align_codes='1sj2A', atom_files='1sj2.pdb')
aln.append(file='catp.ali', align_codes='catp')
aln.align2d()
aln.write(file='catp-1sj2A.ali', alignment_format='PIR')
aln.write(file='catp-1sj2A.pap', alignment_format='PAP')
```

H.2: MODELLER v9.23 python script for model building- "catp_model.py"

```python
from modeller import *
from modeller.automodel import *
#from modeller import soap_protein_od
env = environ()
a = automodel(env, alnfile='catp-1sj2A.ali',
        knowns='1sj2A', sequence='catp',
        assess_methods=(assess.DOPE,
                #soap_protein_od.Scorer(),
                assess.GA341))
a.starting_model = 1
a.ending_model = 10
a.make()
```

References

Abdallah, A. M., Gey van Pittius, N. C., DiGiuseppe Champion, P. A., Cox, J., Luirink, J., Vandenbroucke-Grauls, C. M. J. E., Appelmelk, B. J., & Bitter, W. (2007). Type VII secretion — mycobacteria show the way. *Nature Reviews Microbiology*, *5*(11), 883-891.

Adami, A. J., & Cervantes, J. L. (2015). The microbiome at the pulmonary alveolar niche and its role in *Mycobacterium tuberculosis* infection. *Tuberculosis*, *95*(6), 651-658.

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, *76*(1).

Aggarwal, M., Singh, A., Grover, S., Pandey, B., Kumari, A., & Grover, A. (2018). Role of *pnc* A gene mutations W68R and W68G in pyrazinamide resistance. *Journal of Cellular Biochemistry*, *119*(3), 2567-2578.

Ahmad, S., & Mokaddas, E. (2014). Current status and future trends in the diagnosis and treatment of drug-susceptible and multidrug-resistant tuberculosis. *Journal of Infection and Public Health*, *7*(2), 75-91.

Ahsan, M. J. (2015). Recent advances in the development of vaccines for tuberculosis. *Therapeutic Advances in Vaccines*, *3*(3), 66-75.

Alapati, R., Shuvo, Md. H., & Bhattacharya, D. (2020). SPECS: Integration of side-chain orientation and global distance-based measures for improved evaluation of protein structural models. *PLOS ONE*, *15*(2), e0228245.

Alcaide, F., Pfyffer, G. E., & Telenti, A. (1997). Role of *embB* in natural and acquired resistance to ethambutol in mycobacteria. *Antimicrobial Agents and Chemotherapy*, *41*(10), 2270-2273.

Alderwick, L. J., Lloyd, G. S., Ghadbane, H., May, J. W., Bhatt, A., Eggeling, L., Fütterer, K., & Besra, G. S. (2011). The C-Terminal Domain of the Arabinosyltransferase *Mycobacterium tuberculosis* EmbC Is a Lectin-Like Carbohydrate Binding Module. *PLoS Pathogens*, *7*(2), e1001299.

Almeida Da Silva, P. E., & Palomino, J. C. (2011). Molecular basis and mechanisms of drug resistance in *Mycobacterium tuberculosis*: classical and new drugs. *Journal of Antimicrobial Chemotherapy*, *66*(7), 1417-1430.

Almeida, R. R., Rosa, D. S., Ribeiro, S. P., Santana, V. C., Kallás, E. G., Sidney, J., Sette, A., Kalil, J., & Cunha-Neto, E. (2012). Broad and Cross-Clade CD4+ T-Cell Responses Elicited by a DNA Vaccine Encoding Highly Conserved and Promiscuous HIV-1 M-Group Consensus Peptides. *PLoS ONE*, *7*(9), e45267.

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389-3402.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403-410.

Altschul, S. F., & Koonin, E. v. (1998). Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends in Biochemical Sciences*, *23*(11), 444-447.

Andersen, P., & Kaufmann, S. H. E. (2014). Novel Vaccination Strategies against Tuberculosis. *Cold Spring Harbor Perspectives in Medicine*, *4*(6), a018523.

Andersson, D. I., & Hughes, D. (2010). Antibiotic resistance and its cost: is it possible to reverse resistance? *Nature Reviews Microbiology*, *8*(4), 260-271.

Ando, H., Kondo, Y., Suetake, T., Toyota, E., Kato, S., Mori, T., & Kirikae, T. (2010). Identification of *KatG* Mutations Associated with High-Level Isoniazid Resistance in *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy*, *54*(5), 1793-1799.

Andre, E., Goeminne, L., Cabibbe, A., Beckert, P., Kabamba Mukadi, B., Mathys, V., Gagneux, S., Niemann, S., van Ingen, J., & Cambau, E. (2017). Consensus numbering system for the rifampicin resistance-associated *rpoB* gene mutations in pathogenic mycobacteria. *Clinical Microbiology and Infection*, *23*(3), 167-172.

Ankley, L., Thomas, S., & Olive, A. J. (2020). Fighting Persistence: How Chronic Infections with *Mycobacterium tuberculosis* Evade T Cell-Mediated Clearance and New Strategies to Defeat Them. *Infection and Immunity*, *88*(7), e00916-19.

Arbues, A., Aguilo, J. I., Gonzalo-Asensio, J., Marinova, D., Uranga, S., Puentes, E., Fernandez, C., Parra, A., Cardona, P. J., Vilaplana, C., Ausina, V., Williams, A., Clark, S., Malaga, W., Guilhot, C., Gicquel, B., & Martin, C. (2013). Construction, characterization and preclinical evaluation of MTBVAC, the first live-attenuated M. tuberculosis-based vaccine to enter clinical trials. *Vaccine*, *31*(42), 4867-4873.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, *44*(W1), 344-350.

Assis, P. A., Espíndola, M. S., Paula-Silva, F. W., Rios, W. M., Pereira, P. A., Leão, S. C., Silva, C. L., & Faccioli, L. H. (2014). *Mycobacterium tuberculosis* expressing phospholipase C subverts PGE2 synthesis and induces necrosis in alveolar macrophages. *BMC Microbiology*, *14*(1), 1-10.

Ates, L. S., van der Woude, A. D., Bestebroer, J., van Stempvoort, G., Musters, R. J. P., Garcia-Vallejo, J. J., Picavet, D. I., Weerd, R. van de, Maletta, M., Kuijl, C. P., van der Wel, N. N., & Bitter, W. (2016). The ESX-5 System of Pathogenic Mycobacteria Is Involved in Capsule Integrity and Virulence through Its Substrate PPE10. *PLOS Pathogens*, *12*(6), e1005696.

Bachmair, A., Finley, D., & Varshavsky, A. (1986). In vivo half-life of a protein is a function of its amino-terminal residue. *Science*, *234*(4773), 179-186.

Baddam, R., Kumar, N., Wieler, L. H., Lankapalli, A. K., Ahmed, N., Peacock, S. J., & Semmler, T. (2018). Analysis of mutations in *pncA* reveals non-overlapping patterns among various lineages of *Mycobacterium tuberculosis*. *Scientific Reports*, *8*(1), 1-9.

Bagos, P. G., Nikolaou, E. P., Liakopoulos, T. D., & Tsirigos, K. D. (2010). Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics*, *26*(22), 2811-2817.

Bagos, P. G., Tsirigos, K. D., Liakopoulos, T. D., & Hamodrakas, S. J. (2008). Prediction of Lipoprotein Signal Peptides in Gram-Positive Bacteria with a Hidden Markov Model. *Journal of Proteome Research*, *7*(12), 5082-5093.

Banu, S., Honoré, N., Saint-Joanis, B., Philpott, D., Prévost, M.-C., & Cole, S. T. (2002). Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Molecular Microbiology*, *44*(1), 9-19.

Barreto, M. L., Pereira, S. M., & Ferreira, A. A. (2006). BCG vaccine: efficacy and indications for vaccination and revaccination. *Jornal de Pediatria*, *82*(7), 45-54.

Bendtsen, J. D., Kiemer, L., Fausbøll, A., & Brunak, S. (2005). Non-classical protein secretion in bacteria. *BMC Microbiology*, *5*(1), 1-13.

Bennett, R. M. (2009). Farm costs associated with premovement testing for bovine tuberculosis. *Veterinary Record*, *164*(3), 77-79.

Bertrand, T., Eady, N. A. J., Jones, J. N., Jesmin, Nagy, J. M., Jamart-Grégoire, B., Raven, E. L., & Brown, K. A. (2004). Crystal Structure of *Mycobacterium tuberculosis* Catalase-Peroxidase. *Journal of Biological Chemistry*, *279*(37), 38991-38999.

Bhat, Z. S., Rather, M. A., Maqbool, M., & Ahmad, Z. (2018). Drug targets exploited in *Mycobacterium tuberculosis*: Pitfalls and promises on the horizon. *Biomedicine & Pharmacotherapy*, *103*, 1733-1747.

Bhavsar, A. P., Guttman, J. A., & Finlay, B. B. (2007). Manipulation of host-cell pathways by bacterial pathogens. *Nature*, *449*(7164), 827-834.

Biffa, D., Bogale, A., Godfroid, J., & Skjerve, E. (2012). Factors associated with severity of bovine tuberculosis in Ethiopian cattle. *Tropical Animal Health and Production*, *44*(5), 991-998.

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, *41*(W1), 29-33.

Bowie, J., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, *253*(5016), 164-170.

Braibant, M., Gilot, P., & Content, J. (2000). The ATP binding cassette (ABC) transport systems of *Mycobacterium tuberculosis*. *FEMS Microbiology Reviews*, *24*(4), 449-467.

Brazier, B., & McShane, H. (2020). Towards new TB vaccines. *Seminars in Immunopathology*, *42*(3), 315-331.

Brennan, P. J., & Nikaido, H. (1995). The Envelope of Mycobacteria. *Annual Review of Biochemistry*, *64*(1), 29-63.

Brewer, T. F. (2000). Preventing Tuberculosis with Bacillus Calmette-Guérin Vaccine: A Meta-Analysis of the Literature. *Clinical Infectious Diseases*, *31*(Supplement_3), 64-67.

Briassoulis, G., Karabatsou, I., Gogoglou, V., & Tsorva, A. (2005). BCG vaccination at three different age groups: response and effectiveness. *Journal of Immune Based Therapies and Vaccines*, *3*(1), 1-12.

Brosch, R., Gordon, S. v., Garnier, T., Eiglmeier, K., Frigui, W., Valenti, P., dos Santos, S., Duthoy, S., Lacroix, C., Garcia-Pelayo, C., Inwald, J. K., Golby, P., Garcia, J. N., Hewinson, R. G., Behr, M. A., Quail, M. A.,

Churcher, C., Barrell, B. G., Parkhill, J., & Cole, S. T. (2007). Genome plasticity of BCG and impact on vaccine efficacy. *Proceedings of the National Academy of Sciences*, *104*(13), 5596-5601.

Brosch, R., Gordon, S. v., Pym, A., Eiglmeier, K., Garnier, T., & Cole, S. T. (2000). Comparative genomics of the mycobacteria. *International Journal of Medical Microbiology*, *290*(2), 143-152.

Brusic, V., & Flower, D. R. (2004). Bioinformatics tools for identifying T-cell epitopes. *Drug Discovery Today: BIOSILICO*, *2*(1), 18-23.

Brylinski, M., & Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences*, *105*(1), 129-134.

Buchan, D. W. A., & Jones, D. T. (2019). The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Research*, *47*(W1), 402-407.

Buddle, B. M. (2010). Tuberculosis vaccines for cattle: the way forward. *Expert Review of Vaccines*, *9*(10), 1121-1124.

Buddle, B. M., Vordermeier, H. M., Chambers, M. A., & de Klerk-Lorist, L. M. (2018). Efficacy and Safety of BCG Vaccine for Control of Tuberculosis in Domestic Livestock and Wildlife. *Frontiers in Veterinary Science*, *5*, 1-17.

Bui, H.-H., Sidney, J., Dinh, K., Southwood, S., Newman, M. J., & Sette, A. (2006). Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics*, *7*(1), 1-5.

Bulashevska, A., & Eils, R. (2006). Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinformatics*, *7*(1), 1-13.

Bulut, Y., Michelsen, K. S., Hayrapetian, L., Naiki, Y., Spallek, R., Singh, M., & Arditi, M. (2005). *Mycobacterium Tuberculosis* Heat Shock Proteins Use Diverse Toll-like Receptor Pathways to Activate Pro-inflammatory Signals. *Journal of Biological Chemistry*, *280*(22), 20961-20967.

Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M., & Funkhouser, T. A. (2009). Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLoS Computational Biology*, *5*(12), e1000585.

Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, *33*(Web Server). W306-310.

Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., & Ben-Tal, N. (2013). ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Israel Journal of Chemistry*, *53*(3–4), 199-206.

Chan, E., & Iseman, M. (2002). Current medical treatment for tuberculosis. *BMJ*, *325*(7375), 1282-1286.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2010). *MolProbity* : all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography*, *66*(1), 12-21.

Cheng, A. F. B., Yew, W. W., Chan, E. W. C., Chin, M. L., Hui, M. M. M., & Chan, R. C. Y. (2004). Multiplex PCR Amplimer Conformation Analysis for Rapid Detection of *gyrA* Mutations in Fluoroquinolone-Resistant *Mycobacterium tuberculosis* Clinical Isolates. *Antimicrobial Agents and Chemotherapy*, *48*(2), 596-601.

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, *31*(16), 2745-2747.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, *7*(10), e46688

Churchyard, G., Kim, P., Shah, N. S., Rustomjee, R., Gandhi, N., Mathema, B., Dowdy, D., Kasmar, A., & Cardenas, V. (2017). What We Know About Tuberculosis Transmission: An Overview. *The Journal of Infectious Diseases*, *216*(suppl_6), S629-635.

Clemens, D. L., Lee, B.-Y., & Horwitz, M. A. (2000). Deviant Expression of Rab5 on Phagosomes Containing the Intracellular Pathogens *Mycobacterium tuberculosis* and *Legionella pneumophila* Is Associated with Altered Phagosomal Fate. *Infection and Immunity*, *68*(5), 2671-2684.

Cohen, K. A., Bishai, W. R., & Pym, A. S. (2014). Molecular Basis of Drug Resistance in *Mycobacterium tuberculosis*. *Microbiology Spectrum*, *2*(3).

Colditz, G., Brewer, T., Berkey, C., Wilson, M., Burdick, E., Fineberg, H., & Mosteller, F. (1994). Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *JAMA*, *271*(9), 698–702.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. v., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., … Barrell, B. G. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, *393*(6685), 537-544.

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science*, *2*(9), 1511-1519.

Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., Galagan, J., Niemann, S., & Gagneux, S. (2012). Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics*, *44*(1), 106-110.

Corner, L., & Norton, S. (2003). Resolution of *Mycobacterium bovis* infection in wild brushtail possums (*Trichosurus vulpecula*). *New Zealand Veterinary Journal*, *51*(1), 40-42.

Cosivi, O. (1998). Zoonotic Tuberculosis due to *Mycobacterium bovis* in Developing Countries. *Emerging Infectious Diseases*, *4*(1), 59-70.

Cousins, D. V. (2001). *Mycobacterium bovis* infection and control in domestic livestock. *Revue Scientifique et Technique de l'OIE*, *20*(1), 71-85.

Covián, C., Fernández-Fierro, A., Retamal-Díaz, A., Díaz, F. E., Vasquez, A. E., Lay, M. K., Riedel, C. A., González, P. A., Bueno, S. M., & Kalergis, A. M. (2019). BCG-Induced Cross-Protection and Development of Trained Immunity: Implication for Vaccine Design. *Frontiers in Immunology*, *10*.

Crofton, J., & Mitchison, D. A. (1948). Streptomycin Resistance in Pulmonary Tuberculosis. *BMJ*, *2*(4588), 1009-1015.

da Cunha, E. F. F., Ramalho, T. C., de Alencastro, R. B., & Maia, E. R. (2007). Docking Simulations and QM/MM Studies between Isoniazid Prodrug, Catalase-Peroxidase (*KatG*) and S315T Mutant from *Mycobacterium tuberculosis*. *Computational and Mathematical Methods in Medicine*, *8*(2), 113-124.

Daniel, T. (2006). The history of tuberculosis. *Respiratory Medicine*, *100*(11), 1862-1870.

Daniel, T. (2011). Hermann Brehmer and the origins of tuberculosis sanatoria. *International Journal of Tuberculosis and Lung Disease*, *15*(2), 161–162.

Davies, P. D. O. (2006). Tuberculosis in humans and animals: are we a threat to each other? *Journal of the Royal Society of Medicine*, *99*(10), 539-540.

de Groot, A. S., & Martin, W. (2009). Reducing risk, improving outcomes: Bioengineering less immunogenic protein therapeutics. *Clinical Immunology*, *131*(2), 189-201.

de Rycker, M., Baragaña, B., Duce, S. L., & Gilbert, I. H. (2018). Challenges and recent progress in drug discovery for tropical diseases. *Nature*, *559*(7715), 498-506.

Deenadayalan, A., Maddineni, P., & Raja, A. (2013). Comparison of whole blood and PBMC assays for T-cell functional analysis. *BMC Research Notes*, *6*(1), 1-5.

Deng, Mikusová, K., Robuck, K. G., Scherman, M., Brennan, P. J., & McNeil, M. R. (1995). Recognition of multiple effects of ethambutol on metabolism of mycobacterial cell envelope. *Antimicrobial Agents and Chemotherapy*, *39*(3), 694-701.

Deng, W., Zeng, J., Xiang, X., Li, P., & Xie, J. (2015). PE11 (Rv1169c) selectively alters fatty acid components of *Mycobacterium smegmatis* and host cell interleukin-6 level accompanied with cell death. *Frontiers in Microbiology*, *6*.

Dheda, K., Schwander, S. K., Zhu, B., van Zyl-Smit, R. N., & Zhang, Y. (2010). The immunology of tuberculosis: From bench to bedside. *Respirology*, *15*(3), 433-450.

Dimitrov, I., Bangov, I., Flower, D. R., & Doytchinova, I. (2014). AllerTOP v.2—a server for in silico prediction of allergens. *Journal of Molecular Modeling*, *20*(6), 2278-2284.

Dockrell, H. M., & Smith, S. G. (2017). What Have We Learnt about BCG Vaccination in the Last 20 Years? *Frontiers in Immunology*, *8*.

Doherty, T. M., & Andersen, P. (2005). Vaccines for Tuberculosis: Novel Concepts and Recent Progress. *Clinical Microbiology Reviews*, *18*(4), 687-702.

Donati, C., & Rappuoli, R. (2013). Reverse vaccinology in the 21st century: improvements over the original design. *Annals of the New York Academy of Sciences*, *1285*(1), 115-132.

Doolan, D. L., Apte, S. H., & Proietti, C. (2014). Genome-based vaccine design: the promise for malaria and other infectious diseases. *International Journal for Parasitology*, *44*(12), 901-913.

Dooley, K. E., Obuku, E. A., Durakovic, N., Belitsky, V., Mitnick, C., & Nuermberger, E. L. (2013). World Health Organization Group 5 Drugs for the Treatment of Drug-Resistant Tuberculosis: Unclear Efficacy or Untapped Potential? *Journal of Infectious Diseases*, *207*(9),1352-1358.

Doytchinova, I. A., & Flower, D. R. (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, *8*(1), 1-7.

EL-Manzalawy, Y., & Honavar, V. (2010). Recent advances in B-cell epitope prediction methods. *Immunome Research*, *6*(Suppl 2), 1-9.

Eswar, N., Eramian, D., Webb, B., Shen, M. Y., & Sali, A. (2008). Protein Structure Modeling with MODELLER. *Structural Proteomics*, *426*, 145-159.

Ferber, D. (2005). Biochemistry: Protein That Mimics DNA Helps Tuberculosis Bacteria Resist Antibiotics. *Science*, *308*(5727). 1393.

Ferluga, J., Yasmin, H., Al-Ahdal, M. N., Bhakta, S., & Kishore, U. (2020). Natural and trained innate immunity against *Mycobacterium tuberculosis*. *Immunobiology*, *225*(3).

Ferrari, G., Langen, H., Naito, M., & Pieters, J. (1999). A Coat Protein on Phagosomes Involved in the Intracellular Survival of Mycobacteria. *Cell*, *97*(4), 435-447.

Ferreira, L., dos Santos, R., Oliva, G., & Andricopulo, A. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*, *20*(7), 13384-13421.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*(D1), D222-230.

Fishbein, S., van Wyk, N., Warren, R. M., & Sampson, S. L. (2015). Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Molecular Microbiology*, *96*(5), 901-916.

Flower, D. R., Macdonald, I. K., Ramakrishnan, K., Davies, M. N., & Doytchinova, I. A. (2010). Computer aided selection of candidate vaccine antigens. *Immunome Research*, *6*(Suppl 2), 1-16.

Flynn, J. L. (2004). Immunology of tuberculosis and implications in vaccine development. *Tuberculosis*, *84*(1–2), 93-101.

Flynn, J. L., & Chan, J. (2001). Immunology of Tuberculosis. *Annual Review of Immunology*, *19*(1), 93-129.

Frieden, T. (2007). Promoting adherence to treatment for tuberculosis: the importance of direct observation. *Bulletin of the World Health Organization*, *85*(5), 407-409.

Gagneux, S., Long, C., Small, P., Van, T., Schoolnik, G., & Bohannan, B. (2006). The Competitive Cost of Antibiotic Resistance in *Mycobacterium tuberculosis*. *Science*, *312*(5782), 1944-1946.

Garg, A., & Gupta, D. (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics*, *9*(1), 1-12.

Garnier, T., Eiglmeier, K., Camus, J.-C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., Grondin, S., Lacroix, C., Monsempe, C., Simon, S., Harris, B., Atkin, R., Doggett, J., Mayes, R., Keating, L., Wheeler, P. R., Parkhill, J., Barrell, B. G., … Hewinson, R. G. (2003). The complete genome sequence of Mycobacterium bovis. *Proceedings of the National Academy of Sciences*, *100*(13), 7877-7882.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M., Appel, R., & Bairoch, A. (2005). *The Proteomics Protocols Handbook* (J. Walker, Ed.). Humana Press.

Georghiou, S. B., Magana, M., Garfein, R. S., Catanzaro, D. G., Catanzaro, A., & Rodwell, T. C. (2012). Evaluation of Genetic Mutations Associated with *Mycobacterium tuberculosis* Resistance to Amikacin, Kanamycin and Capreomycin: A Systematic Review. *PLoS ONE*, *7*(3). e33275

Gershoni, J. M., Roitburd-Berman, A., Siman-Tov, D. D., Tarnovitski Freund, N., & Weiss, Y. (2007). Epitope Mapping. *BioDrugs*, *21*(3), 141-156.

Ghaffari-Nazari, H., Tavakkol-Afshari, J., Jaafari, M. R., Tahaghoghi-Hajghorbani, S., Masoumi, E., & Jalali, S. A. (2015). Improving Multi-Epitope Long Peptide Vaccine Potency by Using a Strategy that Enhances CD4+ T Help in BALB/c Mice. *PLOS ONE*, *10*(11), 1-12.

Giguère, S., Drouin, A., Lacoste, A., Marchand, M., Corbeil, J., & Laviolette, F. (2013). MHC-NP: Predicting peptides naturally processed by the MHC. *Journal of Immunological Methods*, *400–401*, 30-36.

Gill, S. C., & von Hippel, P. H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Analytical Biochemistry*, *182*(2), 319-326.

Ginsberg, A. M., & Spigelman, M. (2007). Challenges in tuberculosis drug research and development. *Nature Medicine*, *13*(3), 290-294.

Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., Altermann, U., Angerer, P., Ansorge, S., Balasz, K., Bernhofer, M., Betz, A., Cizmadija, L., Do, K., Gerke, J., Greil, R., Joerdens, V., Hastreiter, M., Hembach, K., … Rost, B. (2014). LocTree3 prediction of localization. *Nucleic Acids Research*, *42*(W1), W350-355.

Good, M., & Duignan, A. (2011). Perspectives on the History of Bovine TB and the Role of Tuberculin in Bovine TB Eradication. *Veterinary Medicine International*, *410470,* 1-11.

Goren, M. B. (1977). Phagocyte Lysosomes: Interactions with Infectious Agents, Phagosomes, and Experimental Perturbations in Function. *Annual Review of Microbiology*, *31*(1), 507-533.

Goulding, C. W., Parseghian, A., Sawaya, M. R., Cascio, D., Apostol, M. I., Gennaro, M. L., & Eisenberg, D. (2009). Crystal structure of a major secreted protein of Mycobacterium tuberculosis-MPT63 at 1.5-Å resolution. *Protein Science*, *11*(12), 2887-2893.

Grandi, G. (2001). Antibacterial vaccine design using genomics and proteomics. *Trends in Biotechnology*, *19*(5), 181-188.

Grange, J., & Stanford, J. (1994). Dogma and innovation in the global control of tuberculosis: discussion paper. *Journal of Royal Society of Medicine*, *87*(5), 272–275.

Gulukota, K. (2008). Immunoinformatics in Personalized Medicine. *Novartis Foundation Symposium*, *254*,43-50.

Gupta, A., Kapil, R., Dhakan, D., & Sharma, V. (2014). MP3: A Software Tool for the Prediction of Pathogenic Proteins in Genomic and Metagenomic Data. *PLoS ONE*, *9*(4), e93907.

Gupta, S., Shenoy, V. P., Mukhopadhyay, C., Bairy, I., & Muralidharan, S. (2011). Role of risk factors and socio-economic status in pulmonary tuberculosis: a search for the root cause in patients in a tertiary care hospital, South India. *Tropical Medicine & International Health*, *16*(1), 74-78.

Haake, D. A., & Levett, P. N. (2015). Leptospirosis in Humans. *Current Topics in Microbiology and Immunology*, *387*, 65-97.

Hardy, A. (1999). Captain of death: the story of tuberculosis. *Medical History*, *43*(1).

He, Y., Rappuoli, R., de Groot, A. S., & Chen, R. T. (2010). Emerging Vaccine Informatics. *Journal of Biomedicine and Biotechnology*, *218590*, 1-26.

Herr, H. W., & Morales, A. (2008). History of Bacillus Calmette-Guerin and Bladder Cancer: An Immunotherapy Success Story. *Journal of Urology*, *179*(1), 53-56.

Hess, B., Kutzner, C., van der Spoel, D., & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, *4*(3), 435-447.

Hesseling, A. C., Marais, B. J., Gie, R. P., Schaaf, H. S., Fine, P. E. M., Godfrey-Faussett, P., & Beyers, N. (2007). The risk of disseminated Bacille Calmette-Guerin (BCG) disease in HIV-infected children. *Vaccine*, *25*(1), 14-18.

Hestvik, A. L. K., Hmama, Z., & Av-Gay, Y. (2005). Mycobacterial manipulation of the host cell. *FEMS Microbiology Reviews*, *29*(5), 1041-1050.

Heym, B., Alzari, P. M., Honore, N., & Cole, S. T. (1995). Missense mutations in the catalase-peroxidase gene, *KatG*, are associated with isoniazid resistance in *Mycobacterium tuberculosis*. *Molecular Microbiology*, *15*(2), 235-245.

Heym, B., Honoré, N., Schurra, C., Cole, S. T., Heym, B., Truffot-Pernot, C., Grosset, J. H., Banerjee, A., Jacobs, W. R., & van Embden, J. D. A. (1994). Implications of multidrug resistance for the future of short-course chemotherapy of tuberculosis: a molecular study. *The Lancet*, *344*(8918), 293-298.

Hirano, K., Takahashi, M., Kazumi, Y., Fukasawa, Y., & Abe, C. (1998). Mutation in *pncA* is a major mechanism of pyrazinamide resistance in *Mycobacterium tuberculosis*. *Tubercle and Lung Disease*, *78*(2). 117-122.

Hoagland, D. T., Liu, J., Lee, R. B., & Lee, R. E. (2016). New agents for the treatment of drug-resistant *Mycobacterium tuberculosis*. *Advanced Drug Delivery Reviews*, *102*, 55-72.

Hoffmann, J. A. (1999). Phylogenetic Perspectives in Innate Immunity. *Science*, *284*(5418), 1313-1318.

Hong, B.-Y., Maulén, N. P., Adami, A. J., Granados, H., Balcells, M. E., & Cervantes, J. (2016). Microbiome Changes during Tuberculosis and Antituberculous Therapy. *Clinical Microbiology Reviews*, *29*(4), 915-926.

Hossain, Md. S., Azad, A. K., Chowdhury, P. A., & Wakayama, M. (2017). Computational Identification and Characterization of a Promiscuous T-Cell Epitope on the Extracellular Protein 85B of *Mycobacterium* spp. for Peptide-Based Subunit Vaccine Design. *BioMed Research International*, *4826030*, 1-14.

Houston, M. (1999). The White Death: A History of Tuberculosis. *BMJ*, *318*(7199).

Ikai, A. J. (1980). Thermostability and aliphatic index of globular proteins. *Journal of Biochemistry*, *88*(6), 1895–1898.

Imai, K., Asakawa, N., Tsuji, T., Akazawa, F., Ino, A., Sonoyama, M., & Mitaku, S. (2008). SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria. *Bioinformation*, *2*(9), 417-421.

Jarlier, V., & Nikaido, H. (1994). Mycobacterial cell wall: Structure and role in natural resistance to antibiotics. *FEMS Microbiology Letters*, *123*(1–2), 11-18.

Jasmer, R. M., Saukkonen, J. J., Blumberg, H. M., Daley, C. L., Bernardo, J., Vittinghoff, E., King, M. D., Kawamura, L. M., & Hopewell, P. C. (2002). Short-Course Rifampin and Pyrazinamide Compared with Isoniazid for Latent Tuberculosis Infection: A Multicenter Clinical Trial. *Annals of Internal Medicine*, *137*(8), 640-647.

Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., Sette, A., Peters, B., & Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*, *154*(3), 394-406.

Jerant, A., Bannon, M., & Rittenhouse, S. (2000). Identification and management of tuberculosis. *American Family Physician*, *61*(9), 2681–2682.

Johnsson, K., & Schultz, P. (1994). Mechanistic Studies of the Oxidation of Isoniazid by the Catalase Peroxidase from Mycobacterium tuberculosis. *Journal of the American Chemical Society*, *116*(16), 7425-7426.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236-1240.

Joshi, J. (2011). Tuberculosis chemotherapy in the 21 st century: Back to the basics. *Lung India*, *28*(3).

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353-361.

Karakousis, P. C., Williams, E. P., & Bishai, W. R. (2008). Altered expression of isoniazid-regulated genes in drug-treated dormant Mycobacterium tuberculosis. *Journal of Antimicrobial Chemotherapy*, *61*(2), 323-331.

Katoch, V. (2004). Newer diagnostic techniques for tuberculosis. *The Indian Journal of Medical Research*, *120*(4), 418–428.

Kaufmann, S. H. E. (2013). Tuberculosis vaccines: Time to think about the next generation. *Seminars in Immunology*, *25*(2), 172-181.

Kennedy, D. A., & Read, A. F. (2017). Why does drug resistance readily evolve but vaccine resistance does not? *Proceedings of the Royal Society B: Biological Sciences*, *284*(1851).

Keshavjee, S., & Farmer, P. (2012). Tuberculosis, Drug Resistance, and the History of Modern Medicine. *New England Journal of Medicine*, *367*(10), 931-936.

Khubaib, M., Sheikh, J. A., Pandey, S., Srikanth, B., Bhuwan, M., Khan, N., Hasnain, S. E., & Ehtesham, N. Z. (2016). Mycobacterium tuberculosis Co-operonic PE32/PPE65 Proteins Alter Host Immune Responses by Hampering Th1 Response. *Frontiers in Microbiology*, *7*.

Kleinnijenhuis, J., Oosting, M., Joosten, L. A. B., Netea, M. G., & van Crevel, R. (2011). Innate Immune Recognition of *Mycobacterium tuberculosis*. *Clinical and Developmental Immunology*, *405310*, 1-12.

Knechel, N. A. (2009). Tuberculosis: Pathophysiology, Clinical Features, and Diagnosis. *Critical Care Nurse*, *29*(2), 34-43.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., & Wishart, D. S. (2011). DrugBank 3.0: a comprehensive resource for "Omics" research on drugs. *Nucleic Acids Research*, *39*(Database). D1035-1041.

Ko, J., Park, H., Heo, L., & Seok, C. (2012). GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Research*, *40*(W1), W294-297.

Koch, A., Mizrahi, V., & Warner, D. F. (2014). The impact of drug resistance on *Mycobacterium tuberculosis* physiology: what can we learn from rifampicin? *Emerging Microbes & Infections*, *3*(1), 1-11.

Kopp, J., Bordoli, L., Battey, J. N. D., Kiefer, F., & Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins: Structure, Function, and Bioinformatics*, *69*(S8), 38-56.

Kovacs-Simon, A., Titball, R. W., & Michell, S. L. (2011). Lipoproteins of Bacterial Pathogens. *Infection and Immunity*, *79*(2), 548-561.

Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D., & Vajda, S. (2017). The ClusPro web server for protein–protein docking. *Nature Protocols*, *12*(2), 255-278.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*, *305*(3), 567-580.

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073-1081.

Kumar, S., & Jena, L. (2014). Understanding Rifampicin Resistance in Tuberculosis through a Computational Approach. *Genomics & Informatics*, *12*(4), 276-282.

Kuria, J. K. (2019). Diseases Caused by Bacteria in Cattle: Tuberculosis. *Bacterial Cattle Diseases*. IntechOpen.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105-132.

Lahariya, C. (2014). A brief history of vaccines & vaccination in India. *The Indian Journal of Medica Research*, *139*(4), 491–511.

Lakshmipathy, D., Ramasubban, G., Therese, L., Vetrivel, U., Muthukumaran, S., & Rajendiran, S. (2013). In silico Analysis of Novel Mutation ala102pro Targeting *pncA* Gene of M. Tuberculosis. *Journal of Computer Science & Systems Biology*, *06*(02), 083–087.

Lamichhane, G. (2011). Novel targets in M. tuberculosis: search for new drugs. *Trends in Molecular Medicine*, *17*(1), 25-33.

Lange, C., Chesov, D., Heyckendorf, J., Leung, C. C., Udwadia, Z., & Dheda, K. (2018). Drug-resistant tuberculosis: An update on disease burden, diagnosis and treatment. *Respirology*, *23*(7), 656-673.

Lange, C., & Mori, T. (2010). Advances in the diagnosis of tuberculosis. *Respirology*, *15*(2), 220-240.

Larsen, J., Lund, O., & Nielsen, M. (2006). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Immunome Research*, *2*(1).

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, *26*(2), 283-291.

Lemos, A. C. M., & Matos, E. D. (2013). Multidrug-resistant tuberculosis. *The Brazilian Journal of Infectious Diseases*, *17*(2), 239-246.

Letunic, I., & Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research*, *46*(D1), D493-496.

Lindestam Arlehamn, C. S., Paul, S., Mele, F., Huang, C., Greenbaum, J. A., Vita, R., Sidney, J., Peters, B., Sallusto, F., & Sette, A. (2015). Immunological consequences of intragenus conservation of *Mycobacterium tuberculosis* T-cell epitopes. *Proceedings of the National Academy of Sciences*, *112*(2), E147-155.

Lipsitch, M., & Siber, G. R. (2016). How Can Vaccines Contribute to Solving the Antimicrobial Resistance Problem? *MBio*, *7*(3).

López-Blanco, J. R., Aliaga, J. I., Quintana-Ortí, E. S., & Chacón, P. (2014). iMODS: internal coordinates normal mode analysis server. *Nucleic Acids Research*, *42*(W1), W271-276.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S., & Richardson, D. C. (2003). Structure validation by Cα geometry: φ,ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics*, *50*(3), 437-450.

Luabeya, A. K. K., Kagina, B. M. N., Tameris, M. D., Geldenhuys, H., Hoff, S. T., Shi, Z., Kromann, I., Hatherill, M., Mahomed, H., Hanekom, W. A., Andersen, P., Scriba, T. J., Schoeman, E., Krohn, C., Day, C. L., Africa, H., Makhethe, L., Smit, E., Brown, Y., … Hussey, G. D. (2015). First-in-human trial of the post-exposure tuberculosis vaccine H56:IC31 in Mycobacterium tuberculosis infected and non-infected healthy adults. *Vaccine*, *33*(33), 4130-4140.

Mariam, D. H., Mengistu, Y., Hoffner, S. E., & Andersson, D. I. (2004). Effect of *rpoB* Mutations Conferring Rifampin Resistance on Fitness of *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy*, *48*(4).

Marshall, J. S., Warrington, R., Watson, W., & Kim, H. L. (2018). An introduction to immunology and immunopathology. *Allergy, Asthma & Clinical Immunology*, *14*(S2), 5-14.

Mashiach, E., Schneidman-Duhovny, D., Andrusier, N., Nussinov, R., & Wolfson, H. J. (2008). FireDock: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Research*, *36*(Web Server), W229-232.

Mazandu, G. K., & Mulder, N. J. (2012). Function Prediction and Analysis of Mycobacterium tuberculosis Hypothetical Proteins. *International Journal of Molecular Sciences*, *13*(6), 7283-7302.

McGrath, M., Gey van Pittius, N. C., van Helden, P. D., Warren, R. M., & Warner, D. F. (2014). Mutation rate and the emergence of drug resistance in *Mycobacterium tuberculosis*. *Journal of Antimicrobial Chemotherapy*, *69*(2), 292-302.

McLaughlin, B., Chon, J. S., MacGurn, J. A., Carlsson, F., Cheng, T. L., Cox, J. S., & Brown, E. J. (2007). A Mycobacterium ESX-1–Secreted Virulence Factor with Unique Requirements for Export. *PLoS Pathogens*, *3*(8), e105.

McMaster, W. G., Kirabo, A., Madhur, M. S., & Harrison, D. G. (2015). Inflammation, Immunity, and Hypertensive End-Organ Damage. *Circulation Research*, *116*(6), 1022-1033.

McNerney, R., Maeurer, M., Abubakar, I., Marais, B., Mchugh, T. D., Ford, N., Weyer, K., Lawn, S., Grobusch, M. P., Memish, Z., Squire, S. B., Pantaleo, G., Chakaya, J., Casenghi, M., Migliori, G.-B., Mwaba, P., Zijenah, L., Hoelscher, M., Cox, H., … Zumla, A. (2012). Tuberculosis Diagnostics and Biomarkers: Needs, Challenges, Recent Advances, and Opportunities. *The Journal of Infectious Diseases*, *205*(suppl_2), S147-158.

McShane, H. (2014). Editorial Commentary: Understanding BCG Is the Key to Improving It. *Clinical Infectious Diseases*, *58*(4), 481-482.

Medzhitov, R., & Janeway, C. (2000). Innate Immunity. *New England Journal of Medicine*, *343*(5), 338-344.

Mehta, P. K., Raj, A., Singh, N., & Khuller, G. K. (2012). Diagnosis of extrapulmonary tuberculosis by PCR. *FEMS Immunology & Medical Microbiology*, *66*(1), 20-36.

Melo, F., Sánchez, R., & Sali, A. (2009). Statistical potentials for fold assessment. *Protein Science*, *11*(2), 430-448.

Middlebrook, G. (1954). Isoniazid-resistance and catalase activity of tubercle bacilli; a preliminary report. *American Review of Tuberculosis*, *69*(3), 471–472.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., … Finn, R. D. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, *47*(D1), D351-360.

Mitchison, D. A. (1985). The action of antituberculosis drugs in short-course chemotherapy. *Tubercle*, *66*(3), 219-225.

Moliva, J. I., Turner, J., & Torrelles, J. B. (2017). Immune Responses to Bacillus Calmette–Guérin Vaccination: Why Do They Fail to Protect against Mycobacterium tuberculosis? *Frontiers in Immunology*, *8*.

Monterrubio-López, G. P., González-Y-Merchand, J. A., & Ribas-Aparicio, R. M. (2015). Identification of Novel Potential Vaccine Candidates against Tuberculosis Based on Reverse Vaccinology. *BioMed Research International*, *483150*, 1-16.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, *35*(Web Server), W182-185.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, *19*(14), 1639-1662.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, *30*(16), 2785-2791.

Movahedi, A. (Reza), & Hampson, D. J. (2008). New ways to identify novel bacterial antigens for vaccine development. *Veterinary Microbiology*, *131*(1–2), 1-13.

Murray, C., & Lopez, A. (1994). Global and regional cause-of-death patterns in 1990. *Bulletin of the World Health Organization*, *72*(3), 447–480.

Nachega, J. B., & Chaisson, R. E. (2003). Tuberculosis Drug Resistance: A Global Threat. *Clinical Infectious Diseases*, *36*(Supplement_1), S24-30.

Nagabhushanam, V., Solache, A., Ting, L.-M., Escaron, C. J., Zhang, J. Y., & Ernst, J. D. (2003). Innate Inhibition of Adaptive Immunity: *Mycobacterium tuberculosis* -Induced IL-6 Inhibits Macrophage Responses to IFN-γ. *The Journal of Immunology*, *171*(9), 4750-4757.

Nakajima, C., Rahim, Z., Fukushima, Y., Sugawara, I., van der Zanden, A. G., Tamaru, A., & Suzuki, Y. (2010). Identification of Mycobacterium tuberculosis clinical isolates in Bangladesh by a species distinguishable multiplex PCR. *BMC Infectious Diseases*, *10*(1), 1-7.

Narita, M., Ashkin, D., Hollender, E., & Pitchenik, A. (1998). Paradoxical Worsening of Tuberculosis Following Antiretroviral Therapy in Patients with AIDS. *American Journal of Respiratory and Critical Care Medicine*, *158*(1), 157-161.

Naveen, A., Araga, U., & Narasu, M. (2014). Vaccinology: Reflections and the way forward. *Journal of Medical & Allied Sciences*, *4*(1), 45–53.

Ng, P. C., & Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. *Genome Research*, *11*(5), 863-874.

Nguyen, L. (2016). Antibiotic resistance mechanisms in M. tuberculosis: an update. *Archives of Toxicology*, *90*(7), 1585-1604.

Nugent, G., Gormley, A. M., Anderson, D. P., & Crews, K. (2018). Roll-Back Eradication of Bovine Tuberculosis (TB) From Wildlife in New Zealand: Concepts, Evolving Approaches, and Progress. *Frontiers in Veterinary Science*, *5*.

O'Reilly, L. M., & Daborn, C. J. (1995). The epidemiology of Mycobacterium bovis infections in animals and man: A review. *Tubercle and Lung Disease*, *76*, 1-46.

Ortega-Tirado, D., Niño-Padilla, E. I., Arvizu-Flores, A. A., Velazquez, C., Espitia, C., Serrano, C. J., Enciso-Moreno, J. A., Sumoza-Toledo, A., & Garibay-Escobar, A. (2020). Identification of immunogenic T-cell peptides of Mycobacterium tuberculosis PE_PGRS33 protein. *Molecular Immunology*, *125*, 123-130.

Ortiz, A. R., Strauss, C. E. M., & Olmea, O. (2009). MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Science*, *11*(11), 2606-2621.

Ottenhoff, T. H. M., & Kaufmann, S. H. E. (2012). Vaccines against Tuberculosis: Where Are We and Where Do We Need to Go? *PLoS Pathogens*, *8*(5), e1002607.

Pal, R., Ansari, M. A., Hameed, S., & Fatima, Z. (2016). Diabetes Mellitus as Hub for Tuberculosis Infection: A Snapshot. *International Journal of Chronic Diseases*, *5981574*, 1-7.

Palmer, M. v., & Thacker, T. C. (2018). Use of the Human Vaccine, Mycobacterium bovis Bacillus Calmette Guérin in Deer. *Frontiers in Veterinary Science*, *5*.

Pang, Y., Lu, J., Wang, Y., Song, Y., Wang, S., & Zhao, Y. (2013). Study of the Rifampin Monoresistance Mechanism in Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy*, *57*(2).

Parlane, N. A., & Buddle, B. M. (2015). Immunity and Vaccination against Tuberculosis in Cattle. *Current Clinical Microbiology Reports*, *2*(1), 44-53.

Pasquale, A., Preiss, S., Silva, F., & Garçon, N. (2015). Vaccine Adjuvants: from 1920 to 2015 and Beyond. *Vaccines*, *3*(2), 320-343.

Patronov, A., & Doytchinova, I. (2013). T-cell epitope vaccine design by immunoinformatics. *Open Biology*, *3*(1).

Paulsen, I., Chen, J., Nelson, K., & Saier, M. (2001). Comparative genomics of microbial drug efflux systems. *Journal of Molecular Microbiology and Biotechnology*, *3*(2), 145–150.

Penn-Nicholson, A., Tameris, M., Smit, E., Day, T. A., Musvosvi, M., Jayashankar, L., Vergara, J., Mabwe, S., Bilek, N., Geldenhuys, H., Luabeya, A. K.-K., Ellis, R., Ginsberg, A. M., Hanekom, W. A., Reed, S. G., Coler, R. N., Scriba, T. J., & Hatherill, M. (2018). Safety and immunogenicity of the novel tuberculosis vaccine ID93 + GLA-SE in BCG-vaccinated healthy adults in South Africa: a randomised, double-blind, placebo-controlled phase 1 trial. *The Lancet Respiratory Medicine*, *6*(4), 287-298.

Peñuelas-Urquides, K., Castorena-Torres, F., Ramírez, B. S., & León, M. B. de. (2018). Drug Resistance in Mycobacterium tuberculosis. *Mycobacterium - Research and Development*. InTech.

Pergande, M., & Cologna, S. (2017). Isoelectric Point Separations of Peptides and Proteins. *Proteomes*, *5*(4), 1-14.

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785-786.

Petit, J., Wieetzerbin, J., Das, B., & Lederer. (1975). Chemical structure of the cell wall of Mycobacterium tuberculosis var. bovis, strain BCG. *Z Immunitatsforsch Exp Klin Immunol*, *149*, 118–125.

Petrella, S., Gelus-Ziental, N., Maudry, A., Laurans, C., Boudjelloul, R., & Sougakoff, W. (2011). Crystal Structure of the Pyrazinamidase of Mycobacterium tuberculosis: Insights into Natural and Acquired Resistance to Pyrazinamide. *PLoS ONE*, *6*(1), e15785.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera- A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605-1612.

Piddington, D. L., Fang, F. C., Laessig, T., Cooper, A. M., Orme, I. M., & Buchmeier, N. A. (2001). Cu,Zn Superoxide Dismutase of *Mycobacterium tuberculosis* Contributes to Survival in Activated Macrophages That Are Generating an Oxidative Burst. *Infection and Immunity*, *69*(8).

Pieters, J. (2008). Mycobacterium tuberculosis and the Macrophage: Maintaining a Balance. *Cell Host & Microbe*, *3*(6), 399-407.

Piovesan, D., Minervini, G., & Tosatto, S. C. E. (2016). The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Research*, *44*(W1), W367-374.

Pires, D. E. v., Blundell, T. L., & Ascher, D. B. (2016). mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Scientific Reports*, *6*(1), 1-8.

Pizza, M. (2000). Identification of Vaccine Candidates Against Serogroup B Meningococcus by Whole-Genome Sequencing. *Science*, *287*(5459), 1816-1820.

Portelli, S., Phelan, J. E., Ascher, D. B., Clark, T. G., & Furnham, N. (2018). Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Scientific Reports*, *8*(1), 1-12.

Price-Carter, M., Brauning, R., de Lisle, G. W., Livingstone, P., Neill, M., Sinclair, J., Paterson, B., Atkinson, G., Knowles, G., Crews, K., Crispell, J., Kao, R., Robbe-Austerman, S., Stuber, T., Parkhill, J., Wood, J., Harris, S., & Collins, D. M. (2018). Whole Genome Sequencing for Determining the Source of Mycobacterium bovis Infections in Livestock Herds and Wildlife in New Zealand. *Frontiers in Veterinary Science*, *5*.

Puigbò, P., Bravo, I. G., & Garcia-Vallvé, S. (2008). E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). *BMC Bioinformatics*, *9*(1), 1-7.

Quemard, A., Sacchettini, J. C., Dessen, A., Vilcheze, C., Bittman, R., Jacobs, W. R., & Blanchard, J. S. (1995). Enzymic Characterization of the Target for Isoniazid in Mycobacterium tuberculosis. *Biochemistry*, *34*(26), 8235-8241.

Quesniaux, V. J., Nicolle, D. M., Torres, D., Kremer, L., Guérardel, Y., Nigou, J., Puzo, G., Erard, F., & Ryffel, B. (2004). Toll-Like Receptor 2 (TLR2)-Dependent-Positive and TLR2-Independent-Negative Regulation of Proinflammatory Cytokines by Mycobacterial Lipomannans. *The Journal of Immunology*, *172*(7), 4425-4434.

Ramaswamy, S. v., Amin, A. G., Göksel, S., Stager, C. E., Dou, S.-J., el Sahly, H., Moghazeh, S. L., Kreiswirth, B. N., & Musser, J. M. (2000). Molecular Genetic Analysis of Nucleotide Polymorphisms Associated with Ethambutol Resistance in Human Isolates of *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy*, *44*(2).

Ramensky, V. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, *30*(17), 3894-3900.

Rammensee, H.-G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A., & Stevanović, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, *50*(3–4), 213-219.

Rapin, N., Lund, O., Bernaschi, M., & Castiglione, F. (2010). Computational Immunology Meets Bioinformatics: The Use of Prediction Tools for Molecular Binding in the Simulation of the Immune System. *PLoS ONE*, *5*(4), e9862.

Rappuoli, R. (2000). Reverse vaccinology. *Current Opinion in Microbiology*, *3*(5), 445-450.

Rauf, Mohd. A., Zubair, S., & Azhar, A. (2015). Ligand docking and binding site analysis with pymol and autodock/vina. *International Journal of Basic and Applied Sciences*, *4*(2), 168-177.

Rautio, J., Meanwell, N. A., Di, L., & Hageman, M. J. (2018). The expanding role of prodrugs in contemporary drug design and development. *Nature Reviews Drug Discovery*, *17*(8), 559-587.

Rawat, R., Whitty, A., & Tonge, P. J. (2003). The isoniazid-NAD adduct is a slow, tight-binding inhibitor of InhA, the Mycobacterium tuberculosis enoyl reductase: Adduct affinity and drug resistance. *Proceedings of the National Academy of Sciences*, *100*(24), 13881-13886.

Reche, PedroA., Glutting, J.-P., Zhang, H., & Reinherz, EllisL. (2004). Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics*, *56*(6), 405-419.

Reis, A. C., Tenreiro, R., Albuquerque, T., Botelho, A., & Cunha, M. v. (2020). Long-term molecular surveillance provides clues on a cattle origin for Mycobacterium bovis in Portugal. *Scientific Reports*, *10*(1), 1-18.

Renwick, A. R., White, P. C. L., & Bengis, R. G. (2007). Bovine tuberculosis in southern African wildlife: a multi-species host–pathogen system. *Epidemiology and Infection*, *135*(4), 529-540.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, *48*(W1), W449-454.

Reynolds, M. G. (2000). Compensatory Evolution in Rifampin-Resistant *Escherichia coli*. *Genetics*, *156*(4), 1471-1481.

Riccardi, G., Pasca, M. R., & Buroni, S. (2009). *Mycobacterium tuberculosis* : drug resistance and future perspectives. *Future Microbiology*, *4*(5), 597-614.

Richardson, J. S. (1981). The Anatomy and Taxonomy of Protein Structure. *Advances in Protein Chemistry*, *34*, 167-339.

Rivas-Santiago, B., Hernandez-Pando, R., Carranza, C., Juarez, E., Contreras, J. L., Aguilar-Leon, D., Torres, M., & Sada, E. (2008). Expression of Cathelicidin LL-37 during *Mycobacterium tuberculosis* Infection in Human Alveolar Macrophages, Monocytes, Neutrophils, and Epithelial Cells. *Infection and Immunity*, *76*(3).

Robert, F., & Pelletier, J. (2018). Exploring the Impact of Single-Nucleotide Polymorphisms on Translation. *Frontiers in Genetics*, *9*.

Rodo, M. J., Rozot, V., Nemes, E., Dintwe, O., Hatherill, M., Little, F., & Scriba, T. J. (2019). A comparison of antigen-specific T cell responses induced by six novel tuberculosis vaccine candidates. *PLOS Pathogens*, *15*(3), 1-20.

Rodrigues, C. M. C., & Plotkin, S. A. (2020). Impact of Vaccines; Health, Economic and Social Perspectives. *Frontiers in Microbiology*, *11*.

Rohde, K., Yates, R. M., Purdy, G. E., & Russell, D. G. (2007). Mycobacterium tuberculosis and the environment within the phagosome. *Immunological Reviews*, *219*(1), 37-54.

Rook, G. A. W., & Hernandez-Pando, R. (1996). THE PATHOGENESIS OF TUBERCULOSIS. *Annual Review of Microbiology*, *50*(1), 259-284.

Rouse, D. A., DeVito, J. A., Li, Z., Byer, H., & Morris, S. L. (1996). Site-directed mutagenesis of the *KatG* gene of *Mycobacterium tuberculosis* : effects on catalase-peroxidase activities and isoniazid resistance . *Molecular Microbiology*, *22*(3), 583-592.

Roy, A., Yang, J., & Zhang, Y. (2012). COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*, *40*(W1), W471-477.

Sachdeva, G., Kumar, K., Jain, P., & Ramachandran, S. (2005). SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics*, *21*(4), 483-491.

Saha, S., & Raghava, G. P. S. (2006a). AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Research*, *34*(Web Server), W202-209.

Saha, S., & Raghava, G. P. S. (2006b). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins: Structure, Function, and Bioinformatics*, *65*(1), 40-48.

Sahu, R., Singh, K., & Subodh, S. (2015). Adverse Drug Reactions to Anti-TB Drugs: Pharmacogenomics Perspective for Identification of Host Genetic Markers. *Current Drug Metabolism*, *16*(7), 538-552.

Salfinger, M., Crowle, A. J., & Reller, L. B. (1990). Pyrazinamide and Pyrazinoic Acid Activity against Tubercle Bacilli in Cultured Human Macrophages and in the BACTEC System. *Journal of Infectious Diseases*, *162*(1), 201-207.

Schaible, U. E., & Kaufmann, S. H. E. (2000). CD1 molecules and CD1-dependent T cells in bacterial infections: a link from innate to acquired immunity? *Seminars in Immunology*, *12*(6), 527-535.

Schatz, A., Bugie, E., Waksman, S. A., Hanssen, A. D., Patel, R., & Osmon, D. R. (2005). The Classic: Streptomycin, a Substance Exhibiting Antibiotic Activity against Gram-Positive and Gram-Negative Bacteria. *Clinical Orthopaedics and Related Research*, *437*, 3-6.

Schindler, H., Lutz, M. B., Röllinghoff, M., & Bogdan, C. (2001). The Production of IFN-γ by IL-12/IL-18-Activated Macrophages Requires STAT4 Signaling and Is Inhibited by IL-4. *The Journal of Immunology*, *166*(5), 3075-3082.

Schirle, M., Weinschenk, T., & Stevanović, S. (2001). Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens. *Journal of Immunological Methods*, *257*(1–2), 1-16.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., & Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Research*, *33*(Web Server), W363-367.

Scorpio, A., Lindholm-Levy, P., Heifets, L., Gilman, R., Siddiqi, S., Cynamon, M., & Zhang, Y. (1997). Characterization of *pncA* mutations in pyrazinamide-resistant Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy*, *41*(3).

Scorpio, A., & Zhang, Y. (1996). Mutations in *pncA*, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nature Medicine*, *2*(6), 662-667.

Sebastian, J., Swaminath, S., Nair, R. R., Jakkala, K., Pradhan, A., & Ajitkumar, P. (2017). *De Novo* Emergence of Genetically Resistant Mutants of Mycobacterium tuberculosis from the Persistence Phase Cells Formed against Antituberculosis Drugs *In Vitro*. *Antimicrobial Agents and Chemotherapy*, *61*(2).

Seib, K. L., Zhao, X., & Rappuoli, R. (2012). Developing vaccines in the era of genomics: a decade of reverse vaccinology. *Clinical Microbiology and Infection*, *18*(5), 109-116.

Seung, K. J., Keshavjee, S., & Rich, M. L. (2015). Multidrug-Resistant Tuberculosis and Extensively Drug-Resistant Tuberculosis. *Cold Spring Harbor Perspectives in Medicine*, *5*(9).

Shah, S., & Briken, V. (2016). Modular Organization of the ESX-5 Secretion System in Mycobacterium tuberculosis. *Frontiers in Cellular and Infection Microbiology*, *6*.

Sharma, S. K., Katoch, K., Sarin, R., Balambal, R., Kumar Jain, N., Patel, N., Murthy, K. J. R., Singla, N., Saha, P. K., Khanna, A., Singh, U., Kumar, S., Sengupta, A., Banavaliker, J. N., Chauhan, D. S., Sachan, S., Wasim, M., Tripathi, S., Dutt, N., … Rani, R. (2017). Efficacy and Safety of Mycobacterium indicus pranii as an adjunct therapy in Category II pulmonary tuberculosis in a randomized trial. *Scientific Reports*, *7*(1), 1-12.

Sharma, S., & Mohan, A. (2004). Extrapulmonary tuberculosis. *The Indian Journal of Medical Research*, *120*(4), 316–353.

Shen, M., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, *15*(11), 2507-2524.

Sherman, D. R., Mdluli, K., Hickey, M. J., Arain, T. M., Morris, S. L., Barry, C. E., & Stover, C. K. (1996). Compensatory *ahpC* Gene Expression in Isoniazid-Resistant Mycobacterium tuberculosis. *Science*, *272*(5268), 1641-1643.

Shoichet, B. K., McGovern, S. L., Wei, B., & Irwin, J. J. (2002). Lead discovery using molecular docking. *Current Opinion in Chemical Biology*, *6*(4), 439-446.

Simeone, R., Bottai, D., & Brosch, R. (2009). ESX/type VII secretion systems and their role in host–pathogen interaction. *Current Opinion in Microbiology*, *12*(1), 4-10.

Singh, H., & Raghava, G. P. S. (2001). ProPred: prediction of HLA-DR binding sites. *Bioinformatics*, *17*(12), 1236-1237.

Singh, S., Singh, H., Tuknait, A., Chaudhary, K., Singh, B., Kumaran, S., & Raghava, G. P. S. (2015). PEPstrMOD: structure prediction of peptides containing natural, non-natural and modified residues. *Biology Direct*, *10*(1), 1-19.

Singh, V., & Mizrahi, V. (2017). Identification and validation of novel drug targets in Mycobacterium tuberculosis. *Drug Discovery Today*, *22*(3), 503-509.

Skeiky, Y. A. W., & Sadoff, J. C. (2006). Advances in tuberculosis vaccine strategies. *Nature Reviews Microbiology*, *4*(6), 469-476.

Smith, I. (2003). *Mycobacterium tuberculosis* Pathogenesis and Molecular Determinants of Virulence. *Clinical Microbiology Reviews*, *16*(3), 463-496.

Sobolev, B., Olenina, L., Kolesanova, E., Poroikov, V., & Archakov, A. (2005). Computer Design of Vaccines: Approaches, Software Tools and Informational Resources. *Current Computer Aided-Drug Design*, *1*(2).

Somoskovi, A., Parsons, L. M., & Salfinger, M. (2001). The molecular basis of resistance to isoniazid, rifampin, and pyrazinamide in Mycobacterium tuberculosis. *Respiratory Research*, *2*(3).

Sreevatsan, S., Pan, X., Zhang, Y., Kreiswirth, B. N., & Musser, J. M. (1997). Mutations associated with pyrazinamide resistance in *pncA* of Mycobacterium tuberculosis complex organisms. *Antimicrobial Agents and Chemotherapy*, *41*(3).

Srinivasan, S., Easterling, L., Rimal, B., Niu, X. M., Conlan, A. J. K., Dudas, P., & Kapur, V. (2018). Prevalence of Bovine Tuberculosis in India: A systematic review and meta-analysis. *Transboundary and Emerging Diseases*, *65*(6), 1627-1640.

Stavrinides, J., McCann, H. C., & Guttman, D. S. (2007). Host–pathogen interplay and the evolution of bacterial effectors. *Cellular Microbiology*, *10*(2), 285-292.

Stranzl, T., Larsen, M. V., Lundegaard, C., & Nielsen, M. (2010). NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics*, *62*(6), 357-368.

Strohmeier, G. R., & Fenton, M. J. (1999). Roles of lipoarabinomannan in the pathogenesis of tuberculosis. *Microbes and Infection*, *1*(9), 709-717.

Sun, Z., Zhang, Y., & Permar, S. (2002). Conditions that may affect the results of susceptibility testing of Mycobacterium tuberculosis to pyrazinamide. *Journal of Medical Microbiology*, *51*(1), 42-49.

Sutcliffe, I. C., & Harrington, D. J. (2004). Lipoproteins of *Mycobacterium tuberculosis* : an abundant and functionally diverse class of cell envelope components. *FEMS Microbiology Reviews*, *28*(5), 645-659.

Szakacs, T. A., Wilson, D., Cameron, D. W., Clark, M., Kocheleff, P., Muller, F. J., & McCarthy, A. E. (2006). Adherence with isoniazid for prevention of tuberculosis among HIV-infected adults in South Africa. *BMC Infectious Diseases*, *6*(1), 1-7.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. von. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607-613.

Tameris, M., Mearns, H., Penn-Nicholson, A., Gregg, Y., Bilek, N., Mabwe, S., Geldenhuys, H., Shenje, J., Luabeya, A. K. K., Murillo, I., Doce, J., Aguilo, N., Marinova, D., Puentes, E., Rodríguez, E., Gonzalo-Asensio, J., Fritzell, B., Thole, J., Martin, C., … Veldsman, A. (2019). Live-attenuated *Mycobacterium tuberculosis* vaccine MTBVAC versus BCG in adults and neonates: a randomised controlled, double-blind dose-escalation trial. *The Lancet Respiratory Medicine*, *7*(9), 757-770.

Tani, K., Murphy, W. J., Chertov, O., Salcedo, R., Koh, C. Y., Utsunomiya, I., Funakoshi, S., Asai, O., Herrmann, S. H., Wang, J. M., Kwak, L. W., & Oppenheim, J. J. (2000). Defensins act as potent adjuvants that promote cellular and humoral immune responses in mice to a lymphoma idiotype and carrier antigens. *International Immunology*, *12*(5), 691-700.

Taniguchi, H., Aramaki, H., Nikaido, Y., Mizuguchi, yasuo, Nakamura, M., Koga, T., & Yoshida, S. (1996). Rifampicin resistance and mutation of the *rpoB* gene in *Mycobacterium tuberculosis*. *FEMS Microbiology Letters*, *144*(1), 103-108.

Telenti, A., Imboden, P., Marchesi, F., Matter, L., Schopfer, K., Bodmer, T., Lowrie, D., Colston, M. J., & Cole, S. (1993). Detection of rifampicin-resistance mutations in *Mycobacterium tuberculosis*. *The Lancet*, *341*(8846), 647-651.

Telenti, A., Philipp, W. J., Sreevatsan, S., Bernasconi, C., Stockbauer, K. E., Wieles, B., Musser, J. M., & Jacobs, W. R. (1997). The emb operon, a gene cluster of *Mycobacterium tuberculosis* involved in resistance to ethambutol. *Nature Medicine*, *3*(5), 567-570.

Thomas, J., Balseiro, A., Gortázar, C., & Risalde, M. A. (2021). Diagnosis of tuberculosis in wildlife: a systematic review. *Veterinary Research*, *52*(1), 1-23.

Timmins, G., & Deretic, V. (2006). Mechanisms of action of isoniazid. *Molecular Microbiology*, *62*(5), 1220-1227.

Tomar, N., & De, R. K. (2014). A Brief Outline of the Immune System. *Immunoinformatics*, 3-12.

Torgerson, P. R., & Torgerson, D. J. (2010). Public health and bovine tuberculosis: what's all the fuss about? *Trends in Microbiology*, *18*(2), 67-72.

Torrelles, J. B., & Schlesinger, L. S. (2010). Diversity in *Mycobacterium tuberculosis* mannosylated cell wall determinants impacts adaptation to the host. *Tuberculosis*, *90*(2), 84-93.

Torres-Gonzalez, P., Soberanis-Ramos, O., Martinez-Gamboa, A., Chavez-Mazari, B., Barrios-Herrera, M. T., Torres-Rojas, M., Cruz-Hervert, L. P., Garcia-Garcia, L., Singh, M., Gonzalez-Aguirre, A., Ponce de Leon-Garduño, A., Sifuentes-Osornio, J., & Bobadilla-del-Valle, M. (2013). Prevalence of Latent and Active

Tuberculosis among Dairy Farm Workers Exposed to Cattle Infected by *Mycobacterium bovis*. *PLoS Neglected Tropical Diseases*, *7*(4), e2177.

Trias, J., & Benz, R. (1994). Permeability of the cell wall of *Mycobacterium smegmatis*. *Molecular Microbiology*, *14*(2), 283-290.

Trias, J., Jarlier, V., & Benz, R. (1992). Porins in the cell wall of mycobacteria. *Science*, *258*(5087), 1479-1481.

Tusnady, G. E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics*, *17*(9), 849-850.

Unissa, A. N., Doss C., G. P., Kumar, T., Swathi, S., Lakshmi, A. R., & Hanna, L. E. (2017). Analysis of interactions of clinical mutants of catalase-peroxidase (*KatG*) responsible for isoniazid resistance in Mycobacterium tuberculosis with derivatives of isoniazid. *Journal of Global Antimicrobial Resistance*, *11*, 57-67.

Vadwai, V., Boehme, C., Nabeta, P., Shetty, A., Alland, D., & Rodrigues, C. (2011). Xpert MTB/RIF: a New Pillar in Diagnosis of Extrapulmonary Tuberculosis? *Journal of Clinical Microbiology*, *49*(7), 2540-2545.

van Crevel, R., Ottenhoff, T. H. M., & van der Meer, J. W. M. (2002). Innate Immunity to *Mycobacterium tuberculosis*. *Clinical Microbiology Reviews*, *15*(2), 294-309.

van den Boogaard, J., Kibiki, G. S., Kisanga, E. R., Boeree, M. J., & Aarnoutse, R. E. (2009). New Drugs against Tuberculosis: Problems, Progress, and Evaluation of Agents in Clinical Development. *Antimicrobial Agents and Chemotherapy*, *53*(3), 849-862.

Vergne, I., Chua, J., & Deretic, V. (2003). *Mycobacterium tuberculosis* Phagosome Maturation Arrest: Selective Targeting of PI3P-Dependent Membrane Trafficking. *Traffic*, *4*(9), 600-606.

Vergne, I., Chua, J., Lee, H.-H., Lucas, M., Belisle, J., & Deretic, V. (2005). Mechanism of phagolysosome biogenesis block by viable Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, *102*(11), 4033-4038.

Versteeg, L., Almutairi, M. M., Hotez, P. J., & Pollet, J. (2019). Enlisting the mRNA Vaccine Platform to Combat Parasitic Infections. *Vaccines*, *7*(4), 1-19.

Vidal Pessolani, M. C., de Melo Marques, M. A., Reddy, V. M., Locht, C., & Menozzi, F. D. (2003). Systemic dissemination in tuberculosis and leprosy: do mycobacterial adhesins play a role? *Microbes and Infection*, *5*(7), 677-684.

Vilchèze, C., Morbidoni, H. R., Weisbrod, T. R., Iwamoto, H., Kuo, M., Sacchettini, J. C., & Jacobs, W. R. (2000). Inactivation of the *inhA* -Encoded Fatty Acid Synthase II (FASII) Enoyl-Acyl Carrier Protein Reductase Induces Accumulation of the FASI End Products and Cell Lysis of *Mycobacterium smegmatis*. *Journal of Bacteriology*, *182*(14), 4059-4067.

Vivona, S., Gardy, J. L., Ramachandran, S., Brinkman, F. S. L., Raghava, G. P. S., Flower, D. R., & Filippini, F. (2008). Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. *Trends in Biotechnology*, *26*(4), 190-200.

Volmink, J., & Garner, P. (2007). Directly observed therapy for treating tuberculosis. In J. Volmink (Ed.), *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd.

von Reyn, C. F., Lahey, T., Arbeit, R. D., Landry, B., Kailani, L., Adams, L. v., Haynes, B. C., Mackenzie, T., Wieland-Alter, W., Connor, R. I., Tvaroha, S., Hokey, D. A., Ginsberg, A. M., & Waddell, R. (2017). Safety and immunogenicity of an inactivated whole cell tuberculosis vaccine booster in adults primed with BCG: A randomized, controlled trial of DAR-901. *PLOS ONE*, *12*(5), 1-16.

Wan, S., Mak, M.-W., & Kung, S.-Y. (2017). Gram-LocEN: Interpretable prediction of subcellular multi-localization of Gram-positive and Gram-negative bacterial proteins. *Chemometrics and Intelligent Laboratory Systems*, *162*, 1-9.

Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M., & Peters, B. (2010). Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics*, *11*(1), 1-12.

Wang, S., Li, W., Liu, S., & Xu, J. (2016). RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research*, *44*(W1), W430-435.

Wass, M. N., Kelley, L. A., & Sternberg, M. J. E. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Research*, *38*(suppl_2), W469-473.

Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, *54*(1).

Wehrli, W., Knusel, F., Schmid, K., & Staehelin, M. (1968). Interaction of rifamycin with bacterial RNA polymerase. *Proceedings of the National Academy of Sciences*, *61*(2), 667-673.

Wells, R. M., Jones, C. M., Xi, Z., Speer, A., Danilchanka, O., Doornbos, K. S., Sun, P., Wu, F., Tian, C., & Niederweis, M. (2013). Discovery of a Siderophore Export System Essential for Virulence of Mycobacterium tuberculosis. *PLoS Pathogens*, *9*(1), e1003120.

Wetlaufer, D. B. (1973). Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proceedings of the National Academy of Sciences*, *70*(3), 697-701.

Wilkie, M., Satti, I., Minhinnick, A., Harris, S., Riste, M., Ramon, R. L., Sheehan, S., Thomas, Z.-R. M., Wright, D., Stockdale, L., Hamidi, A., O'Shea, M. K., Dwivedi, K., Behrens, H. M., Davenne, T., Morton, J., Vermaak, S., Lawrie, A., Moss, P., & McShane, H. (2020). A phase I trial evaluating the safety and immunogenicity of a candidate tuberculosis vaccination regimen, ChAdOx1 85A prime – MVA85A boost in healthy UK adults. *Vaccine*, *38*(4), 779-789.

Wlodarska, M., Johnston, J. C., Gardy, J. L., & Tang, P. (2015). A Microbiological Revolution Meets an Ancient Disease: Improving the Management of Tuberculosis with Genomics. *Clinical Microbiology Reviews*, *28*(2), 523-539.

Wu, Q., Peng, Z., Zhang, Y., & Yang, J. (2018). COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Research*, *46*(W1), W438-442.

Xiao, X., Cheng, X., Chen, G., Mao, Q., & Chou, K.-C. (2019). pLoc_bal-mGpos: Predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. *Genomics*, *111*(4), 886-892.

Yang, J., Roy, A., & Zhang, Y. (2013). Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, *29*(20), 2588-2595.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, *12*(1), 7-8.

You, L., Brusic, V., Gallagher, M., & Boden, M. (2010). Using Gaussian Process with Test Rejection to Detect T-Cell Epitopes in Pathogen Genomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *7*(4), 741-751.

Yu, C., Chen, Y., Lu, C., & Hwang, J. (2006). Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics*, *64*(3), 643-651.

Yu, N., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., & Brinkman, F. S. L. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, *26*(13), 1608-1615.

Yuk, J. M., & Jo, E. K. (2014). Host immune responses to mycobacterial antigens and their implications for the development of a vaccine to control tuberculosis. *Clinical and Experimental Vaccine Research*, *3*(2), 155-167.

Zagursky, R. J., & Russell, D. (2001). Bioinformatics: Use in Bacterial Vaccine Discovery. *BioTechniques*, *31*(3).

Zhang, L., Chen, Y., Wong, H., Zhou, S., Mamitsuka, H., & Zhu, S. (2012). TEPITOPEpan: Extending TEPITOPE for Peptide Binding Prediction Covering over 700 HLA-DR Molecules. *PLoS ONE*, *7*(2), e30483.

Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., Beaver, J., Bourne, P., Bui, H., Buus, S., Frankild, S., Greenbaum, J., Lund, O., Lundegaard, C., Nielsen, M., Ponomarenko, J., Sette, A., Zhu, Z., & Peters, B. (2008). Immune epitope database analysis resource (IEDB-AR). *Nucleic Acids Research*, *36*(Web Server), W513-518.

Zhang, R. (2004). DEG: a database of essential genes. *Nucleic Acids Research*, *32*(90001), D271-272.

Zhang. Y. (2005). *Annual Review of Pharmacology and Toxicology*, *45*(1), 529-564.

Zhang, Y., Heym, B., Allen, B., Young, D., & Cole, S. (1992). The catalase—peroxidase gene and isoniazid resistance of Mycobacterium tuberculosis. *Nature*, *358*(6387), 591-593.

Zhang, Y., & Yew, W. (2015). Mechanisms of drug resistance in *Mycobacterium tuberculosis*: update 2015. *The International Journal of Tuberculosis and Lung Disease*, *19*(11), 1276-1289.

Zhang, & Yew. (2009). Mechanisms of drug resistance in *Mycobacterium tuberculosis*. *International Journal of Tuberculosis and Lung Disease*, *13*(11), 1320–1330.

Zimmerman, M. (1979). Pulmonary and osseous tuberculosis in an Egyptian mummy. *Bulletin of New York Academy of Medicine*, *55*(6), 604–608.