UNDERSTANDING OF THE UNDERLYING RESISTANCE MECHANISM OF THE KAT-G PROTEIN AGAINST ISONIAZID IN *Mycobacterium tuberculosis* USING BIOINFORMATICS APPROACHES

A thesis submitted in partial fulfilment of the requirements for the degree

of

Master of Science in Bioinformatics and Computational Molecular Biology

(Coursework and Thesis)

of

RHODES UNIVERSITY, SOUTH AFRICA

Research Unit in Bioinformatics (RUBi)

DEPARTMENT OF BIOCHEMISTRY and MICROBIOLOGY

Faculty of Science

by

Victor Barozi

19B9790

08/01/2020

ABSTRACT

Tuberculosis (TB) is a multi-organ infection caused by rod-shaped acid-fast *Mycobacterium tuberculosis*. The World Health Organization (WHO) ranks TB among the top 10 fatal infections and the leading the cause of death from a single infection. In 2017, TB was responsible for an estimated 1.3 million deaths among both the HIV negative and positive populations worldwide (WHO, 2018). Approximately 23% (roughly 1.7 billion) of the world's population is estimated to have latent TB with a high risk of reverting to active TB infection. In 2017, an estimated 558,000 people developed drug resistant TB worldwide with 82% of the cases being multi-drug resistant TB (WHO, 2018). South Africa is ranked among the 30 high TB burdened countries with a TB incidence of 322,000 cases in 2017 accounting for 3% of the world's TB cases. TB is curable and is clinically managed through a combination of intensive and continuation phases of first-line drugs (isoniazid, rifampicin, ethambutol, and pyrazinamide). Second-line drugs which include fluoroquinolones, injectable aminoglycoside and injectable polypeptides are used in cases of first line drug resistance. The third-line drugs include amoxicillin, clofazimine, linezolid and imipenem. These have variable but unproven efficacy to TB and are the last resort in cases of total drug resistance (Jilani et al., 2019). TB drug resistance to first-line drugs especially isoniazid in *M. tuberculosis* has been attributed to single nucleotide polymorphisms (SNPs) in the catalase peroxidase enzyme (katG), a protein important in the activation of the pro-drug isoniazid. The SNPs especially at position 315 of the katG enzyme are believed to reduce the sensitivity of the *M. tuberculosis* to isoniazid while still maintaining the enzyme's catalytic activity - a mechanism not completely understood. KatG protein is important for protecting the bacteria from hydro peroxides and hydroxyl radicals present in an aerobic environment. This study focused on understanding the mechanism of isoniazid drug resistance in *M. tuberculosis* as a result of high confidence mutations in the katG through modelling the enzyme with its respective variants, performing

MD simulations to explore the protein behaviour, calculating the dynamic residue network analysis (DRN) of the variants in respect to the wild type katG and finally performing alanine scanning. From the MD simulations, it was observed that the high confidence mutations i.e. S140R, S140N, G279D, G285D, S315T, S315I, S315R, S315N, G316D, S457I and G593D were not only reducing the backbone flexibility of the protein but also reducing the protein's conformational variation and space. All the variant protein structures were observed to be more compact compared to the wild type. Residue fluctuation results indicated reduced residue flexibility across all variants in the loop region (position 26-110) responsible for katG dimerization. In addition, mutation S315T is believed to reduce the size of the active site access channel in the protein. From the DRN data, residues in the interface region between the N and C-terminal domains were observed to gain importance in the variants irrespective of the mutation location indicating an allosteric effect of the mutations on the interface region. Alanine scanning results established that residue Leucine at position 48 was not only important in the protein communication but also a destabilizing residue across all the variants. The study not only demonstrated change in the protein behaviour but also showed allosteric effect of the mutations in the katG protein.

DECLARATION

I, **Victor Barozi,** hereby declare that this thesis submitted to Rhodes University is my original work and has never been submitted to any institution for a degree or diploma. All sources, references and literature used during preparation of this work are properly cited and listed in complete reference to the due source.

…..   ..…..

Signature   Date

DEDICATION

This thesis is dedicated to Miss Bridget Tusiime Jolly for always supporting and believing in me.

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude to my supervisor Prof Özlem Tastan Bishop for giving me the opportunity to work on this study and for the invaluable guidance offered throughout the course of the study.

I would also like to acknowledge my co-supervisors; Dr. Musyoka, Thommas Mutemi and Dr. Sheik Amamuddy, Olivier for the support and insight offered towards completion of this thesis.

Lastly, I am eternally grateful to the all the members of the Research Unit in Bioinformatics (RUBi), Rhodes University especially my fellow MSc colleagues.

Funding acknowledgement

This work was supported through the Grand Challenges Africa: Drug Discovery Programme [GCA/DD/rnd3/023]. Grand Challenges Africa is a program of the African Academy of Sciences (AAS) implemented through the Alliance for Accelerating Excellence in Science in Africa (AESA) platform, an initiative of the AAS and the African Union Development Agency (AUDA-NEPAD). GC Africa is supported by the Bill & Melinda Gates Foundation (BMGF), Swedish International Development Cooperation Agency (SIDA), German Federal Ministry of Education and Research (BMBF), Medicines for Malaria Venture (MMV), and Drug Discovery and Development Centre of University of Cape Town (H3D). The content of this thesis is solely the responsibility of the author and does not necessarily represent the official views of the funders.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF WEB SERVER AND SOFTWARE TOOLS

Web server	**link**
ATB	https://atb.uq.edu.au/
BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch
HHpred	http://toolkit.tuebingen.mpg.de/hhpred
LigPlus	https://www.ebi.ac.uk/thornton-srv/software/LigPlus/
MD-TASK	https://md-task.readthedocs.io/en/latest/home.html
MODE-TASK	https://mode-task.readthedocs.io/en/latest/
NCBI	https://www.ncbi.nlm.nih.gov/
PIC	http://pic.mbu.iisc.ernet.in/
PRIMO	https://primo.rubi.ru.ac.za/
PROCHECK	https://servicesn.mbi.ucla.edu/PROCHECK/
ProSA	https://prosa.services.came.sbg.ac.at
ROBETTA	http://robetta.bakerlab.org/
SWISS MODEL	http://swissmodel.expasy.org/SWISS-MODEL.html
TBDReaMDB	https://tbdreamdb.ki.se/Info/
TCoffee	https://www.ebi.ac.uk/Tools/msa/tcoffee/
VERIFY3D	https://servicesn.mbi.ucla.edu/Verify3D/

LIST OF ABBREVIATIONS

AIDS	Acquired Immune Deficiency Syndrome
ATB	Automated Topology Builder
BLAST	Basic Local Alignment Search Tool
BC	*Betweenness Centrality*
DNA	Deoxy Ribonucleic Acid
DOPE	Discrete Optimized Protein Energy
DRN	Dynamic Residue Network
EM	Electron Microscopy
GROMACS	GROningen MAchine for Chemical Simulations
HIV	Human Immunodeficiency Virus
HMMs	Hidden Markov Models
INH	Isoniazid
ID	Identification
LAM	Lipoarabinomannan
L	Reachability
MD	Molecular Dynamics
MDR	Multidrug resistant Tuberculosis
MM	Molecular Mechanics
MTB	*Mycobacterium tuberculosis*

MQAP	Model Quality Evaluation Programs
NAD	Nicotinamide Adenine Dinucleotide
NMR	Nuclear Magnetic Resonance
NCBI	National Centre for Biotechnology Information
PCA	Principal Component Analysis
PDB	Protein Data Bank
PIC	Protein Interaction Calculator
QM	Quantum Mechanics
RIN	Residue Interaction Network
Rg	Radius of gyration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RNA	Ribonucleic Acid
RR	Rifampicin resistant
TB	Tuberculosis
WHO	World Health Organization
XDR	Extensive Drug Resistance
ZN	Ziehl-Neelsen
3D	3-Dimensional

CHAPTER ONE

LITERATURE REVIEW

1.1 INTRODUCTION

Tuberculosis (TB) is an infectious disease caused by bacteria called *Mycobacterium tuberculosis.* The infection is one of the leading causes of deaths worldwide and the leading cause from a single infectious agent (WHO, 2018). TB is transmitted through the air when a person with active TB infection in their lungs coughs or sneezes and the other inhales the aerosols that contain the *M. tuberculosis.* TB mainly affects the lungs however it is also known to damage other body organs like the spleen and the spine (WHO, 2018). Symptoms of active lung TB infection include; coughing with sputum and blood, malaise, loss of body weight, fever and general body weakness (Churchyard et al., 2017).

1.2 BACKGROUND

1.2.1 Global and national TB burden

In almost every country, there is a documented case of TB and nearly one-third of the world's population has been infected by TB (Singh et al., 2018). Recent estimates put the total number of people infected with TB at 1.7 billion with a small proportion of them (5-10%) expected to develop tuberculosis disease. This proportion is expectedly higher among people living with Human Immune deficiency Virus (HIV) (Ismail et al., 2018). As per the WHO, two-thirds of the world's TB incidence is contributed by India, sub-Saharan Africa, China, Indonesia, Pakistan and Philippines (Figure 1.1). A susceptible TB infection is the type of infection that is sensitive to available TB therapies whereas a resistant TB infection can be resistant to either first line, second line or sometimes third line TB drugs. More than 47% of the global burden of resistant TB is accounted for by three countries i.e. Russia, China and India. The highest prevalence was found to be in Russia at 27% among new patients and 63% among patients previously on TB treatment. A similar TB survey in China revealed that multidrug resistant TB (MDR-TB) prevalence was at 5.7% and 25.6% among new cases and previously treated cases respectively while data from India showed a prevalence of 2.84% and 11.64% among new and previously treated patients respectively (Ismail et al., 2018).

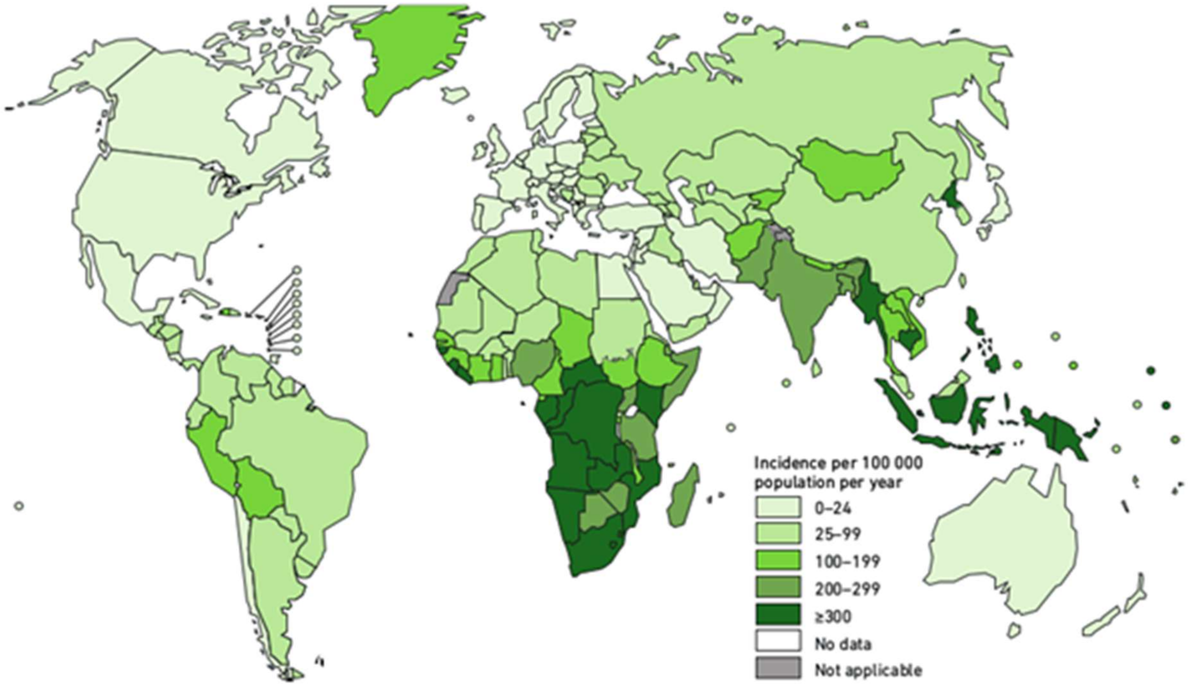


Figure 1.1: World map of the estimated global TB incidence distribution as per 2017 (Adapted from the WHO global TB report 2018).

On a national level, the incidence of TB in South Africa is among the highest in the world with up to 450,000 individuals diagnosed in 2015 alone (Glaziou et al., 2009). In 2015, TB was also the leading cause of death in South Africa accounting for 8.4% of the natural deaths that year (Kweza et al., 2018). The mortality rates could be explained by the low access to TB treatment. A study in 2016 estimated that only 72.8% of the people diagnosed with TB were started on treatment (Skinner et al., 2016). Among the key drivers for the TB burden in South Africa is poor access to health facilities, poor health seeking habits and poor living conditions (Hartel et al., 2018). The TB burden is disproportionately higher among people living with HIV with

63% of the reported TB cases being HIV positive (Robertson et al., 2018). TB is also significantly higher among the black population with more than 91% of the TB patients being black (Oni et al., 2015). Other vulnerable populations include health workers, miners, incarcerated populations, pregnant women, children < 5 years of age, and people with diabetes (Oni et al., 2015). In 2017, there was a 2.1% prevalence among inmates and people living in enclosed spaces, 2760 per 100,000 among people with diabetics and the prevalence among gold miners was at 3000-7000 per 100,000 (South African National AIDS Council., 2019).

One key barrier to ending the TB epidemic in South Africa is the emerging drug resistance to the commonly used first-line TB drugs. A recent survey reported that 2.1% (n=5423) of the new TB cases and 25.6% of previously treated cases had MDR-TB with up to 4.9% of the drug resistance cases being XDR-TB (Extensively Drug Resistant TB: further explained under section 1.2.5) (Sharp et al., 2018). Rifampicin resistance testing is at almost 80% coverage with the use of Gene-Xpert machines and although this helps in screening MDR cases, isoniazid mono-resistance is often left undiagnosed, which often impacts patient management. This is evident in the fact that currently, only 14% of Gene-Xpert negative cases are followed up for further drug resistance testing (Cox et al., 2017, Ismail et al 2018).

The economic burden of TB is heavy on resource-limited countries like South Africa. In 2017 alone, the estimated public expenditure for the TB program was approximately $415 million, of which $310 million was used for direct service delivery. This figure was projected to shoot up to $332 million by 2020 (Bozzani et al., 2018). These numbers raise a serious concern and urgency to find novel ways of managing and preventing TB infection.

1.2.2 Etiology of TB

Historically, TB has been documented as far back as the 17th Century. The etiology of TB was mysterious until 1882 when the *Mycobacterium tuberculosis* was discovered by Dr. Robert

Koch (Barberis et al., 2017). As mentioned before, TB is caused by *M. tuberculosis*; a weakly gram positive, non-motile, aerobic and acid-fast bacillus (Barberis et al., 2017). The disease is highly contagious and spreads when infected individuals eject aerosolized droplets containing the bacteria into the surrounding, for example by coughing or sneezing and a new host inhales the bacteria containing aerosols. The probability of spread is contingent on the infectivity of the source case (i.e. how diseased the source individual is), the level of exposure (i.e. closeness, air circulation, and the length of exposure) as well as susceptibility of the person who is in close proximity to an infected case.

The development of TB requires both infection by *M. tuberculosis* and a weakened immune system. Studies have shown that not every individual exposed to TB develops the infection (Chai et al., 2018). The immune system is capable of clearing the body of the infection most of the time with only 5%-10% of exposures resulting in a primary infection (Petruccioli et al., 2016). In addition, many of the cases of TB infection do not become diseased, the majority of patients infected with *M. tuberculosis* exist with no clinical, bacteriologic or radiographic evidence of active TB and this is what is commonly referred to as latent TB infection (LTB). If conditions allow, previously latent infection may progress to active TB. People living with HIV, expectant mothers, injection drug users, cigarette smokers and all immuno-compromised individuals are at a higher risk of TB reactivation and primary TB progression (Holt et al., 2016). Another group at risk of TB reactivation is people from TB endemic countries, a study in the UK showed that reactivation among immigrants six years after moving to the UK accounted for 60% of the cases in the UK (Public Health England, 2016).

Both clinical and research evidence has demonstrated that TB primarily affects the lungs (pulmonary TB), but is also capable to infecting several other sites in the body including lymph nodes, spleen, urinary system, bones and the central nervous system (CNS).

1.2.3 TB diagnosis

Over the years, a number of diagnostic methods have been put forward as a means to both screen and confirm TB infection. It should be noted that some methods are more sensitive and specific than others and therefore that should always be taken into consideration when selecting a diagnostic method. Some of the most common methods include; the Ziehl-Neelsen technique (ZN) that involves microscopically examining sputum stained slides. This method regardless of the low sensitivity, it is relatively affordable for developing countries, highly specific and can be done under field conditions (Ahmed et al., 2019) hence making it the most commonly used method especially in resource limited settings.

Whereas cultures are considered the gold standard in TB diagnosis, culture results take too long with an average turnaround time of up to one week (Ogwang et al., 2009). Another hindrance to TB cultures is the expenses associated with the test and in fact, TB cultures are not done outside reference labs and research institutions in some African countries (Agrawal et al., 2016).

Chest X-ray is another TB diagnostic method which focuses on radiographic changes in the lungs associated with positive and negative Ziehl-Neelsen stained smears. In one study, calcification, hilar adenopathy, incomplete destruction and bronchiectasis were found to be associated with positive Ziehl-Neelsen stained smears in TB infected persons (Ebrahimzadeh et al., 2014).

Currently, newer techniques that are either molecular or immunology based are being utilized in a number of countries. One such test is the TB LAM test, which is an immunology-based rapid assay that utilizes urine samples (Broger et al., 2019). Research has shown that the TB LAM test is highly insensitive, meaning it would not make an effective screening test, however, given its applicability as a low-cost point of care urine-based tool, TB LAM has been said to be innovative for limited resource settings (Lawn et al., 2017). In a study done in South Africa, TB LAM correctly identified only 39% of the positive TB cases while correctly identifying 98.9% of the TB negative cases. Based on the results, Lawn et al., (2017) argued that TB LAM would be useful as a tool for identifying poor prognosis outcomes rather than a diagnostic tool.

Over the past few years, molecular diagnostic methods like the Gene-Xpert have been developed to better diagnose TB. In the Gene-Expert method, the *M. tuberculosis* genetic material is isolated from the sputum sample through sonication and amplified by polymerase chain reaction (PCR). In addition to identifying the bacteria, it tests for rifampicin drug resistance. The DNA amplification test has gained popularity for its high sensitivity, batch running of samples and ability to measure rifampicin resistance. By 2017, there were 314 Gene-Xpert machines in 207 microscopy centers and about 8 million assays conducted in South Africa. It is also important to note that South Africa accounts for at least 50% of the Global consumption of Gene-Xpert cartridges.

1.2.4 TB treatment

The first TB registered clinical management approaches were sunlight exposure, adequate nutrition and isolation of the infected persons to prevent transmission of the infection to the healthy individuals (Sotgiu et al., 2015). With advances in the medical treatment and a better understanding of TB, the surgical management method was discovered and set as the standard around 1927 (Sakula, 1983). The discovery of the TB etiological agent by Robert Koch in 1882 paved way for use of natural and chemical compounds targeting the mycobacteria for clinical management of TB (Goldsworthy and McFarlane 2002). A number of chemical compounds have since then been used over the years for clinical management of the infection and some of these include; streptomycin and para-aminosal-icylic acid (PAS) which presented bactericidal activity against TB in both humans and animals (Sotgiu et al., 2015). Mono-therapy of TB however, presented resistance challenges over time and as a result, combination therapy involving prescription of at least two drugs was adopted in 1952 based on the combination of streptomycin, PAS and isoniazid (Crofton, 1960, Crofton, 1969). Advances in TB management have been made since then with the discovery and use of rifampicin, isoniazid, ethambutol and pyrazinamide. Factors like effective drug combination and duration of medication are important determinants of eradication, prevention of relapse in infected individuals and also preventing development of drug resistant strains of TB.

Currently, TB drugs are divided into five key groups (Table 1.1) (WHO, 2018) i.e. Group one which are used as the first-line therapy include; isoniazid, rifampicin, ethambutol, pyrazinamide and rifabutin. Group two consists of the injectable agents such as aminoglycosides (kanamycin, amikacin and streptomycin) as well as peptide-based capreomycin. Group two drugs are used in combination with group three fluoroquinolone drugs (ofloxacin, levofloxacin, moxifloxacin and gatifloxacin.). Group four drugs consist of p-aminosalicylic acid, cycloserine, terizidone, ethionamide (ETH) and prothionamide, which are second line oral bacteriostatic drugs. Finally, there is group five whose efficacy is still unclear and is used only if the other four drug groups cannot be used (Zumla et al., 2013). Rifampicin, kanamycin and amikacin work by interfering with RNA synthesis and in the process inhibit the synthesis of host bacterial proteins. Capreomycin, clarithromycin, streptomycin and clofazimine have a similar mode of action that involves binding to the protein ribosomal subunits and in the process inhibiting translation (Zumla et al., 2013).

Generally, TB drugs have four major mechanisms of action i.e. inhibition of RNA synthesis, inhibition of protein synthesis, inhibition of cell wall biosynthesis and interference with the synthesis of cell membranes (Shi et al., 2007).

Table 1.1: Drugs used to treat the various forms of tuberculosis in their respective groups (Zumla et al., 2013).

TB drug group	**TB drug**
Group one: First-line oral agents	Isoniazid
	Rifampicin
	Ethambutol
	Pyrazinamide
Group two: Injectable agents	Kanamycin
	Amikacin
	Capreomycin
	Streptomycin
Group three: Fluoroquinolones	Levofloxacin
	Moxifloxacin
	Gatifloxacin
Group four: Oral bacteriostatic second-line agents	Para-amino salicylic acid
	Cycloserine
	Terizidone
	Ethionamide
	Prothionamide
Group five: Agents with unclear efficacy	Clofazimine
	Linezolid
	Amoxicillin

Linezolid disrupts bacterial growth by inhibiting the initiation process of protein synthesis whereas levofloxacin, moxifloxacin, and gatifloxacin work by inhibiting the DNA gyrase and topoisomerase IV which are responsible for separation of DNA that has been replicated (doubled) prior to bacterial cell division (Shi et al., 2007). Cycloserine and terizidone have a similar mode of action that involves inhibiting cell wall synthesis by competitively inhibiting two enzymes; l-alanine racemase and D-alanine ligase (Shi et al., 2007). Finally, isoniazid acts by inhibiting cell wall biosynthesis, a process that is catalysed by catalase-peroxidase (katG) which facilitates formation of activated isonicotinyl-NAD adduct that binds to InhA (enoyl-acyl carrier protein reductase) and inhibits the biosynthesis of mycolic acids. Mycolic acids are key components of the mycobacterial cell wall (Shi et al., 2007).

Currently, management of TB is done in two phases with varying durations of drug exposure depending on the susceptibility of the isolated TB strain. The initial stage in TB clinical management is the bactericidal phase that involves killing of the mycobacteria with high replication rate hence repairing and regenerating the pulmonary system. Drug adherence at this stage ensures non-infectivity of patients and reduces the chances of resistance (Sotgiu et al., 2015). The second stage is the continuation phase that is characterized by the elimination of the semi-dormant mycobacteria whose reduction in titers in the body and elimination is crucial in preventing drug resistance and relapse of the infection (Kumar & Kon, 2017).

As per the world health organization guidelines on TB management, treatment of newly diagnosed cases with drug-susceptible TB involves a two months intensive phase of a first line four-drug regimen comprised of rifampicin, isoniazid, pyrazinamide and ethambutol followed by a four to seven-months continuation phase consisting of rifampicin and isoniazid (Dheda & Sharma, 2019). It is important to note that the duration of treatment of the infected individuals is dependent on the type of TB infection. For example, patients with meningitis TB have a treatment duration of nine to twelve months (Nahid et al., 2016) whilst individuals with HIV co-infection not receiving antiretroviral treatment, are treated for a duration of seven months (Nahid et al., 2016).

The first line drugs have a range of side effects which are known to contribute to poor adherence to TB treatment and hence development of resistance overtime. Some of the side effects include; gastrointestinal symptoms and thrombocytopenia in rifampicin use, hepatitis and peripheral neuropathy in isoniazid use, gastrointestinal disturbances, and gout among others in pyrazinamide use (Kumar & Kon, 2017).

Treatment of drug resistant TB is also dependent on the type of drug to which the *M. tuberculosis* strain has developed resistance. Isoniazid resistant TB (Hr-TB) refers to the *M.*

tuberculosis strains that are resistant to isoniazid but susceptible to rifampicin. Individuals with confirmed Hr-TB are treated with rifampicin, ethambutol, pyrazinamide and levofloxacin for duration of six months. Rifampicin resistant TB (RR-TB) are *M. tuberculosis* strains that are resistant to rifampicin and as a result, the patients are enrolled to the multi-drug resistant TB (MDR-TB) regimen. According to the WHO, the latent TB infection is treated with isoniazid (INH) for six to nine months ("Updated TB guidelines raise upper age limit for treating latent disease", 2016). Like the first line drugs, second line TB regimen has a range of adverse effects including ototoxicity and hepatotoxicity (Liu et al., 2018) and these adverse effects together with pill burden especially in patients with other comorbidities play an important role in poor drug adherence and drug resistance development as they discourage patients from taking the medication.

1.2.5 TB drug resistance

TB drug resistance as the name suggests is resistance of *M. tuberculosis* to TB drugs. TB drug resistance classification is based on the number of TB drugs that a particular strain of *M. tuberculosis* is resistant to, these classifications include; Mono-resistant TB, Poly-drug resistance TB, Multi-drug resistant TB (MDR-TB) and Extensively Drug Resistant TB (XDR-TB). Mono-resistant TB is resistant to only one first-line drug while Poly-resistant TB is resistant to more than one first-line drug other than both isoniazid and rifampicin. Multi-drug resistant TB is resistant to at least both isoniazid and rifampicin. Further resistance to second-line drugs i.e. fluoroquinolones and at least one second-line injectable in addition to multi-drug resistance is called Extensive drug resistance (WHO, 2019). MDR-TB incidence is the highest (87%) in comparison to non-resistant TB (84%) among the 20 countries with the highest TB burden (WHO, 2018). Development of resistance to both Isoniazid and Rifampicin, the two most effective drugs against TB poses the greatest threat to the WHO global strategy of ending the TB epidemic by 2035 (WHO, 2018).

1.2.5.1 Multi drug resistant TB

Multidrug resistant tuberculosis (MDR-TB) is defined as resistance to at least rifampicin and isoniazid (WHO 2018). As of 2017, 3.5% of the newly TB tested cases globally and 18% of the previously tested TB cases had MDR/Rifampicin Resistant TB (WHO 2018). Research has shown that mutations in specific regions in drug target genes (Table 1.2) are the cause of MDR-TB, such mutations reduce sensitivity of *M. tuberculosis* to anti-tuberculosis therapy (Velayati, et al., 2009).

Table 1.2: **TB Resistance conferring genes** (Ando et al., 2014, Cohen et al., 2019).

Group	TB Drug	TB gene
First line drugs	Isoniazid	*katG, InhA,ahpC, kasA,* NDH and oxyA
	Ethambutol	*embCAB* operon
	Pyrazinamide	*pncA*
	Rifampicin	*rpoB*
Injectable gents	Kanamycin, Amikacin	*rrs* (rRNA gene)
	Capreomycin	*rrs* (rRNA gene)
Fluoroquinolones	Levofloxacin, moxifloxacin	*gyrA,*
	gatifloxacin	*gryB*
Oral bacteriostatic	Ethionamide	*ethA*
	Prothionamide	*ethR, inhA*
	Cycloserine	*Ald*
	Para-amino salicylic acid	*folC, dfrA, thyA, thyX, ribD*
	Terizidone	*Alr, ddl, cycA*
Agents with unclear efficacy	Clofazimine	*pepQ*
	Linezolid	*Rrl, rplC*
	Imipenem	*crfA*

These mutations can be single nucleotide polymorphisms, insertions and deletions (Zhang, 2017). Abuse of TB drugs through non-adherence can cause the *M. tuberculosis* population in the host to mutate for survival and to increase in mutation frequency which results in resistant *M. tuberculosis* (Ando et al., 2014). Mutations that cause resistance to rifampicin and isoniazid are well characterised while mutations leading to second line drug resistance are less understood and hence resistance to second line drugs is harder to predict using sequencing. Rifampicin kills *M. tuberculosis* by inhibiting RNA-polymerase. This action prevents RNA synthesis by physically blocking elongation, and thus preventing the synthesis of host bacterial proteins. Resistance to rifampicin is caused by mutations in the *rpoB* gene coding for RNA polymerase β subunit. When the β subunit structure changes, rifampicin cannot properly bind to RNA polymerase and hence becomes ineffective (Ando et al., 2014). In 96% of the *M. tuberculosis* isolates with resistance, mutations occurring in an 81bp spanning codons 507–533 of the *rpoB* gene lead to rifampicin resistance and are called the rifampicin resistance-determining region (Trauner et al., 2017).

Mutations in the *katG* gene or in the *InhA* gene are the main causes of isoniazid resistance although there are other gene mutations like *ahpC*, *kasA* and NDH responsible for a minority of the cases (Argyrou et al., 2006, Ando et al., 2014).

1.2.6 Isoniazid drug and its mode of action

Isoniazid (isonicotinic acid hydrazide) is a simple chemical compound composed of a pyridine ring and a hydrazide group that are important for the anti-mycobacterial activity against *M. tuberculosis* (Suarez et al., 2009). The drug is prepared through a reaction between 4-cyanopyridine and hydrazine hydrate in an aqueous alkaline environment at a temperature of 100°C. Isoniazid's molecular formula is $C_6H_7N_3O$, with a molecular weight of 137.14 g/mol, melting point of 171.4°C and LogP of −0.64 (Fernandes et al., 2017). The drug is commonly used in combination therapy as a first-line drug in the prevention and treatment of TB. Isoniazid can be administered through both the oral and intravenous route and is widely distributed in the body with a 61% approximated volume of distribution in body fluids and tissues (Weber et al, 1979).

Isoniazid together with rifampicin have long been known to be the key *M. tuberculosis* management drugs (Timmins et al., 2004) with isoniazid acting through inhibition of formation of the mycobacterial cell wall. Isoniazid is a pro-drug activated by the catalase peroxidase enzyme encoded by the *katG* gene in *M. tuberculosis*. Isoniazid activation leads to the formation of oxygen reactive species such as superoxide, hydrogen peroxide, peroxynitrite and the isonicotinoyl radical (Timmins et al., 2004). The isonicotinic acyl radical produced, naturally couples with NADH to form the nicotinoyl-NAD adduct (Figure 1.2). The formed abduct binds tightly to the enoyl-acyl carrier protein reductase inhA, blocking the action of fatty acid synthase. This process inhibits the synthesis of mycolic acids required for the synthesis of the *M. tuberculosis* cell wall (Timmins et al., 2004).

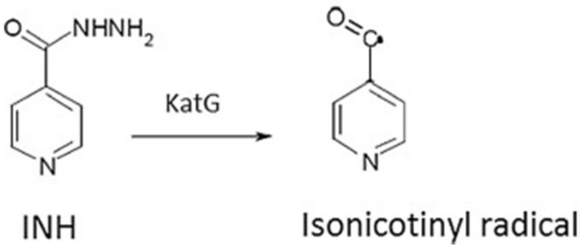


Figure 1.2: **Schematic diagram showing isoniazid reaction in the katG to form the nicotinoyl radical** (Adapted from Unissa et al., 2018).

Most clinical isolates with isoniazid resistance have presented with mutations in *katG* gene, resulting in elimination of or reduced peroxidase activity due to either isoniazid inactivation, or to a decreased affinity of the *M. tuberculosis* to isoniazid (Huard et al. 2003). Research shows that majority of the time in isoniazid resistance, there are either simple base pair changes or small deletions in the *katG* gene (Huard et al., 2003; Jena et al., 2015).

1.2.7 KatG (catalase peroxidase) enzyme

Catalase peroxidase (katG) protein is a homodimer with subunits of about 80kDa. The katG identical subunits are believed to be as a result of a gene duplication event and they consist of an N-terminal domain and a C-terminal domain each consisting of ~40 kDa (Bertrand et al., 2004). The N-terminal domain of the katG protein contains a heme binding motif which acts as the active site of the enzyme, whereas the C-terminal domain lacks this heme group and is believed to confer structural stability to the enzyme as shown in Figure 1.3.

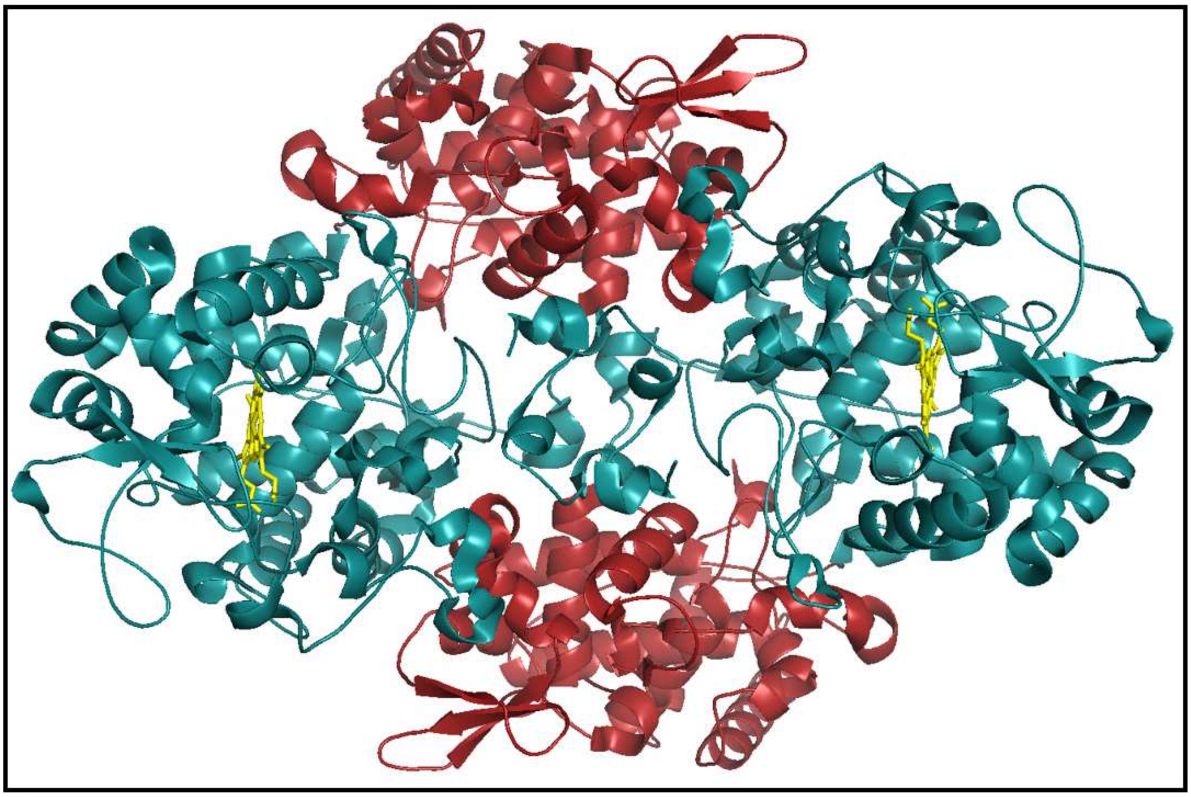


Figure 1.3: KatG (PDB ID: 2CCA) protein structure showing the N-terminal (cyan) and C-terminal (red) domains. N terminal domains contain the heme group (yellow) as viewed in PyMOL.

Although the katG enzyme has catalase activity, it is not homologous to catalases but rather has high sequence similarity to cytochrome c-peroxidase from yeast (Ccp) and ascorbate peroxidase (APX) and as a result, the katG enzyme is classified under the Class I of the superfamily of plant, fungal, and bacterial peroxidases (Welinder, 1992). The catalase function of the enzyme serves to protect the *M. tuberculosis* from hydro-peroxides and hydroxyl radicals present in the aerobic environment which are toxic to the *M. tuberculosis* (Bertrand et al., 2004). The peroxidase activity of the enzyme leads to activation of the first-line pro-drug isoniazid that is used in the treatment and management of TB.

The activation the TB first line drug isoniazid leads to the inhibition of mycobacteria cell wall synthesis and hence inhibition of the bacteria multiplication (Bertrand et al., 2004). Studies have shown that the active binding site for the isoniazid in the catalase peroxidase protein is a round the δ meso heme edge (Carpena et al., 2002) involving the residues Arginine, Histidine, Tryptophan, Tyrosine and Methionine.

1.2.8 KatG enzyme and *M. tuberculosis* isoniazid drug resistance

Isoniazid, being a pro enzyme, requires activation by a catalase peroxidase enzyme encoded from the *katG* gene in mycobacteria under aerobic conditions. As a result, mutations occurring in the *katG* gene that affect activity or encoding of catalase peroxidase enzyme hinder the anti-mycobacterial activity of the drug hence making the *M. tuberculosis* non-susceptible to isoniazid (Singh et al., 2018). Identification of the katG crystal structure (Bertrand et al., 2004) provided insight on the mode of action for catalase peroxidase and the structural similarities of the katG active site with the mono-functional Class I peroxidases such as cytochrome c peroxidase (C*c*P) (Edwards et al., 1987) which is also known to activate isoniazid (Pierattelli et al., 2004). Literature shows that the active site of the *M. tuberculosis* catalase peroxidases is made up of the heme-group surrounded by Arg-104, Trp-107, and His-108 residues on the distal side and His-270, Trp-321, and Asp-381 residues on the proximal side as shown in Figure 1.4. A three residue adduct between Trp, Tyr and Met was also observed to be conserved in all the catalase peroxidases (Bertrand et al., 2004).

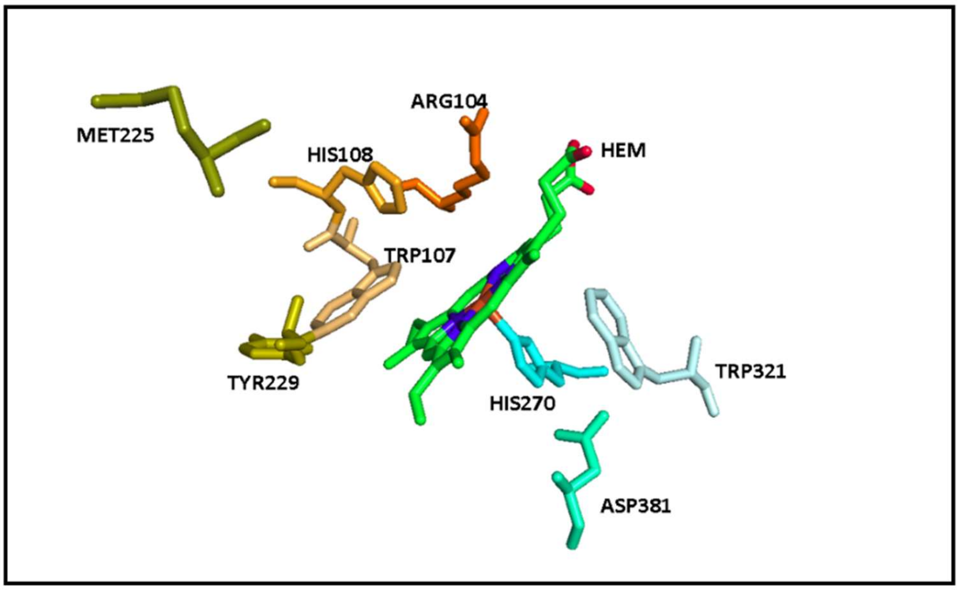


Figure 1.4: ***M. tuberculosis* katG (PDB ID: 2CCA) active site residues.** The active site consists of His-270, Trp-321, and Asp-381 residues on the proximal end and Arg-104, Trp-107, and His-108 residues on the distal end**.** Trp-107 forms adduct with Tyr-229 and Met-255 in all catalase peroxidases. Viewed in PyMOL.

Research has shown that the S315T mutation in the *katG* gene leads to the production of a catalase peroxidase enzyme capable of catalytic activity in the mycobacterium but unable to form the isoniazid-nicotinoyl adenine dinucleotide adduct to inactivate inhA and kasA activity, hence reducing the isoniazid anti-tubercular activity (Yu et al., 2003). A study done on 69 isoniazid resistant samples from three Brazilian states showed percentage resistances of 87.1%, 60.9%, and 60% attributed to the codon 315 *katG* mutation (Silva et al., 2003). In another study done on a 894bp central fragment of the *katG* gene in 212 samples across Africa; South Africa, Uganda and Sierra Leone, showed that 68% of the isoniazid resistant samples from South Africa and 54.1% of the isoniazid resistant samples in Uganda and Sierra Leone were as a result of codon 315 mutation in the *katG* gene (Haas et al., 1997). Missense AGC-ACC mutation (Ser-Thr), AGC-ATC (Ser-Ile) and AGC-CGC (Ser-Arg) were observed at codon 315 while at codon 328, TGG-TTG (Trp-Leu) and TGG-TGC (Trp-Cys) were observed (Haas et al., 1997). Furthermore, a study done in Iran on 25 isoniazid resistant strains showed 56% of the resistance was as a result of mutations of the *katG* gene position 315 (Bostanabad, 2011). From all this data, codon 315 is shown to play a significant role in isoniazid resistance in *M. tuberculosis*. Other documented *katG* gene mutations leading to isoniazid resistance are T275C (Pretorius et al., 1995), L587M (Saint-Joanis et al., 1999), H108E (Mdluli, 1998) and N138S (Morris et al., 1995).

It is important to note that besides the *katG* gene mutations, studies have shown that point mutations in the *inhA* gene at position −15C/T increase inhA mRNA level in the *M. tuberculosis* wild type which results in inhA overexpression and an eight-fold increase in resistance to isoniazid (Larsen et al., 2002). As per Lari et al., (2006), of the 45 isoniazid resistant samples analysed in Italy, 37.8% were as a result of *katG* gene mutation at codon position 315 while substitutions C-T at position 15 and G-T at position 24 in inhA accounted for 20.0% and 2.2% of the isoniazid resistance respectively. The *ahpC* and *kasA* gene mutations accounted for 2.2% and 4.4% of the isoniazid resistance. According to Seifert et al (2015), a combination of *katG* S315T and *fabGI-inhA* c-15t mutations account for 83% of the isoniazid resistance.

Data from these studies agrees with most literature that indicates the majority of the isoniazid resistance in *M. tuberculosis* being as a result of mutations in the *katG* gene especially at codon position 315. Even with that compelling evidence, the mechanism of isoniazid resistance by *M. tuberculosis* as a result of *katG* gene mutations still remains greatly unexplored and an understanding of these resistance mechanisms is bound to provide great insight in next generation drug design and discovery.

1.2.8 High confidence (HC) *katG* gene mutations

High confidence mutations are the genetic variations in katG protein that are highly correlated with the observed phenotypic resistance of *M. tuberculosis* to isoniazid. A systematic literature review of the association of sequencing and phenotypic drug susceptibility testing (DST) data for *M. tuberculosis* (Miotto et al., 2017) identified katG protein mutations: S315I, S315N, S315T, pooled frameshifts and premature stop codons as the high confidence mutations based on likelihood ratios (LRs > 1) i.e. the strength of association between the mutation and the phenotypic drug resistance, and odds ratios (ORs > 1) i.e. the measure of association between exposure and outcome. A study done using 129 *M. tuberculosis* isolates from WHO/TDR-TB Strain Bank and 47 isolates from the MDR-TB patients from the Bangladesh (Lempens et al., 2018) classified katG protein mutations; S315T and S315N as high-confidence mutations basing on the PhyResSE (Phylo-Resistance Search Engine) (Ssengooba et al., 2016) variant catalogue which classifies high confidence mutations as those for which strong experimental evidence is available linking them to phenotypic isoniazid drug resistance. The high-confidence mutations mentioned in these studies are further supported by the data from the TB Drug Resistance Mutation Database (TBDReaMDB) which classified 15 out of the 273 *katG* mutations as high-confidence mutations leading to isoniazid drug resistance as of August 2019 (Sandgren et al., 2009). TBDReaMDB classifies mutations as high-confidence using the same criteria as the PhyResSE which relies on the experimental data from the reviewed literature associating the mutations to phenotypic drug resistance. Furthermore, different studies i.e. Yu et al., 2003, Silva et al., 2003, Haas et al., 1997 and Bostanabad, 2011 done in different parts of the world attribute the observed phenotypic isoniazid drug resistance to katG protein mutation S315T. This study focused on understanding the mechanism of resistance of the documented high confidence katG protein mutations leading to isoniazid resistance (Table 1.3).

Table 1.3: **High confidence mutations in the *M. tuberculosis* katG and the supporting literature**.

Mutation	Supporting literature	Country of study
S140R	Purkan et al., 2016	Indonesia
S140N	Ramaswamy et al., 1998	USA, Spain
G279D	Morlock et al., 2003	Brazil
G285D	Abe et al., 2008	Japan
S315I	Lempens et al., 2018	Bangladesh
	Haas et al., 1997	Lesotho, Sierra Leone
S315N	Lempens et al., 2018	Bangladesh
	Ssengooba et al., 2016	Uganda
	Haas et al., 1997	South Africa (Free state, Gauteng), Sierra Leone
S315R	Lipin et al., 2007	Russia
S315T	Lempens et al., 2018	Bangladesh
	Ssengooba et al., 2016	Uganda
	Heym et al., 1995	Mali, Ivory coast, France, South Africa
	Marttila et al., 1996	Finland
S316D	Hazbon et al., 2006	South Africa
S457I	Bolotin et al., 2009	Canada
G593D	Ramaswamy et al., 1998	USA, Spain

1.3 PROBLEM STATEMENT

As per 2017, an estimated 10 million people contracted TB and of these, 3% were from South Africa. An estimated 558 000 MDR/RR-TB cases were reported in 2017 with 3.5% of those being new cases and 18% of the old TB cases reverting to MDR/RR-TB (WHO, 2018). South Africa is ranked among the 30 high TB burdened countries with an estimated 3.4% incidence in new MDR/RR-TB cases and a 7% incidence of MDR/RR-TB in the old TB cases as per 2017 (WHO,2018). In order to achieve the World Health Organization target of an 90% reduction in TB related deaths and an 80% reduction the TB incidence rate by 2030 (WHO

2018), a lot of research needs to be done in understanding of the mechanism of TB drug resistance and innovation of new avenues to manage the TB infection. Despite knowing the gene mutations; especially the *katG* gene mutation leading to isoniazid resistance, the mechanism by which these mutations in the *katG* gene lead to the reduced activity of catalase peroxidase enzyme and isoniazid resistance is still an open discussion. Currently, only a few katG protein mutations occurring at codon positions 315 (Unissa et al., 2018), 275 (Pretorius et al., 1995), 587 (Saint-Joanis et al., 1999), 108 (Mdluli, 1998) and 138 (Morris et al., 1995) and a few others have been studied leaving the rest unexplored. This study will focus on understanding the mechanism of isoniazid drug resistance by the mutations in the katG protein with strong experimental evidence (high confidence mutations) linking them to the phenotypic isoniazid drug resistance in *M. tuberculosis*.

1.4 AIM OF THE STUDY

This study aims to determine the mechanism of resistance of katG protein mutations against isoniazid in *M. tuberculosis* using bioinformatics approaches. The understanding of the katG mutations mechanism of resistance to isoniazid will provide insights in the role of the mutations in isoniazid resistance and also inform new approaches to MDR-TB drug design.

1.5 OBJECTIVES

1. Retrieval of the TB protein drug target structure from PDB database for mapping of mutation and further analysis.
2. Identification and retrieval high confidence katG isoniazid drug resistant mutations from TB Drug Resistance Database and literature for modelling and further analysis.
3. Modelling the wild type and the respective variants katG protein structures for molecular dynamics studies.

4. Performing molecular dynamics (MD) simulations and trajectory analysis on the modelled wild type and variants katG proteins to study the protein structural behaviour over time.
5. Dynamic residue network (DRN) analysis to investigate the effect of the variations on the protein communication network.
6. Merging MD, DRN data to understand the resistance mechanism of the variants.

CHAPTER TWO

HOMOLOGY MODELLING OF WILD TYPE KAT-G PROTEIN AND VARIANTS

2.1 INTRODUCTION

Homology modelling, also known as comparative modelling, is one of the techniques used in studying and calculating the 3D structure of proteins. Besides homology modelling, other techniques used are; X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy (EM). Unlike in these other methods (X-ray crystallography, NMR and EM) which require expensive equipment and a lot of time to obtain protein structures, homology modelling uses the concept of sequence similarity to calculate and model protein structures. As a result, homology modelling can be done using a set of computers and algorithms within hours.

Homology has had a number of definitions from different schools of thought over the years however, the most applicable definition was by Osche in 1982 who defined it as the non-random similarities between complex structures as a result of common genetic information and ancestry (Haszprunar, 1992). As a consequence of this, homology modelling method relies on the conservation of the protein structure during evolution (Chothia et al., 1986). The principle behind homology modelling is that if two protein sequences have a high enough sequence similarity, then it is highly likely that they have the same secondary structure (homologous). Two protein sequences are said to be homologous depending on their sequence identity score and the length of the sequences. Irrespective of the sequence identity score between two sequences, two sequences are said to be homologous if they also satisfy a minimum sequence length requirement (Rost, 1999). The argument being that, the shorter the sequence, the higher the chances that the alignment is as a result of random chance. Since protein sequences are made up of 20 amino acids, two unrelated sequences can match up to 5% of the residues as a result of random chance. Hence, the longer the sequence the less likely the sequence match being as a result of chance. Therefore, shorter sequences have higher cut offs for inferring sequence homology while for sequences aligned a full length (i.e. 100 residues long), a sequence identity score of 30% or higher can safely be regarded as close homology (Rost, 1999).

Exploiting that fact, homology modelling uses protein homologs having a sequence similarity of ≥ 30% to the target sequence with existent crystal structures as a template to calculate the 3D structure of the target protein sequence. The process of homology modelling can be broken down to four steps i.e. template identification, sequence alignment, model construction using template 3D coordinates and validation of the modelled target 3D structure.

2.1.1 Template identification

The first step and one of the most crucial steps in homology modelling is identification of the protein template. In homology modelling, the backbone atoms of the template are used to model the target structure and therefore the identified template must have a high sequence similarity (≥ 30%) or at least secondary structure conservation to the target sequence. It should also have an already determined crystal structure of good quality (Xiang, 2006). Homologs to the target sequence can be identified by submitting the target sequence to programs such as BLAST (Altschul et al., 1990), HHpred (Soding et al., 2005), PRIMO (Hatherley et al., 2016), PSI-BLAST (Altschul, 1997) and ScanPS (Barton, 1992). Programs like PSI-BLAST, HHpred and ScanPS employ complex homology search tools. ScanPS uses a modified version of the Smith–Waterman algorithm optimized for parallel processing. PSI-BLAST builds profiles and performs database searches in an iterative fashion similar to ScanPS whereas HHpred uses Hidden Markov Models (HMMs) to retrieve distantly related homologs to the target sequence.

The template search tools provide the user with information of all the target homologs together with the sequence similarity, identity and alignment scores. Choice of the template sequence is dependent on its identity to the target sequence and its quality. The template's quality can be accessed using PDB validation features i.e. wwPDB validation, resolution, R-value free and R-value work. A template with wwPDB validation percentile ranks in blue, high resolution and a low R-value is considered as a good template. In addition to template quality, presence of appropriate ligands and/or cofactors of interest in the template structure is another important factor in template selection. The next step after template identification is target and template sequence alignment.

2.1.2 Sequence alignment

In this step, the target and selected template protein sequences are aligned to obtain the most accurate alignment for creating the model. Since the backbone Cα atoms of the template sequence are used in modelling the target structure, this makes the sequence alignment step the most crucial step in the modelling process (Prasad et al., 2003). A sequence alignment with 100% coverage and a similarity score of ≥ 30% is considered high enough for modelling target structures (Xiang, 2006). A number of alignment programs can be used to obtain a custom alignment between the target and template protein sequences and these include; Clustal (Higgins et al., 1988), Muscle (Edgar, 2004) and TCoffee (Notredame et al., 2000) among others. Other programs like 3DCoffee (O'Sullivan et al., 2004) and FUGUE (Shi et al., 2001) can be used to generate structural alignments between the target and template proteins especially for distantly related homologs. It is advised to manually optimize the target and template alignment while accounting for biological information to improve on the model quality. Homology modelling also allows the user to use more than one template for target modelling to improve on the model quality (Larsson et al., 2008). In this case, missing sequence sections in one template are compensated for by the sections from the second template creating a more accurate alignment.

2.1.3 Target modelling

During modelling, the backbone atoms and the similar residues between the template and target sequences are extracted and used to model the target structure backbone and side chains while respecting the spatial restraints between the atoms. For sections with missing residues due to insertions and deletions, only the backbone atoms are copied (Fiser, 2010). A number of homology modelling programs are available and they include web-based servers; PRIMO, SWISS-MODEL (Waterhouse et al., 2018) and HHpred. Standalone programs like MODELLER (Sali et al., 1993) and WHATIF (Vriend, 1990) are also available for homology modelling. Available modelling methods include; the segment matching method and satisfying spatial restraints method to mention but a few. MODELLER uses the satisfying spatial restraints method where the distances or optimization techniques are used to satisfy the spatial restraints. In this study, MODELLER 9.22 was used for modelling the wild type and variant katG structures. Model validation and refinement follows the modelling step.

2.1.4 Model validation

After modelling step, the 3D models needs to be checked for the accuracy. This has been done by looking at various metrics including if their structures satisfy the protein physicochemical rules, stereo-chemical properties, chemical correctness, planarity and also by checking the atomic interaction energies in comparison to those of the protein X-ray structures in the database. Different validation tools focus on specific model properties, hence it is good practice to validate the model using more than one validation program in a bid to get consensus results.

Model validation can be classification as global validation and local validation. Global validation focuses on the quality of the entire protein structure. In MODELLER, global quality measurement is assessed using the discrete optimized protein energy (DOPE) score. The DOPE score is an atomic distance-dependent statistical potential from a sample of native structures. The score takes into account the finite and spherical shape of the native structures and was derived using a non-redundant set of 1472 crystallographic structures (Shen & Sali, 2006). The DOPE score is usually normalized to create a uniform scale, facilitating quality comparison between different protein structures.

Unlike global validation, local quality validation focuses on individual residues or local regions in the modelled structure i.e. fit to the local electron density map or steric clashes between atoms. Local validation is particularly important especially in identifying and improving problematic sections of the protein structure i.e. the loop regions and binding site. Good model validation takes into account both the biological data and results from analytical validation tools. A number of programs are available for both global and local validation and these include; WHAT_CHECK (Hooft et al., 1996), WHAT IF (Vriend, 1990), ProSA (Sippl, 1993), PROCHECK (Laskowski et al., 1993) and VERIFY3D (Eisenberg et al., 1997) to mention but a few. In this study, ProSA, PROCHECK and VERIFY3D tools were used to validate the generated models.

2.1.4.1 PROCHECK

PROCHECK is a protein structure validation program that checks the physiochemical properties of the structures which include; psi and phi angles, chirality, bond angles and bond length. The program uses parameters derived from the Morris et al., (1992) and bond length and angles (Engh and Huber, 1991) derived from the analyzed small molecule structures from Cambridge structure database. Ultimately, PROCHECK assesses how normal and abnormal the protein structure is compared to the well-refined, high-resolution structures from Cambridge structure database. The input file is protein structure file while the output is a number of plots including a detailed residue by residue listing (Laskowski et al., 1993).

2.1.4.2 ProSA

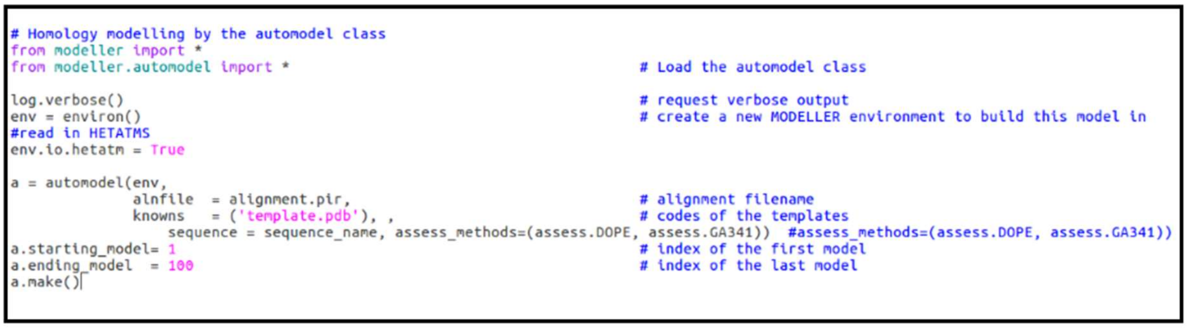
ProSA is a web-based program that validates protein structures by calculating the overall quality score of the structure. ProSA displays the obtained score in comparison with all the experimentally determined structures of similar size present in the protein databank categorized by their techniques of study i.e. X-ray crystallography and Nuclear Magnetic Resonance) (Wiederstein and Sippl, 2007). The input file is a protein coordinate file or the PDB ID and output is a z-score of the overall quality of the structure and a graph of respective residue energy scores in the 10 residue and 40 residue windows.

2.1.4.3 VERIFY3D

VERIFY3D like ProSA is also a web-based server that validates the quality of the protein structures using a statistical approach. The server uses a precomputed database containing 18 environmental profiles from high resolution structures. The profiles used are based on the solvent exposure and secondary structures of the proteins. If the structure under validation falls within the environmental profiles, it receives a high score. A structure is considered to have a favorable environment if the score is above 0 (Xiong, 2006). The output of the validation is a 2-D graph showing the folding quality of the protein structure.

2.1.5 MODELLER

MODELLER is a computer program that is used for modelling the 3D structures of proteins using homology modelling. Basically, three ingredients are needed for homology modelling when using MODELLER i.e. the target protein sequence, the template protein sequence and an alignment of the two sequences (Sali et al., 1995). A MODELLER script (Figure 2.1) containing the name of the alignment, the templates and the desired number of models to be made is run to create the specified models.



```
# Homology modelling by the automodel class
from modeller import *
from modeller.automodel import *                                        # Load the automodel class

log.verbose()                                                           # request verbose output
env = environ()                                                         # create a new MODELLER environment to build this model in
#read in HETATMS
env.io.hetatm = True

a = automodel(env,
              alnfile  = alignment.pir,                                 # alignment filename
              knowns   = ('template.pdb'), ,                            # codes of the templates
                  sequence = sequence_name, assess_methods=(assess.DOPE, assess.GA341))  #assess_methods=(assess.DOPE, assess.GA341))
a.starting_model= 1                                                     # index of the first model
a.ending_model  = 100                                                   # index of the last model
a.make()
```

Figure 2.1: **MODELLER script used in homology modelling**. The script takes an alignment file and templates IDs as augments and returns the specified number of models, in this case 100.

MODELLER models protein structures by optimally satisfying the spatial constraints expressed by the template and target sequence alignment using CHARMM energy terms. The program uses a set of commands that are edited to suit the structures of interest. Besides homology modelling, MODELLER can be used for the comparison of protein structures, *de novo* modelling of loops in protein structures and optimization of protein structure models (Sali et al., 1993).

2.2 METHODOLOGY

2.2.1 Template identification

The *M. tuberculosis* katG sequence was retrieved from the National Center for Biotechnology Information (NCBI) database by searching using key words: *M. tuberculosis* and katG protein sequence. Results from the search linked the *M. tuberculosis* katG sequence to the X-ray crystal structure in the PDB database with a PDB ID; 2CCA. The PDB protein structure was identified as the wild type structure of the *M. tuberculosis* katG enzyme. The *M. tuberculosis* katG homolog search was done using BLASTP (McGinnis et al, 2004) and HHpred (Zimmermann et al., 2018) similarity search tools. The templates were then selected based on sequence similarity, E-value, query sequence coverage and presence of co-crystallized ligand of interest (isoniazid) (Table 2.1). Two templates with the highest sequence similarity to the wild type katG, lowest E-value, good quality and coverage and with a co-crystallized isoniazid ligand were selected.

2.2.2 Template-target alignment

The *M. tuberculosis* sequence and the templates (2CCA and 2V2E) protein sequences were retrieved from NCBI and aligned using TCoffee alignment tool (Notredame et al, 2000). The alignment file was then edited to retain all the coordinate information from template PDB ID: 2CAA and only the ligand (isoniazid) coordinate information from template PDB ID: 2V2E. Finally, the alignment was then converted to MODELLER format and saved as a pir file for the homology modelling step.

For the katG variants, the respective mutations were introduced in the target sequence using a Python script. The alignment between the modelled wild type and the respective variants was done using TCoffee alignment tool and the generated alignment file was converted to MODELLER format and saved as a pir file.

2.2.3 Target homology modelling

For both the wild type and the respective variants, MODELLER version 9.22 was used to generate the 3D structures using the generated alignment files. The respective alignment pir files were passed to MODELLER together with their specified template PDB files. 100 models were generated for each variant and wild type using auto model and slow refinement MODELLER options.

2.2.4 Model validation

From the 100 models of each structure, the best model was selected based on MODELLER normalized z-DOPE score. The model structures with the lowest normalized z-DOPE scores were considered as the best and these proceeded to be validated with ProSA, VERIFY3D and PROCHECK validation programs using default settings.

2.3 RESULTS AND DISCUSSION

2.3.1 KatG wild type homology modelling

2.3.1.1 Template identification

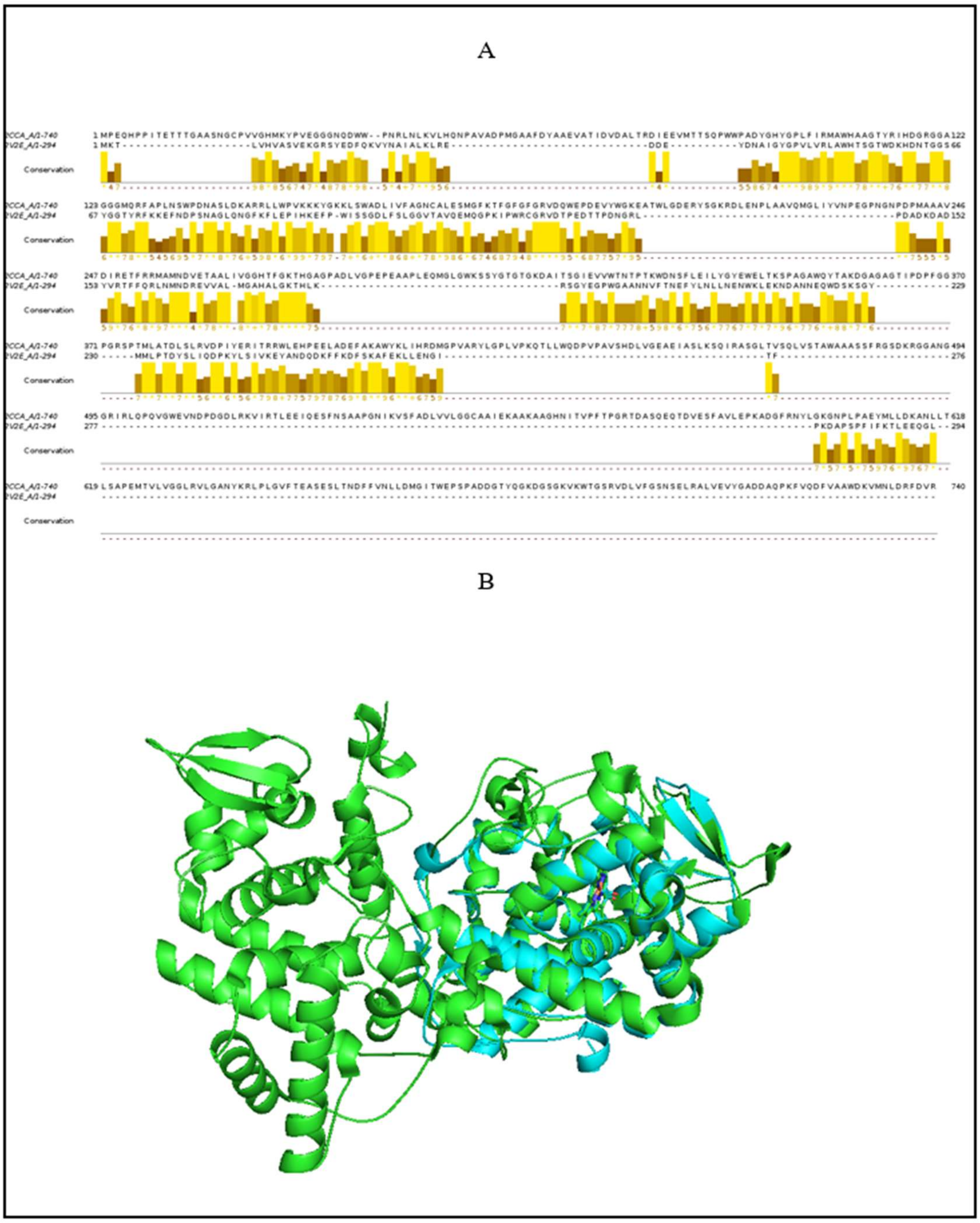
The *M. tuberculosis* and *S. cerevisiae* protein structures were identified as the templates for *M. tuberculosis* katG-isoniazid complex structure modelling based on their sequence identity to the target, structure resolution, coverage (Table 2.1) and the predicted isoniazid coordination residues in the *M. tuberculosis* katG active site.

Table 2.1: KatG model templates global quality scores.

	M. tuberculosis	***S. cerevisiae***
PDB ID	2CAA	2V2E
E-value	0.0	3E-18
R-value Free	0.225	0.187
R-value Work	0.199	0.165
Resolution	2 Å	1.68 Å
z-DOPE score	-1.468	-2.528

Even with a low sequence identity of *S. cerevisiae* cytochrome c peroxidase structure to *M. tuberculosis* katG, the two structures have secondary structure and active site conservations (Figure 2.2).

Isoniazid in the *M. tuberculosis* katG protein is proposed to interact with residues; His-108, Arg-104 and Try-107 on the distal end of the heme group (Henriksen et al., 1998, Aitken et al., 2001and Bertrand et al., 2004) as is the case in the *S. cerevisiae*.

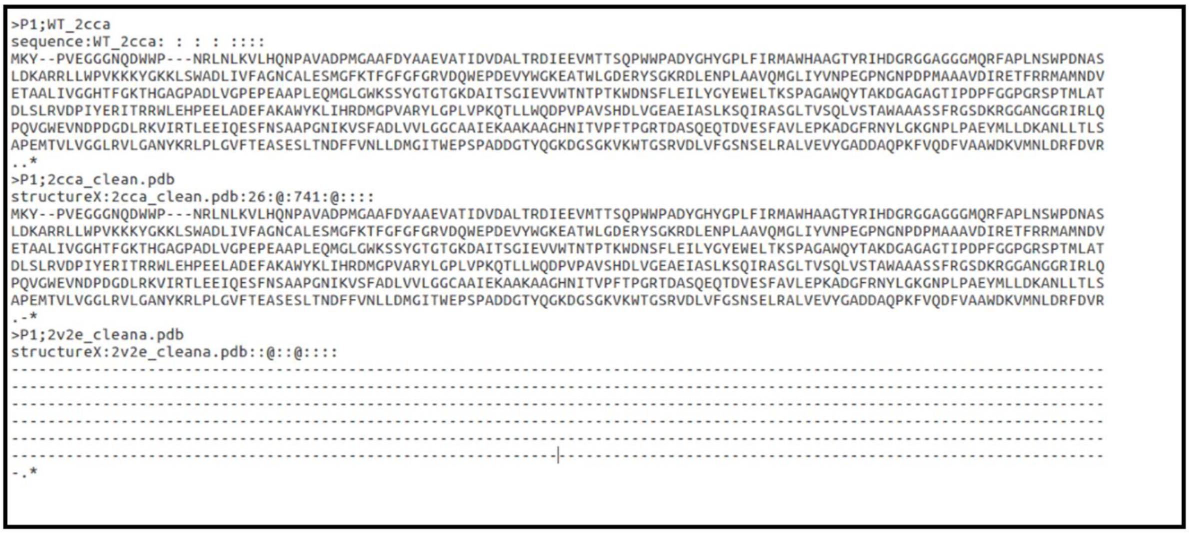


```
2CCA_A/1-740   1 MPEQHPPITETTTGAASNGCPVVGHMKYPVEGGGNQDWW--PNRLNLKVLHQNPAVADPMGAAFDYAAEVATIDVDALTRDIEEVMTTSQPWWPADYGHYGPLFIRMAWHAAGTYRIHDGRGGA 122
2V2E_A/1-294   1 MKT-------------------LVHVASVEKGRSYEDFQKVYNAIALKLRE-----------------------------DDE----------YDNAIGYGPVLVRLAWHTSGTWDKHDNTGGS 66
Conservation

2CCA_A/1-740 123 GGGMQRFAPLNSWPDNASLDKARRLLWPVKKKYGKKLSWADLIVFAGNCALESMGFKTFGFGFGRVDQWEPDEVYWGKEATWLGDERYSGKRDLENPLAAVQMGLIYVNPEGPNGNPDPMAAAV 246
2V2E_A/1-294  67 YGGTYRFKKEFNDPSNAGLQNGFKFLEPIHKEFP-WISSGDLFSLGGVTAVQEMQGPKIPWRCGRVDTPEDTTPDNGRL-------------------------------------PDADKQAD 152
Conservation

2CCA_A/1-740 247 DIRETFRRMAMNDVETAALIVGGHTFGKTHGAGPADLVGPEPEAAPLEQMGLGWKSSYGTGTGKDAITSGIEVVWTNTPTKWDNSFLEILYGYEWELTKSPAGAWQYTAKDGAGAGTIPDPFGG 370
2V2E_A/1-294 153 YVRTFFQRLNMNDREVVAL-MGAHALGKTHLK-----------------------------------RSGYEGPWGAANNVFTNEFYLNLLNENWKLEKNDANNEQWDSKSGY----------- 229
Conservation

2CCA_A/1-740 371 PGRSPTMLATDLSLRVDPIYERITRRWLEHPEELADEFAKAWYKLIHRDMGPVARYLGPLVPKQTLLWQDPVPAVSHDLVGEAEIASLKSQIRASGLTVSQLVSTAWAAASSFRGSDKRGGANG 494
2V2E_A/1-294 230 -----MMLPTDYSLIQDPKYLSIVKEYANDQDKFFKDFSKAFEKLLENGI-----------------------------------------------TF------------------------- 276
Conservation

2CCA_A/1-740 495 GRIRLQPQVGWEVNDPDGDLRKVIRTLEEIQESFNSAAPGNIKVSFADLVVLGGCAAIEKAAKAAGHNITVPFTPGRTDASQEQTDVESFAVLEPKADGFRNYLGKGNPLPAEYMLLDKANLLT 618
2V2E_A/1-294 277 --------------------------------------------------------------------------------------------------------PKDAPSPFIFKTLEEQGL-- 294
Conservation

2CCA_A/1-740 619 LSAPEMTVLVGGLRVLGANYKRLPLGVFTEASESLTNDFFVNLLDMGITWEPSPADDGTYQGKDGSGKVKWTGSRVDLVFGSNSELRALVEVYGADDAQPKFVQDFVAAWDKVMNLDRFDVR 740
2V2E_A/1-294     --------------------------------------------------------------------------------------------------------------------------
Conservation
```

Figure 2.2 Secondary structure conservation between *M. tuberculosis* and *S. cerevisiae* katG. (**A**) Shows the level of sequence conservation in the TCoffee alignment as viewed in Jalview (Clamp et al., 2004). The dotted lines indicate the insertions in 2CCA katG structure that are not present in the 2V2E structure. (**B**) Shows structural conservation between the *M. tuberculosis* (green) and *S. cerevisiae* (cyan) katG structures.

2.3.1.2 Template-target alignment

The target and template sequences were aligned using TCoffee, an alignment programme that makes use of a library of heterogeneous data sources to do multiple sequence alignment through a progressive method. The generated alignment was then edited to only include the ligand section from second template PDB ID: 2V2E (Figure 2.3) and all the residue information from first template PDB ID 2CCA.



```
>P1;WT_2cca
sequence:WT_2cca: : : : ::::
MKY--PVEGGGNQDWWP---NRLNLKVLHQNPAVADPMGAAFDYAAEVATIDVDALTRDIEEVMTTSQPWWPADYGHYGPLFIRMAWHAAGTYRIHDGRGGAGGGMQRFAPLNSWPDNAS
LDKARRLLWPVKKKYGKKLSWADLIVFAGNCALESMGFKTFGFGFGRVDQWEPDEVYWGKEATWLGDERYSGKRDLENPLAAVQMGLIYVNPEGPNGNPDPMAAAVDIRETFRRMAMNDV
ETAALIVGGHTFGKTHGAGPADLVGPEPEAAPLEQMGLGWKSSYGTGTGKDAITSGIEVVWTNTPTKWDNSFLEILYGYEWELTKSPAGAWQYTAKDGAGAGTIPDPFGGPGRSPTMLAT
DLSLRVDPIYERITRRWLEHPEELADEFAKAWYKLIHRDMGPVARYLGPLVPKQTLLWQDPVPAVSHDLVGEAEIASLKSQIRASGLTVSQLVSTAWAAASSFRGSDKRGGANGGRIRLQ
PQVGWEVNDPDGDLRKVIRTLEEIQESFNSAAPGNIKVSFADLVVLGGCAAIEKAAKAAGHNITVPFTPGRTDASQEQTDVESFAVLEPKADGFRNYLGKGNPLPAEYMLLDKANLLTLS
APEMTVLVGGLRVLGANYKRLPLGVFTEASESLTNDFFVNLLDMGITWEPSPADDGTYQGKDGSGKVKWTGSRVDLVFGSNSELRALVEVYGADDAQPKFVQDFVAAWDKVMNLDRFDVR
..*
>P1;2cca_clean.pdb
structureX:2cca_clean.pdb:26:@:741:@::::
MKY--PVEGGGNQDWWP---NRLNLKVLHQNPAVADPMGAAFDYAAEVATIDVDALTRDIEEVMTTSQPWWPADYGHYGPLFIRMAWHAAGTYRIHDGRGGAGGGMQRFAPLNSWPDNAS
LDKARRLLWPVKKKYGKKLSWADLIVFAGNCALESMGFKTFGFGFGRVDQWEPDEVYWGKEATWLGDERYSGKRDLENPLAAVQMGLIYVNPEGPNGNPDPMAAAVDIRETFRRMAMNDV
ETAALIVGGHTFGKTHGAGPADLVGPEPEAAPLEQMGLGWKSSYGTGTGKDAITSGIEVVWTNTPTKWDNSFLEILYGYEWELTKSPAGAWQYTAKDGAGAGTIPDPFGGPGRSPTMLAT
DLSLRVDPIYERITRRWLEHPEELADEFAKAWYKLIHRDMGPVARYLGPLVPKQTLLWQDPVPAVSHDLVGEAEIASLKSQIRASGLTVSQLVSTAWAAASSFRGSDKRGGANGGRIRLQ
PQVGWEVNDPDGDLRKVIRTLEEIQESFNSAAPGNIKVSFADLVVLGGCAAIEKAAKAAGHNITVPFTPGRTDASQEQTDVESFAVLEPKADGFRNYLGKGNPLPAEYMLLDKANLLTLS
APEMTVLVGGLRVLGANYKRLPLGVFTEASESLTNDFFVNLLDMGITWEPSPADDGTYQGKDGSGKVKWTGSRVDLVFGSNSELRALVEVYGADDAQPKFVQDFVAAWDKVMNLDRFDVR
.-*
>P1;2v2e_cleana.pdb
structureX:2v2e_cleana.pdb::@::@::::
------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------------
-.*
```

Figure 2.3: **MODELLER input alignment file (pir file)**. The *M. tuberculosis* katG target sequence is labeled as WT-2cca, the templates are labelled 2cca_clean.pdb and 2v2e_cleana.pdb. The alignment sequence for 2v2e_cleana.pdb was edited to only include the isoniazid ligand information.

2.3.1.3 Model building

The wild type katG models were calculated and refined using the MODELLER slow refinement method. 100 models were generated for the wild type protein, and the best selected based on the normalized z-DOPE score. The best model had a score of -1.006. Visualization using PyMOL showed polar interactions between isoniazid and Asp-137, Ser-315 and Val-230 residues in the modelled *M. tuberculosis* katG active site (Figure 2.4) unlike in the template PDB ID: 2V2E where the isoniazid interacted with residues Pro-132 and Arg-48. This could have been as a consequence of model optimization, a MODELLER function aimed at obtaining the model with the least conformation energy.

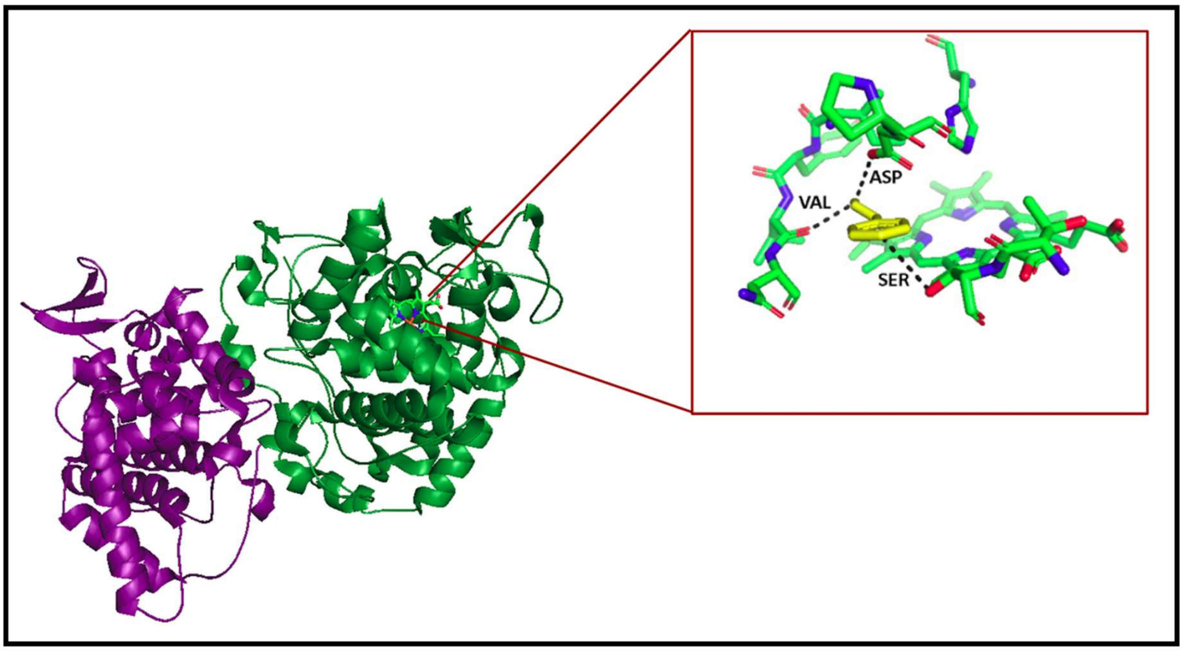


Figure 2.4: Modelled *M. tuberculosis* katG protein in complex with isoniazid (yellow). N-terminal: green, C- terminal: purple. Isoniazid (yellow) has polar contacts with Asp-137, Val-230 and Ser-315.

Serine residue interaction with isoniazid in the model katG has a well-documented SNP (S315T) known for conferring isoniazid resistance in *M. tuberculosis* (Lempens et al., 2018). A 2D residue interaction analysis using LigPlus (Wallace et al., 1995) showed isoniazid hydrogen bond interactions with residues Asp-137 and Val-230 (Figure 2.5) in the katG wild type as was indicated in PyMOL.

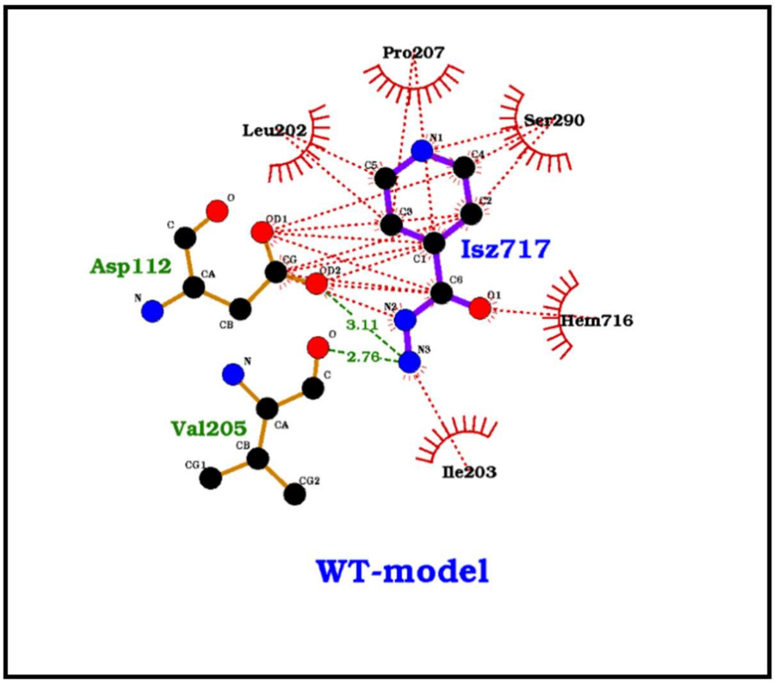


Figure 2.5: **LigPlot of isoniazid's coordinating residues in katG model**. Isoniazid had hydrogen bond interactions (green dotted lines) between Val-230 and Asp-137. The drug also was noticed to have hydrophobic interactions (red dotted lines) with Leu-227, Pro-232, Ser-315, Ile-228 and the heme group. The residue numbers in the Figure 2.5 are model numbers.

2.3.1.4 Model validation

Validation of the katG model with PROCHECK showed 93.2% of the residues in most favourable regions, 6.6% in additionally allowed region and none in the disallowed regions. ProSA scored the model a z-score of –11.72 which was well within katG size similar protein structures present in the Protein Data Bank. Finally, VERIFY3D showed 95.24% of the wild type katG model residues to be above the average 3D-ID score of ≥0.2. An average 3D-ID score of ≥0.22 indicates that the sequence of the katG model wild type is compatible with the protein model as per the 3D profile of the structure (Figure 2.6).

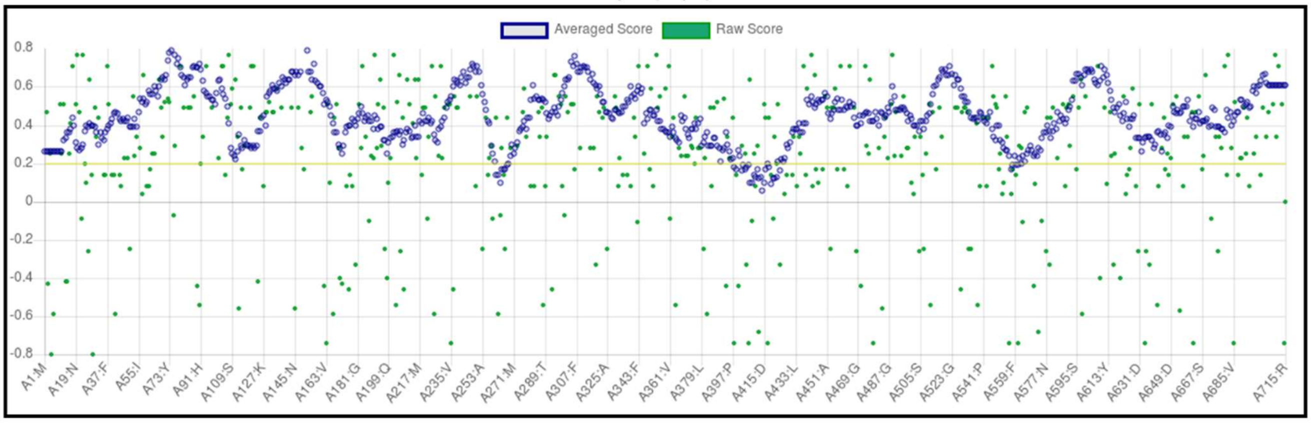


Fig 2.6: VERIFY3D validation result for the *M. tuberculosis* katG wild type model.

2.3.2 Variant katG homology modelling

A vast number of mutations leading to katG isoniazid resistance have been documented over time and as of August 2019, the TB drug resistance database alone had 273 such documented mutations. A number of these mutations in the TB drug resistance database have compelling evidence (literature) supporting the notion that they lead to katG isoniazid resistance and such mutations have been referred to as high confidence mutations.

In this research, 11 high confidence mutations located in seven different positions of the katG protein were studied (Table 2.2). Isoniazid resistance conferring high confidence mutations in the katG are not only located near the active site but also more than 10 Å away from the active site in the C-terminal domain (Figure 2.7). Mutations at position 315 are part of N-terminal loop region 10.4 Å from the Fe in the centre of the heme. Residues at position 315 are believed to form part of the isoniazid access channel to the active site. Mutations G279D, G285D and G316D from part of the N-terminal loop region and in respect to the heme group, they are 19.7, 22.5 and 13.2 Å away from the centre of the heme group respectively. Mutations at position 140 in the N-terminal domain are the only high confidence mutations located in the helix region in this domain. These mutations; S140N and S140R are 13.3 Å from the heme group in the active site. In the C-terminal domain, mutation S457I is located in the helix region 58 Å away from the heme while G593D is located in the loop region 40.9 Å from the heme.

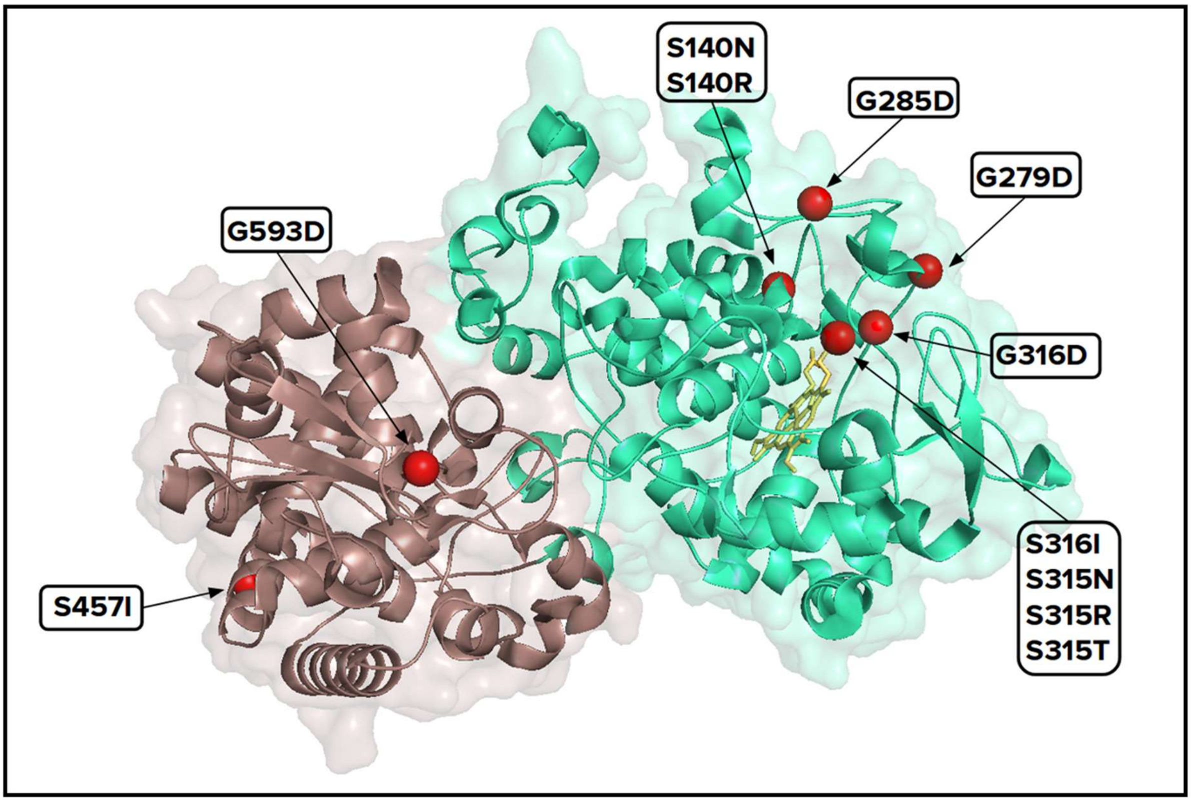


Figure 2.7: **katG structure with mapped isoniazid resistance conferring high confidence mutations.** The N-terminal domain is green, C-terminal domain is brown and the variant position shown as red spheres, the heme group (yellow) is shown as sticks.

Research has shown that mutation location plays a role in the isoniazid resistance mechanisms. In addition, different mutations at the same location have shown to have different isoniazid resistance mechanisms (Cade et al., 2010).

The modelled katG wild type protein structure in complex with isoniazid was used as the template for modelling the respective variants using MODELLER. Alignments were made between the modelled wild type structure sequence and the respective variant sequences using TCoffee (Notredame et al., 2000) and edited into MODELLER format (pir files).

For each variant, 100 models were generated and the best selected based on their z-DOPE scores (Table 2.2).

Table 2.2: **The z-DOPE scores of the best MODELLER generated models**.

Homology model	**z-DOPE score**
Wild type	-1.006
S140R	-0.816
S140N	-0.798
G279D	-0.787
G285D	-0.803
S315T	-0.814
S315R	-0.824
S315N	-0.817
S315I	-0.831
G316D	-0.804
S457I	-0.808
G593D	-0.799

All the z-DOPE scores were close to the template (2CCA) z-DOPE score of -1.006 indicating high structural similarity.

KatG variant model validation was done using ProSA, PROCHECK and VERIFY3D webservers. PROCHECK results obtained from the Ramachandran plot statistics showed that variant structures were in agreement with the protein physicochemical rules (Table 2.3).

Table 2.3: PROCHECK Ramachandran plot statistics for each variant katG model.

Structure model	-Residues in most favored regions [A,B,L]	Residues in additional allowed regions [a,b,l,p]	Residues in generously allowed regions [~a,~b,~l,~p]	Residues in disallowed regions
S140R	92.4%	7.4%	0.2%	0.0%
S140N	92.2%	7.8%	0.0%	0.0%
G279D	92.1%	7.9%	0.0%	0.0%
G285D	92.1%	7.8%	0.0%	0.2%
S315T	91.9%	7.8%	0.0%	0.0%
S315R	92.6%	7.4%	0.0%	0.0%
S315N	92.7%	6.9%	0.3%	0.0%
S315I	92.6%	7.3%	0.2%	0.0%
G316D	92.6%	7.1%	0.3%	0.0%
S457I	92.6%	7.4%	0.0%	0.0%
G593D	91.9%	7.6%	0.5%	0.0%

PROCHECK validation indicated that most of the residues in the models were in the favored regions and additionally allowed regions with no residues in disallowed region except for model G285D. This model had 0.2% of the residues in the disallowed region. The PROCHECK results indicated that there was no steric hindrance of the atoms in the model structures.

ProSA validation scored all the katG variant models a z-score below the recommended cut-off of 0.5 indicating good model quality. Mutations S140N, G279D and S315I had a z-core of -11.82 while mutations S315R and S315N had a z-score of -11.77. G316D and G593 had a ProSA z-score of -11.78 and lastly mutations S140R, G285D, S315T and S457I had z-scores of -11.79, -11.96, -11.84 and -11.68 respectively.

Finally, VERIFY3D validation passed all the generated katG variant models (Table 2.4). In VERIFY3D validation, at least 80% of the protein residues need to score ≥ 0.2 for the model to be passed. Models with more 80% or more residues with 3D-ID score ≥ 0.2 are considered as good models and those below 80% as bad.

Table 2.4: Percentage of residues with 3D-ID score ≥ 0.2 in the respective katG variants.

Variant model	Percentage residues with 3D-ID score ≥ 0.2	VERIFY3D validation
S140R	92.03%	PASS
S140N	93.43%	PASS
G279D	94.97%	PASS
G285D	95.10%	PASS
S315T	95.94%	PASS
S315R	92.87%	PASS
S315N	93.85%	PASS
S315I	93.29%	PASS
G316D	95.10%	PASS
S457I	92.87%	PASS
G593D	93.99%	PASS

2.4 CHAPTER SUMMARY

In order to study the katG protein behavior after undergoing known resistance conferring mutations, the high confidence katG variant structures were modelled using homology modelling. Homology modelling is a technique for modelling protein structures using templates with a high enough (≥ 25%) sequence identity to the target protein sequence. In this chapter, high confidence resistance conferring katG mutations were modelled using a homology modelling tool; MODELLER. The *M. tuberculosis* katG (PDB ID: 2CCA) and *S. cerevisiae* (PDB ID: 2V2E) protein structures were used as modelling templates for the wild type katG and variants.

From the models, variants at position 315 were closest to the active site (10.4 Å from the heme group) forming part of the isoniazid access channel in the N-terminal domain. Other N-terminal variants included S140R, S140N, G279D and G285D that surrounded the active site region in the N-terminal domain bringing the total number of variants in the N-terminal to nine. Two of the high confidence variants; S457I and G593D were located in the C-terminal domain at a distance of 58 Å and 40.9 Å away from the active site heme group. Most of the mutations; G279D, G285D, S315I, S315R, S315N, S315T, G316D and G593D formed part of the loop regions in the katG structure. The other three high confidence mutations; S140N, S140R and S457I formed part of the helix regions in the N and C-terminal domains.

A number of model quality evaluation programs (MQAPs) were used to access the quality of the katG wild type and the respective variant protein models to enable model disqualification based on combined results. All MQAPs used indicated that the modelled structures had the correct stereochemistry as per the Ramachandran statistics, acceptable energy as per ProSA and sequence compatible to 3D structures as per VERIFY3D. The verdict from the model validation was that the models were of good quality and could further be used to study the katG active site residue interaction with isoniazid through molecular dynamics. The next chapter (molecular dynamics) focused on establishing the structural and functional changes in the binding characteristics of isoniazid in the *M. tuberculosis* katG protein.

CHAPTER THREE

MOLECULAR DYNAMICS SIMULATIONS OF THE KAT-G PROTEIN AND VARIANTS

3.1 INTRODUCTION

Based on the fact that proteins are dynamic molecules (Özen et al., 2011), molecular dynamic calculations provide a platform to study protein movements and conformational changes of biological macromolecules as a function of time (Alonso et al., 2006). Therefore, molecular dynamics can be used to investigate protein-protein and protein-ligand interactions in a system. In this chapter, the katG wild type and variants protein conformational changes were examined over time using molecular dynamic simulations.

3.2 MOLECULAR DYNAMIC SIMULATIONS

Molecular dynamics (MD) is a study of the movement of molecules as a function of time. Molecular dynamic simulation is a technique used to produce dynamic trajectories of a system composed of a specific number of particles by applying Newton's laws of motion (González, 2011).

Molecular dynamics were first explored and introduced by Alder and Wainwright in the 1950's (Alder & Wainwright, 1959). Later in 1974, Rahman and Stillinger performed the first MD simulation which was carried out on a bovine trypsin inhibitor and it lasted a duration of 8.8 ps (Stillinger & Rahman, 1974). The principle behind molecular dynamics is solving Newton's classical equation of motion (Equation 3.1) by using forces acting between atoms in an initial configuration to find the next configuration. To solve Newton's equation of motion, initial conditions like the positions of the atoms in a structure and their velocities need to be established first. The atom positions are obtained from the coordinate information in the structure PDB file while their velocities are calculated using the Maxwellian distribution centered on the desired temperature.

$$\mathbf{F = ma}$$

Equation 3.1: **Newton's classical equation of motion used in MD simulations**. **F** is the force acting on the atom, **m** is the mass of the atom, and **a** is the acceleration of the atom.

Forces acting between atoms in a structure are dependent on the bond length, bond angle and the torsional angle between the atoms. For bonded interactions, Hooke's law (**F= -*kx***, where **F** is the force applied, **-k** is the spring constant and **x** is the extension) is used to calculate the forces acting on the bonded atoms while for non-bonded interactions, Lennard-Jones functions are applied. The collection of force equations (bond-angle, bond-length, torsional-angle and non-bonded interactions) and their respective constants is called force field equations and these are used to reproduce molecular geometry and selected properties of tested structures.

With the use of force field parameters, MD simulations can be applied in molecular docking and drug design, understanding allosteric effect, refining structure predictions and studying protein-protein and protein-ligand interactions (Gelpi et al., 2015). Currently three approaches can be used when performing MD simulations and these include; molecular mechanics (MM), quantum mechanics (QM) and hybridization of MM and QM.

When using MM in molecular dynamics, the atom nuclei are treated as charged spheres and the bonds between atoms as springs (Hooke's law). The advantage to MM is that it can be applied on large systems however, it is not as accurate as QM and it has limited predictive power. Unlike MM, QM describes molecule interactions through calculating the electronic forces acting between nuclei and electrons and hence it is capable of calculating properties of atoms, molecules, crystalline solids and even disordered solids. This makes QM a more accurate technique in predicting molecule interaction however, it is not only computationally expensive but also time consuming creating the need for hybridization.

When using a hybrid of QM and MM, the active site in the structure is described using QM and the rest of the structure described using MM. In this way, the system is described in a more accurate manner while requiring less computational resources.

The most commonly used force fields in MD simulations include AMBER (Case et al., 2005), CHARMM (Brooks et al., 1983), NAMD (Phillips et al., 2005) and GROMACS (Pronk et al., 2013). In this study, the GROMACS force fields were used for molecular dynamic simulations.

3.3 GROMACS

GROningen MAchine for Chemical Simulations (GROMACS) is a software package used to perform molecular dynamic simulations for systems like proteins and lipids. The software package uses a set of commands in the command-line interface and takes files as input and output (Abraham et al., 2015). The free software comprises of a large pool of flexible tools for trajectory analysis that require no scripting from the user. The program also allows the user to monitor the progress of the simulations, and the output of the trajectory analysis is represented in form of xmgrace graphs (Abraham et al., 2015).

With the growth in computing power and development of simulation algorithms like GROMACS, the study and prediction of interactions between receptors and ligands (Chong et al., 1999), prediction of receptor functions and description of transitional states of complex interactions (Huang & Caflisch, 2011) has not only been made possible but also eased. The one challenge with the available force fields is their lack of parameters for non-core molecules like cofactors and inhibitors. However, a number of webservers i.e. ATB (Malde et al., 2011), ProDRG (van Aalten et al., 1996) and ACPYPE (Sousa da Silva & Vranken, 2012) are available to calculate parameters for these molecules.

3.4 AUTOMATED TOPOLOGY BUILDER (ATB)

As much as the different force fields i.e. AMBER, CHARMM, GROMACS etc. have unique parameters, these parameters are for a given set of core molecules i.e. lipids, amino acids, nucleic acids, nucleotides and common sugars. These parameters are configured to reproduce a given set of properties that match across all the force fields. As a result, other non-core molecules such as cofactors, inhibitors and potential drug molecules have to be parameterized individually using other tools. Automated topology builder is a web server that uses quantum mechanical calculations and a knowledge-based approach to generate topologies and molecular dynamic parameters compatible with GROMACS force fields for different molecules (Malde et al., 2011). The ATB is capable of generating MD parameters for hetero-molecules, amino acid, nucleic acid and solvents. Besides MD parameter generation, the ATB also acts as an archive for already parameterized molecules as part of the GROMACS family of force fields. The ATB web server requires a molecule pdb file, the protonation and tautomeric states of the molecule as input and the output are the itp and pdb molecule files.

3.5. MODE-TASK SUITE

Besides the conventional MD trajectory analysis techniques i.e. RMSD, RMSF and Rg, MD data can be further analysed to investigate the global motions and identification of essential motions of macromolecules. The MODE-TASK (Ross et al., 2018) is a software suite comprising of a number of tools used to analyse and compare protein dynamics using MD trajectory data (Ross et al., 2018). The suite is composed of normal mode analysis (NMA) and principal component analysis (PCA) tools used in identifying most prominent conformations in a protein MD trajectory. In addition to conformation analysis tools, MODE-TASK also includes mode visualization tools and a MODE-TASK PyMOL plugin (Ross et al., 2018).

3.5.1 Principal component analysis (PCA)

Principal component analysis is a technique that can be used to identify the most biologically relevant protein motions. The technique analyses a data table in which observations are described by several inter-correlated quantitative dependent variables (Ross et al., 2018). The technique has vast applications including detection of correlated motion in molecular dynamics data. In this study, PCA was used to illustrate the 3D conformational sampling and internal dynamics of the wild type and variant katG proteins using scripts in the MODE-TASK suit. Principal component analysis takes an MD trajectory in terms of a small number of variables, sometimes referred to as essential degrees of freedom (EOF) and extracts the most important nodes in the movement of the molecule and this is performed on the Cartesian coordinates of the molecule (Haider et al., 2008). The mean squared positional fluctuations (variances) of each atom are readily calculated from the total simulation. If the obtained variances do not change significantly over time that would imply that within the observed period, only motion about one native conformation was recorded. Principle component analysis uses a covariance matrix-(C) of the protein MD simulation that is obtained by first superimposing the protein coordinates on a reference structure, which is usually the initial coordinates, or the average coordinates and then a displacement vector for each residue (described by the Cα or Cβ coordinates of the residue ***i***) at a time point ***t*** is obtained (Ross et al., 2018).

3.6 METHODOLOGY

Molecular dynamic trajectories for the katG wild type and variants were calculated using a step-by-step MD standard operating procedure as illustrated in Figure 3.1.

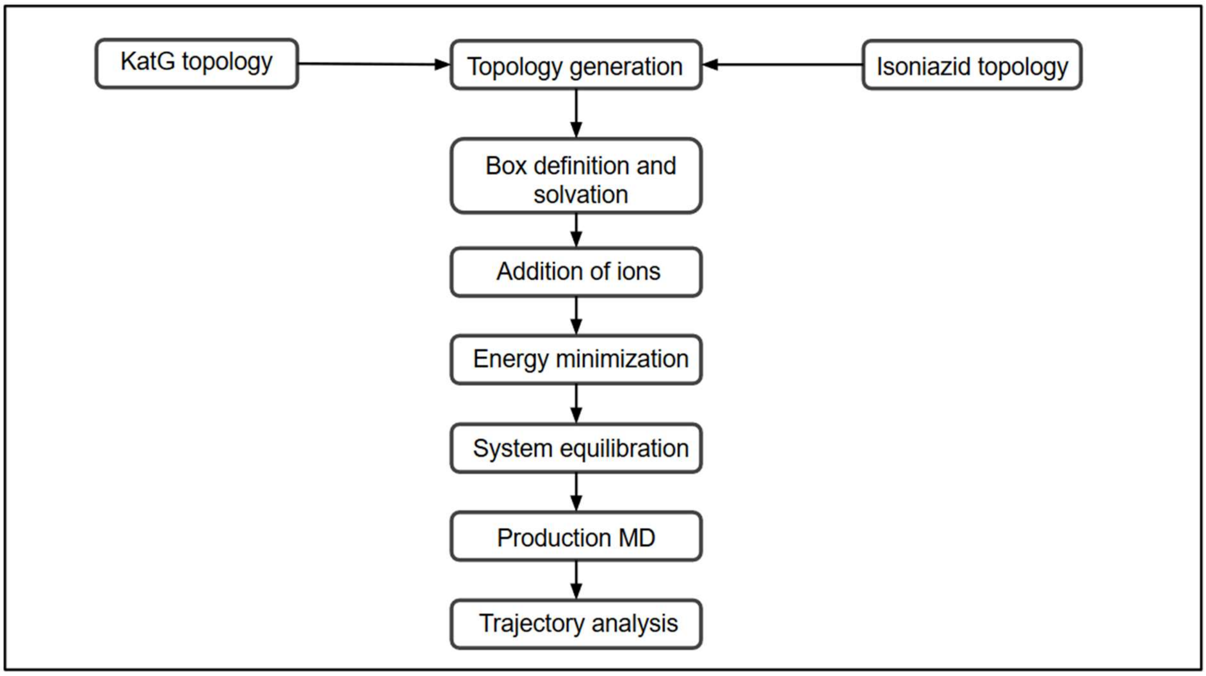


Figure 3.1: **Schematic diagram of the methodology followed in MD calculations**.

3.6.1 Topology generation

3.6.1.1 Protein topology

The protein topology file and processed structure files were generated using the **pbd2gmx** command in the GROMACS 5.1.5 package while using the GROMOS96 54a7 force fields. In addition to having parameters for core molecules, the GROMOS96 54a7 force field parameters also contain parameters for the heme group which is present in the katG active site. The input file for the **pbd2gmx** command was the modelled holo katG coordinate file (containing only the protein and heme atom coordinates) and the outputs were a topology file (.top), a position restraint file (.itp) and a post processed structure file (.gro). The topology file contained parameters for the bonded and non-bonded atom interactions while the processed structure file (.gro) contained the structure atoms coordinate information as defined by the GROMOS96 54a7 force fields and finally the .itp file contained the restraints for the heavy atoms.

3.6.1.2 Ligand topology

The ligand (isoniazid) PDB file was submitted to the ATB topology builder tool to generate a topology file using default settings. The outputs were the ligand .itp file containing the atom connection information and the coordinate (.pdb) file. The ligand pdb file was converted to .gro format using the **editconf** GROMACS command (gmx editconf –f file.pdb -o file.gro).

3.6.2 Box definition and solvation

The protein and ligand coordinate information was combined to create a single structure file containing the protein-ligand complex. This was done by copying and pasting the ligand.gro coordinate information to the protein .gro coordinate file and ensuring that the atom number in the complex file was correct. Molecular dynamic simulations were run under periodic boundary conditions (PBC) using a cubic box to mimic an infinite system. The **editconf** GROMACS command was used to place the protein-ligand complex in the center of the cubic box (cube side length: 11.1 nm) with a clearance space of 1 nm to satisfy the minimum image convention. The system was solvated using SPC water model (spc216) using the **solvate** GROMACS command. Solvation of the system was to mimic the normal protein environment and ensure easy interactions between the protein and the ligand.

3.6.3 Addition of ions

Molecular dynamic simulations require a system of net charge of zero (neutral system) in order to mimic protein behavior *invitro*. The neutral system was achieved by adding 23 NA ions to the system to equilibrate the original structure's net charge of –23. This was done by running the **grompp** command to generate the .tpr file and determine the net charge of the system. The .tpr file was then used as input file for the **genion** tool in GROMACS that replaces a specified number of water molecules in the system with the user specified ions (CL or NA).

3.6.4 Energy minimization

Energy minimization prior to MD simulations is important to ensure that the system has no steric clashes or inappropriate geometry resulting from the addition of water molecules and ions from the prior steps. The **grompp** GROMACs command was used to generate a binary input file (.tpr) which was used as an input file for the GROMACS MD engine **mdrun** minimization command. Minimization was done using the steepest descent minimization integrator for 50000 steps and minimization was stopped when the maximum force of < 1000.0 kJ/mol was achieved. Verification of minimization was done by analyzing the GROMACS energy terms file (em.edr) using the **energy** GROMACS command to generate the potential.xvg file. The potential.xvg shows the energy minimization curve and this was viewed using **xmgrace** plots.

3.6.5 Equilibration

After minimization, system equilibration was required to optimize the solvent together with the solute in the system. Equilibration was done by bringing the solvent to the desired simulation temperature of 300 K under the NVT ensemble (constant Number of particles, Volume, and Temperature) for 100 ps. The desired system density was achieved through pressure equilibration under NPT ensemble (i.e. constant number of particles (N), constant pressure (P) and constant temperature (T)) for 100 ps to stabilize the pressure of the system at around 1.0 bar. Temperature and pressure equilibrations were verified using the GROMACS **energy** command which created temperature and pressure equilibration graphs that were viewed using **xmgrace** plots.

3.6.6 Production MD

Consequent to completion of the temperature and pressure equilibration, the position restraints were released and production MD run using GROMACS commands. Molecular dynamic simulations were run for 200 ns and trajectories dumped every 10 ps. The next step was visualization and trajectory analysis.

3.6.7 Analysis

Upon completion of the production MD runs, the structure was removed from the PBC simulation box and centered using the **trjconv** GROMACS command. The trajectory of the

MD run was then analyzed using, RMSD, RMSF and Rg. The dynamics of the structures over the simulation time were viewed using visual molecular dynamics (VMD) tool (Humphrey et al., 1996).

3.6.7.1 RMSD

The root mean square deviation is the measure of the average distance between the atoms (usually backbone atoms) of superimposed proteins. The RMSD is used to measure the stability of the simulated structure over the simulation time. The RMSD for the protein structures was calculated using the **rms** GROMACS command and visualized using xmgrace plots.

3.6.7.2 Principal component analysis

Principal component analysis was calculated from the MD trajectory of the katG wild type and variants. The input files for PCA were the trajectory files (xtc) and the topology files (tpr). The trajectory files from the MD simulation were prepared by first removing the periodic boundary conditions (PBC), removing the water molecules and fitting atoms to the reference structure using the **gmx trjconv** command. Then, GROMACS commands; **gmx covar** and **gmx anaeig** were used to create the covariance matrix and diagonalize the matrix respectively by selecting the backbone atoms option in both commands. An R script (Appendix 1) written by Arnold Amusengere (RUBi PhD student) was edited to fit the study requirements and used to plot the 3D heat maps showing the most prominent structure conformations.

3.6.7.3 RMSF

The root mean square fluctuation is the standard deviation of the atom position calculated from the average structure. The RMSF is used to measure local chain flexibility and structure stability. The RMSF was calculated using the **rmsf** GROMACS command and visualized using the xmgrace plots.

3.5.7.4 Radius of gyration

This is a measure of the distribution of atoms of a protein around its axis. The Rg was used to describe the overall spread and compactness of the molecule (structure). A stably folded molecule will have tight Rg whereas a molecule that unfolds will have a fluctuating Rg. The GROMACS command **gyrate** was used to calculate the Rg and the gyration graphs were visualized using xmgrace pots.

3.7 RESULTS AND DISCUSSION

The MD simulations were performed to study the katG protein conformational and behavioral changes as a result of the high confidence katG mutations known to confer isoniazid drug resistance. To do this, MD simulations were run on the modelled katG-isoniazid complex, the modelled holo katG and the respective variants as discussed in the following sections.

3.7.1 Modelled katG-isoniazid complex

Analysis of the 200 ns MD simulations showed an unstable isoniazid interaction in the katG active site of both the model and the template (PDB ID: 2v2e) protein structures (Figure 3.2). Isoniazid was observed to maintain interaction within the active site for approximately 22 ns and then changed conformation i.e. isoniazid pyridine ring was interacting with the active site as opposed to the hydrazine group as indicated in literature (Henriksen et al., 1998, Aitken et al., 2001 and Bertrand et al., 2004). The isoniazid conformation change in the active site was attributed to the observed reduction in hydrogen bonds and strength of interactions between the isoniazid and the active site residues as the MD simulations progressed, a scenario also observed by Unissa et al., (2018) during 10 ns katG simulations.

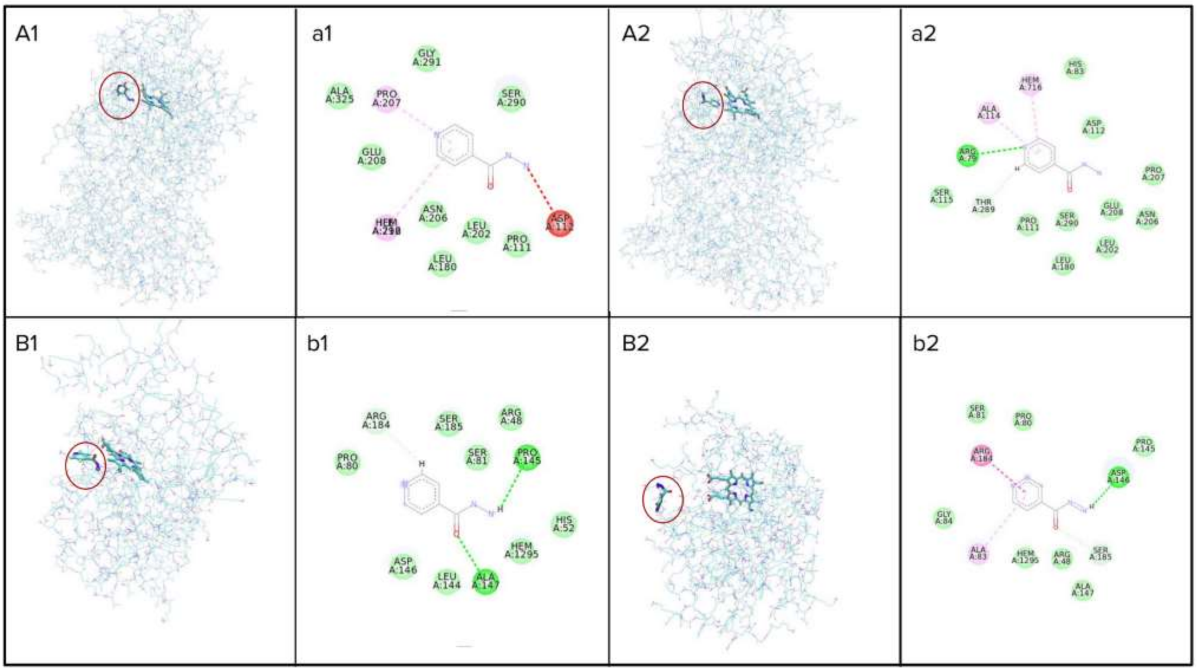


Figure 3.2: **Isoniazid conformational change in the model and template katG active site.** **A1**: Shows model katG-isoniazid conformation before the 22 ns mark, **a1**: shows the model katG-isoniazid's interacting residues before 22 ns mark, **A2**: shows model katG-isoniazid conformation after 22 ns, **a2**: illustrates model katG-isoniazid‘s interacting residues after 22 ns, **B1**: illustrates template katG conformation before 22 ns, **b1**: shows template katG-isoniazid's interacting residues before 22 ns, **B2**: illustrates template katG-isoniazid conformation after 22 ns, and finally, **b2**: shows template isoniazid's interacting residues after 22 ns.

Early in the simulation (before 22 ns) the isoniazid was observed to have two pi-alkyl and one unfavorable bond interaction in the model whereas the template had two hydrogen bonds and one van der Waals interaction in the katG active site. Further into the simulation (after 22 ns) isoniazid was observed to have two pi-alkyl and one hydrogen bond interactions between its pyridine ring and the active site in the model katG. In the template, isoniazid had one hydrogen bond, two pi-alkyl bonds and one van der Waals interaction which are weaker than the hydrogen bond interactions observed early in the MD simulation. Due to the observed weak interactions and in the interest of studying katG variant conformational changes over a significant time (200 ns), katG holo structures were used for further investigation (MD-simulations).

3.7.2 Isoniazid unbound katG and modelled variants

Molecular dynamic simulations were performed on the holo structures of the katG wild type and the respective variants to study the conformational dynamics of the structures at atomic resolution. Post MD analysis was performed using RMSD, RMSF and Rg following the steps as explained in section 3.5.

3.7.2.1 RMSD

The root mean square deviation was used to measure the conformational diversity between the two sets of values in the context of the holo wild type katG protein structure and respective variant holo structures as shown in Figure 3.3.

Analysis of the katG wild type conformational flexibility over a 200 ns trajectory of showed RMSD fluctuations between 0.2 nm and 0.6 nm (Figure 3.3) until the 150 ns mark where the protein stabilizes around 0.45 nm. The 150 ns mark was taken as the protein equilibration point. Variants S315I, S315N, S140R, G593D, G285D, G316D and G279D showed an increased RMSD compared to the wild type. Six of these variants; S315I, S315N, S140R, G285D, G316D and G279D are located in the N-terminal domain around the active site with an average distance of 15.82 Å from the heme group in the active site. Variant G593D is in the C-terminal domain at a distance of 40.9 Å from the active site. Reduced RMSD was observed in variants S315R and in the early stages (up to 100ns) of variant S140R. Variant S315T located in the N-terminal and S457I in the C-terminal showed somewhat similar RMSD as that observed in the wild type (Figure 3.3).

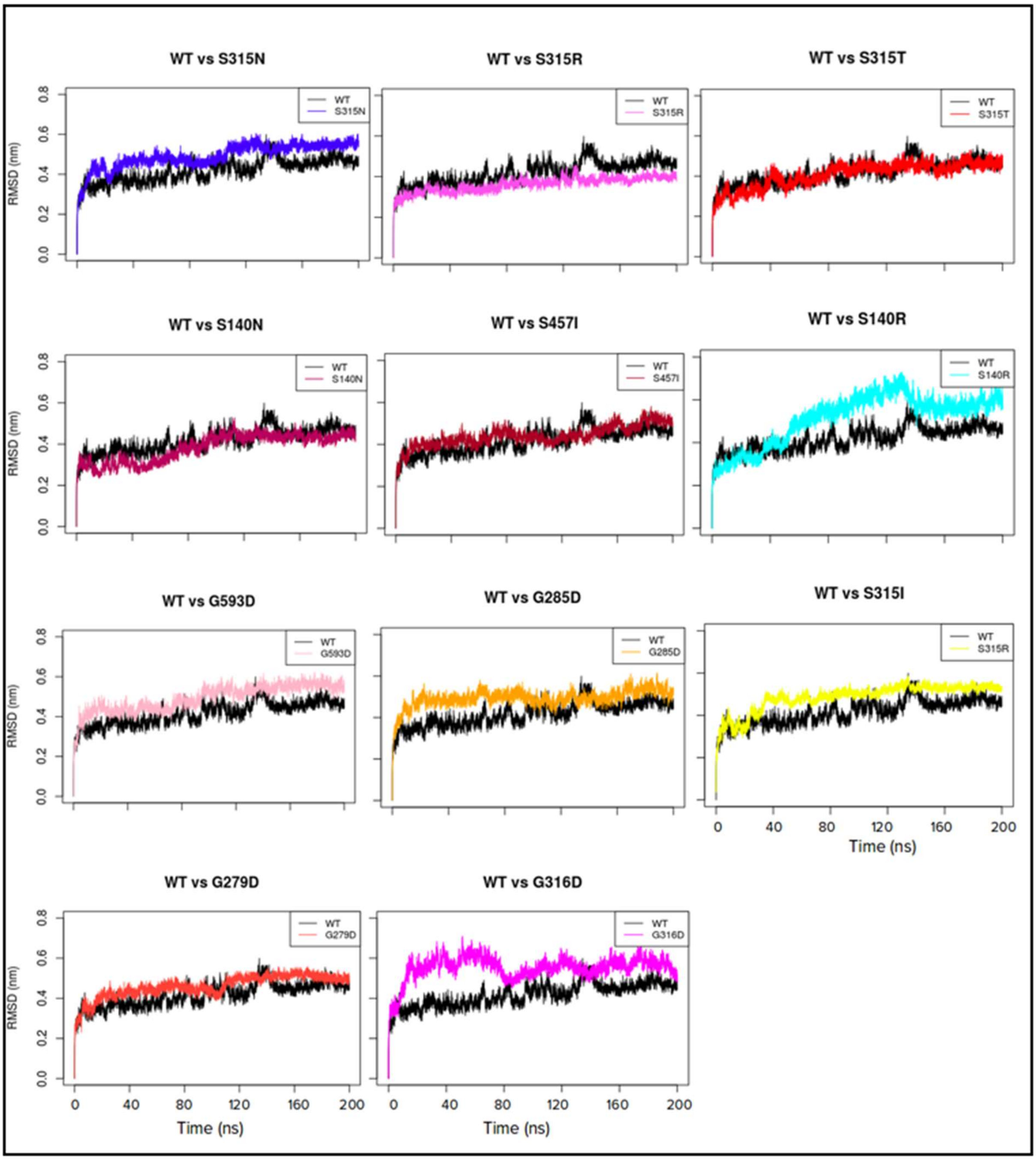


Figure 3.3: RMSD graphs of katG wild type and variant proteins. The graphs show a comparison of the katG wild type (black) and variants' (respective color) structural stability over the 200 ns simulation. The y-axis represents RMSD in nm of the structure throughout the simulation and x-axis represents time in ns.

Pertaining to structural stability, variants S140R and G316D in the N-terminal domain had an increased structural instability compared to the wild type and the rest of the variants. Both S140R and G316D had a fluctuation ranging from 0.2 nm to 0.7 nm suggesting more backbone residue flexibility in the variants. The rest of the variants especially S315R and S315I that had an RMSD range of 0.2 - 0.35 nm and 0.2 - 0.5 nm respectively showed more structural rigidity compared to the wild type.

In as far as the RMSD distribution is concerned; the wild type displayed a bimodal RMSD distribution (Figure 3.4) as was the case for most of the variants indicating two prominent structural conformations of the katG structure. The mutations caused varying backbone conformational variability of the katG structure with variants S315R, S315T and S140N having less variability (RMSD means of 0.36, 0.41 and 0.38 respectively) compared to the wild type with an RMSD mean of 0.42. On the other hand, variants S315N, S315I, S140R, S457I, G279D, G316D, G285D, and G593D showed more conformational diversity (RMSD mean of 0.49, 0.49, 0.51, 0.44, 0.46, 0.55, 0.49 and 0.49 respectively) compared to the wild type structure. Variants S140R and G316D had the most conformational space compared to all the variants and the wild type which was also observed earlier in the structural flexibility graphs (Figure 3.3). The variant S140R is located on the distal end of the heme group and G316D is on the proximal end (Figure 2.7). The observed wide histogram bases signify more sampled conformations in the MD simulation for the respective variants (S140R and G316D).

From the data, the high confidence mutations seem to be affecting not only the stability of the katG protein structure but also, there is a significant conformational change between the katG wild type structure and that of the variants. Further analysis of the backbone flexibility and conformational changes was done using PCA.

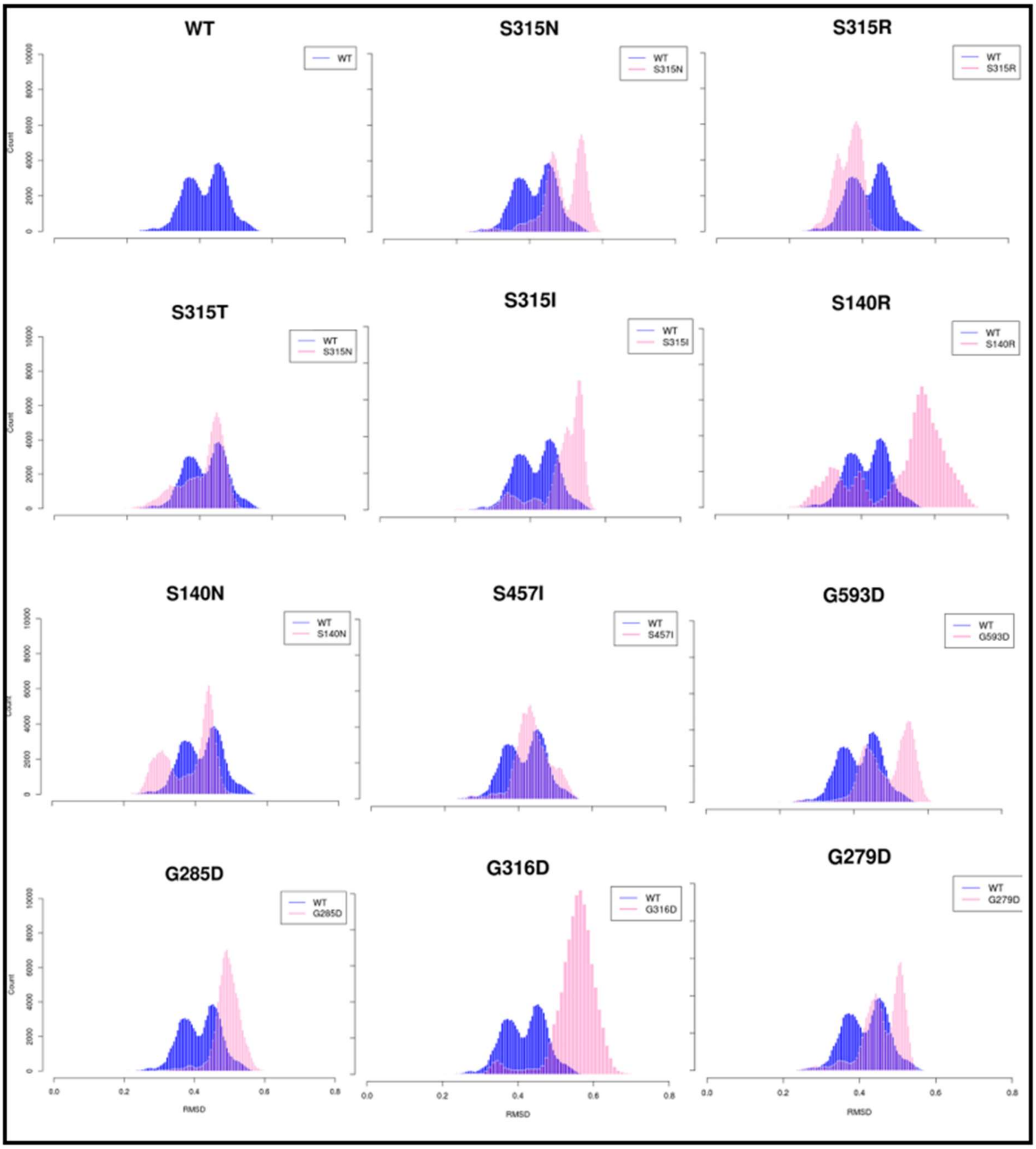


Figure 3.4: Histogram presentation of the differences in the Cα atom RMSD between the wild type holo katG and variants. The y-axis represents the frequency of the sampled conformations and x-axis represents the RMSD of the sampled conformations.

3.7.2.2 Principal component analysis

Principal component analysis was used to further explain the observed RMSD results and to examine the conformational changes and energy distribution over the course of the simulation. Principal component analysis identified both clockwise and anti-clockwise protein conformational movements over the 200 ns trajectory with variants: S315R, S315N, S315I, S140R, S140N, G593D, G285D and G279D having a clockwise conformational shift while G316D, S315T and the wild type exhibiting a somewhat similar anti-clockwise shift (Appendix 2). This means that the mentioned variants caused a change in the protein's behavior leading to a directional shift in protein's conformation sampling from anti-clockwise to clockwise. Conformational change patterns (i.e. clockwise and anti-clockwise) are of paramount importance in as far as protein interactions are concerned as they are essential for activities like ligand binding, bimolecular recognition, enzymatic catalysis and determining protein functionality (Kurplus & McCammon, 1983). The difference in conformational shift between the variants and the wild type could be impacting the reaction of isoniazid in the katG protein active site of the respective variants leading to resistance.

Further analysis of the PCA data using energy heat maps illustrated difference between the wild type and variant conformational space and energy over the trajectory (Figure 3.5). From the heat maps, the wild type structure is observed to have a larger conformational variation as compared to variants S315R, S457I, S315T, S140N and G279D. The energy heat map observations are in agreement the RMSD data that showed a lower backbone flexibility in the wild type and the mentioned variants.

In the wild type RMSD data, a bimodal distribution is observed whereas the PCA data shows three prominent conformations. The third observed conformation in the wild type is believed to be the transitional conformation between the two major ones.

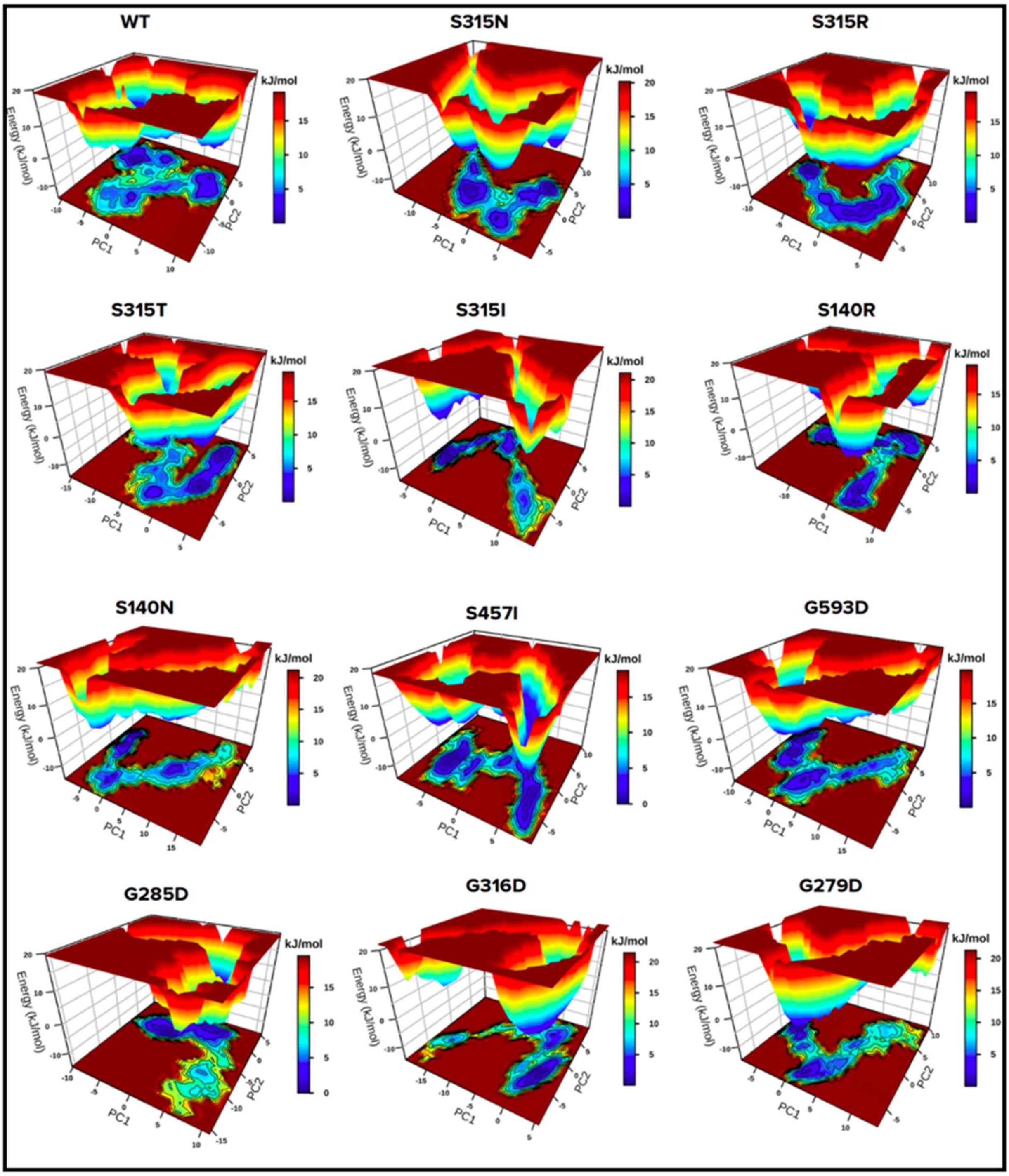


Figure 3.5: Energy heat maps of the katG wild type and variants over a 200 ns trajectory. Surfaces are colored based on the Gibbs free energy levels from maroon (high energy/ maxima) to blue (low energy/ minima). The contours represent an increase in the free energy (kJ/mol).

Variants S140R and S315I that showed more structural instability in section 3.7.2.1, are observed having a greater conformational space compared to the wild type and other variants.

This indicates that these two variants had a more divergent conformation sampling that was not observed in the wild type.

Variants S140R, S140N, G279D, G285D, S315I and G316D showed higher energy barriers between their prominent conformations as compared to S315N, S315R, S457I and the wild type. Structures with high energy barriers between conformations require more energy to move from one conformation to the next. Since in nature low energy interactions are favored, structures requiring more energy would have limited conformational variation. As observed in section 3.7.2.1, the mutations have a varying effect in terms of conformational space and variability on the katG structure.

3.7.2.3 RMSF

The protein residue flexibility of the katG wild type and the respective variants was compared using RMSF KatG residues at positions 26 - 110 (Figure 3.6) showed reduced residue flexibility across all variants as compared to the wild type. This region includes active site residues Arg-104, His-108 and Trp-107 that coordinate isoniazid forming hydrogen bonds.

Variant S315T was observed to have the least RMSF around the katG active site (26 - 110) compared to all the other studied variants. Reduction of the katG region 26 - 110 residue flexibility especially the active site residues could be impacting the reaction of isoniazid in the active site. In addition to the significantly decreased residue flexibility, mutation of serine to threonine in S315T variant has been shown to reduce the size of the active site access channel causing steric hindrance as a result of the additional methyl group in threonine, hence preventing isoniazid from accessing the active site (Marney et al., 2017).

Studies also show that variant S315T has a high catalytic activity as compared to other variants hence increasing its virulence since catalase activity is important in protecting the bacteria from toxic oxygen radicals (Suarez et al., 2009).

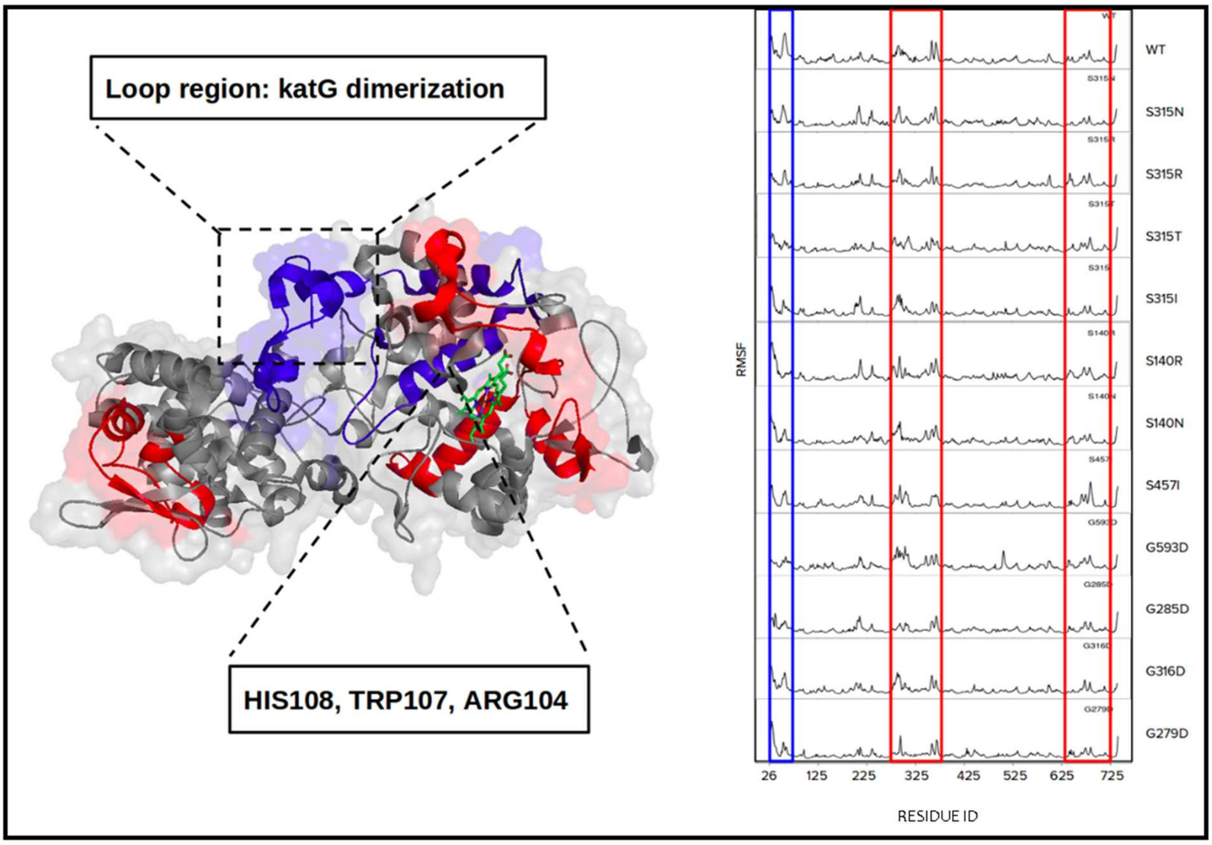


Figure 3.6: RMSF in the holo katG wild type and variant structures. The regions with reduced RMSF across the variants in comparison to the wildtype are marked blue. Regions with increased residue fluctuation in both the wild type and variants are marked red (left).

The reduction in residue flexibility at positions 26-110 across all the variants, even those that are in the C-terminal domain (G593D and S457I) implies an allosteric effect of the mutations on the katG structure. KatG residues 250 - 300 and 330 - 350 showed increased RMSF in both the katG wild type and the respective variants. These regions are loop regions hence explaining the observed high residue flexibility.

In the C-terminal, increased RMSF was observed in both the wild type and variant katG at positions 630 - 725 with variant S457I having the highest residue flexibility at these positions. This is region is part of a loop region which explains the observed residue flexibility. A general RMSF analysis showed reduced residue flexibility across all the variants when compared to the wild type.

3.7.2.4 Radius of gyration

The intermolecular compactness of the katG structures and the respective variants was compared using the Rg. Rg determined the protein's folding and unfolding state by measuring the mean distance of the atom collection from its center of mass.

Interestingly, Rg analysis (Figure 3.7) showed all high confidence mutations structures having a lower Rg as compared to the wild type. In as far as the protein structure is concerned, this meant that the mutations lead to contraction of the katG structure resulting in a more compact variant protein as compared to the wild type. Bearing in mind the effects of protein compactness on the katG, decreases would result in active site and access channel residues moving away from each other, whereas increases in protein compactness would result in the residues moving closer to each other.

Variant S316D showed the least change in the Rg as compared to the other variations whereas variants S315T, S315R and S140R had the highest gyration deviation compared to the wild type. The observed changes in the katG protein compactness especially for variant S315T are in agreement with a study by Unissa et al, (2018) and Marney et al., (2017) that showed narrowing of the isoniazid access channel and increased steric hindrance as a result of the katG mutations especially at position 315. The increase in structure compactness as an effect of the mutations was consistent across all the variants irrespective of their location in the protein. Despite studies showing that mutations in the katG have different mechanisms of conferring resistance to isoniazid (Cade et al., 2010), this data shows a uniform effect in at least among high confidence mutations.

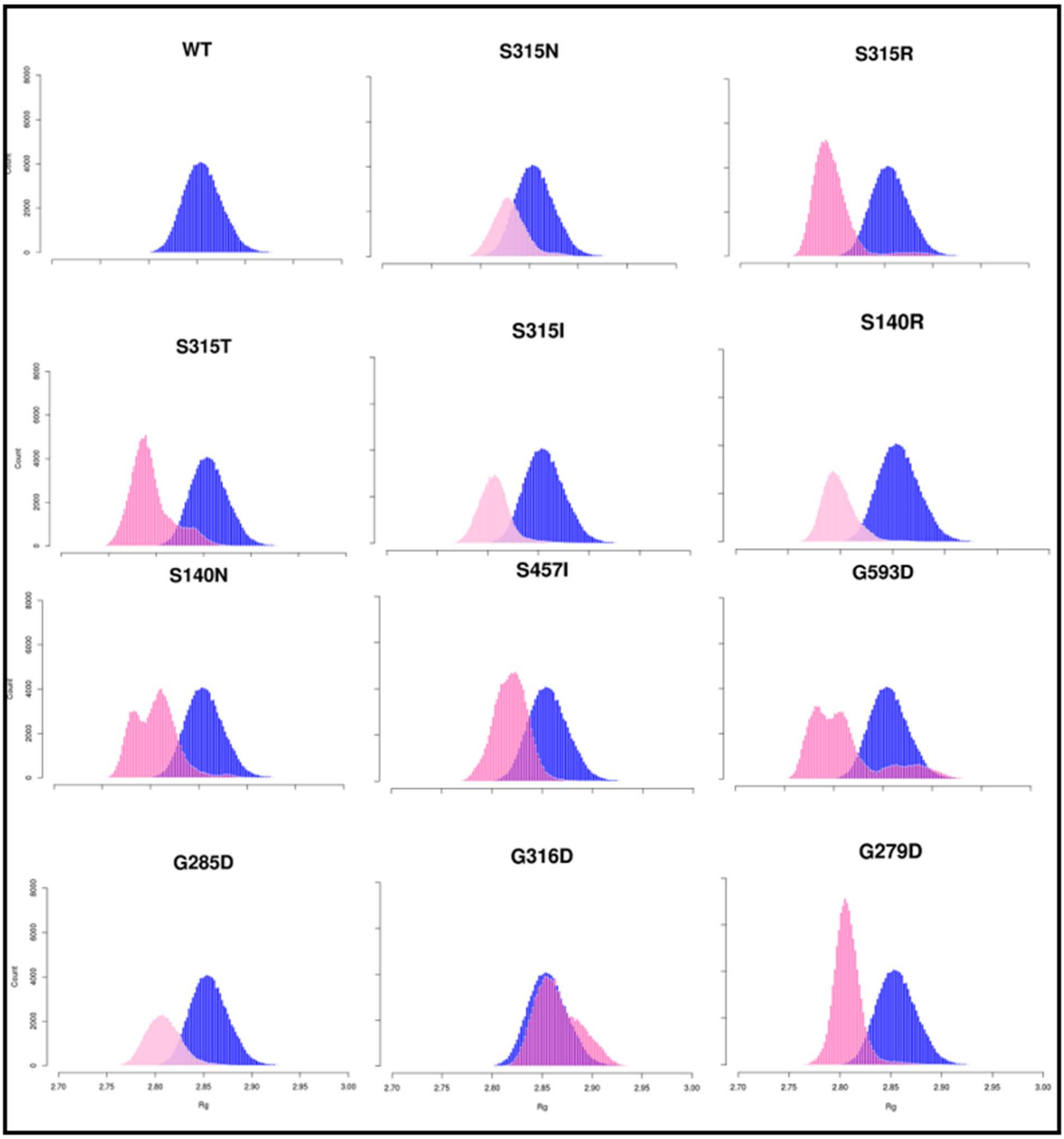


Figure 3.7: Histogram presentation of the wild type katG structure (blue) and the respective variants (deep pink) Rg. The y -axis represents the degree of protein compactness in nm and the y-axis represents the frequency.

3.8 CHAPTER SUMMARY

Molecular dynamics simulations were applied on the katG-isoniazid complex, the holo wild type katG and the holo variants to katG structures to study the protein conformational changes over a 200 ns trajectory. Trajectory analysis was used to identify and describe protein patterns, stability, interactions and conformational changes. In the katG-isoniazid complex, ligand instability was observed in the both the wild type and variants around the 25 ns mark into the MD simulations. This was attributed to less hydrogen bonds formed between isoniazid and active site residues further into the simulation (after 22 ns) in the models compared to the template. The change in orientation of isoniazid in the active site resulted in a reduction in the hydrogen bonds between isoniazid and the katG active site.

Molecular dynamic simulations were further performed on the holo katG wild type and variant protein structures for 200 ns. Trajectory analysis of the holo simulations indicated a varying backbone flexibility of the katG structure across the different variants with majority of the N-terminal variants; S315I, S315N, S140R, G285D, G316D and G279D having an increased RMSD while S315R showing a reduced RMSD compared to the wild type. In as far as structure stability is concerned, variants S140R and G316D in the N-terminal domain showed increased structural instability compared to the wild type and the rest of the variants. S140R is located in the distal side of the heme group while G316D in the proximal end.

Results from conformation analysis using PCA supported the RMSD findings that showed variants S315R, S457I, S315T, S140N and G279D having less conformational variability compared to the wild type katG. In addition, variants S140R, S140N, G279D, G285D, S315I and G316D showed higher energy barriers between their prominent conformations implying more energy requirement between conformational shifts.

Residue flexibility (RMSF) analysis showed reduced residue flexibility from position 26 to 110 across all the variants. This region contains the *M. tuberculosis* katG unique loop hook region (position 26 - 30) believed to be involved in not only mediating interdomain interactions but also responsible for the katG structure dimer assembly (Bertrand et al., 2004). Region 22 - 110 also encompasses isoniazid coordinating residues His-108, Arg-104 and Trp-107 in the katG active site.

Protein compactness calculations (Rg) showed increased structure compactness across all the variants as compared to the katG wild type. The implication of this is a more folded katG structure in the variants than in the wild type. This data therefore suggests that the studied mutations in the katG protein could be conferring resistance not only through changing the protein stability but also affecting the active site residue flexibility, structural stericity and increasing the compactness of the of the protein structure. The data also identified uniform effect of all the mutations in the katG (increase in compactness) suggesting to an extent, a common mechanism of conferring resistance across high confidence mutations.

In the next chapter, the protein residue network communication of the wild type and variants was analyzed through dynamic residue network analysis (DRN) to establish the mutation's effect on protein residue communication.

CHAPTER FOUR

DYNAMIC RESIDUE NETWORK ANALYSIS OF THE KAT-G AND VARIANTS

4.1 INTRODUCTION

In a system composed of macromolecules, biological function is based on macromolecule interactions and dynamics (Özen et al., 2011). These interactions and dynamics have over the years been studied using MD simulations and trajectory analysis techniques like RMSD, RMSF and the Rg. Much as the mentioned techniques are informative, however more insight into residue interactions in MD simulations can be obtained through studying the interactions and behavior of particular residues to establish their importance to the protein behavior as a whole. Through dynamic residue network (DRN) analysis, particular residue interactions and importance throughout the MD simulation can be analyzed.

4.2 RESIDUE INTERACTION NETWORK (RIN)

In RIN, a single residue is considered as a node and the physicochemical interactions, like covalent and non-covalent bonds are represented as edges. Residues are considered to be interacting if they are within a user defined distance of each other which is usually 6.5 - 7.5 Å (Atilgan et al., 2004). A cut-off distance of 6.7 Å is usually used since literature has shown that it includes all the residue neighbors around in the first coordination shell of a central residue (Atilgan et al., 2004). RINs are graphs that are analyzed using mathematical methods like average shortest path (*L*) and *betweenness centrality* (*BC*). Originally, the RIN analysis approach involved analyzing a structure independently and not as a trajectory from MDs however, a new approach of analyzing the RINs over the course of MD trajectory has been developed and this is called dynamic residue network (DRN) analysis (Brown et al., 2017).

4.3 DYNAMIC RESIDUE NETWORK

Like in RIN, in DRN analysis, the C_β atom of each residue (C_α for glycine) in a protein is treated as a node and the interactions between the residues (within a specified cutoff) as edges. In DRN however, residue interaction analysis is done throughout the course of the MD trajectory. In this study, DRN analysis was done using the average shortest path *L* and *betweenness centrality* mathematical approaches under the MD-TASK tool (Brown et al., 2017).

4.3.1 Average shortest path / *Reachability*

Average shortest path or *reachability* is the mean of the shortest paths to a node *i*. The shortest path between two nodes *i* and *j* is the least number of edges between the two nodes. The *reachability* describes the accessibility of the residue in a protein structure and literature (Ozbaykal et al., 2015) suggests that residues with high *reachability* influence protein conformational changes.

4.3.2 *Betweenness centrality*

Betweenness centrality of a node is the number of short paths between all other nodes that pass through the node of interest. *BC* is therefore a measure of how central a residue is in the protein communication network. Ozbaykal, (2015) suggests that residues with high *BC* values play a vital role in the inter-protein and intra-protein domain communication. *BC* is related to *L* as it shows how often a residue of interest is involved in the shortest path and studies have shown an inverse relationship between *BC* and *L* (Penkler et al., 2018).

4.3.3 MD-TASK tool

The MD-TASK is a tool suite capable of analyzing MD trajectories using network analysis techniques, perturbation response scanning (PRS) and dynamic cross-correlation (DCC). The tool was developed using Python for Linux/Unix-based systems and utilizes Python scripts to analyze MD trajectories (Brown et al., 2017). The tool can also use the igraph package of R to plot the residue contact maps and supports various non-standard Python libraries like NumPy, SciPy, Matplotlib, MDTraj and NetworkX.

4.4 METHODOLOGY

4.4.1 *Reachability* and *betweenness centrality*

The *reachability* and *betweenness centrality* of the residues in the wild type katG and variants were determined using the MD-TASK tool. MD-TASK Python script **calc_network.py** was used to calculate and plot the *L* and *BC* of the holo katG wild type and variant protein structures using a node separation distance threshold of 6.7Å and a step of 100. A step in DRN is an integer that defines the trajectory frame to be used in the network calculation. In this case, the network was calculated for every 100th frame in the trajectory.

Furthermore, MD-TASK was used to calculate the average and standard deviation in *L* and *BC* using the MD-TASK **avg_network.py** Python script ("MD-TASK documentation — MD-TASK 1.0.1 documentation", 2019). Reduced topology and trajectory files containing only the C_β atoms of a residue (C_α for glycine) were used for DRN analysis. The files were reduced using the VMD (Humphrey et al., 1996) script in the MD-TASK documentation.

The **avg_network.py** Python script generated the average and standard deviation files for the *L* and *BC*. Using an excel sheet, the change in average *L* (ΔL) and *BC* (ΔBC) was calculated by subtracting the average *L* and average *BC* of the respective variants from that of the wild type (wild type minus variant). A 50 ns trajectory (150 ns – 200 ns) was used to calculate the dynamic residue network as opposed to the whole trajectory because the katG wild type simulation was observed to equilibrate and stabilize from the 150 ns mark onwards (Figure 3.3).

The mean and standard deviation of the ΔL and ΔBC between the wild type and the respective variants were also calculated. For each variant, residues with a ΔL and $\Delta BC \geq$ 2SD and $\leq$ -2SD from the mean were considered to have a significant enough change. A standard deviation of 2SD from the mean was used so as to get only the residues with a significant Δ*L* and Δ*BC*.

4.5 RESULTS

4.5.1 *Reachability*

Dynamic residue network was employed to analyse the katG residue topological spread in relation to other residues and to analyse the protein communication over a trajectory of the last 50 ns. The difference between the average *L* of the wild type and variants was obtained for all the variants. In order to obtain the statistically significant values, a value of 2SD away from the mean was used to identify the residues from variant proteins, which characterizes significant residue connectivity. Those residues that displayed Δ*L* that is ≥ 2SD and ≤ -2SD from the mean were considered to have significant change in the protein residue connectivity (Appendix 3). These residues were further mapped on the katG structure to identify their distribution pattern in the various mutants (Figure 4.1).

A negative Δ*L* (≤ -2SD) difference between the wild type and variant signified an increased inter-residue distance in the protein network of the variants compared to the wild type. The implication of this is less availability of the specific residues for signal transduction in the variant's residue network. A positive Δ*L* (≥ 2SD) difference signified the opposite which is more availability of particular residues of interest in the protein signal transduction network of the variant. A difference of zero meant there was no change in the mean shortest path of the particular residue to the others in the wild type and variant protein structures.

From the DRN average *L* analysis, all katG variants showed decreased inter-residue distance at katG loop region 280 - 303 (Figure 4.2). These results show a linear correlation between the *L* and RMSF results observed in section 3.7.2.3 that indicated increased residue flexibility around the same region. Prior studies (Brown et al.2017) have indicated a similar relationship between *L* and RMSF. The region 280 - 303 forms part of the surface loop region made up of residues at positions 278 to 312.

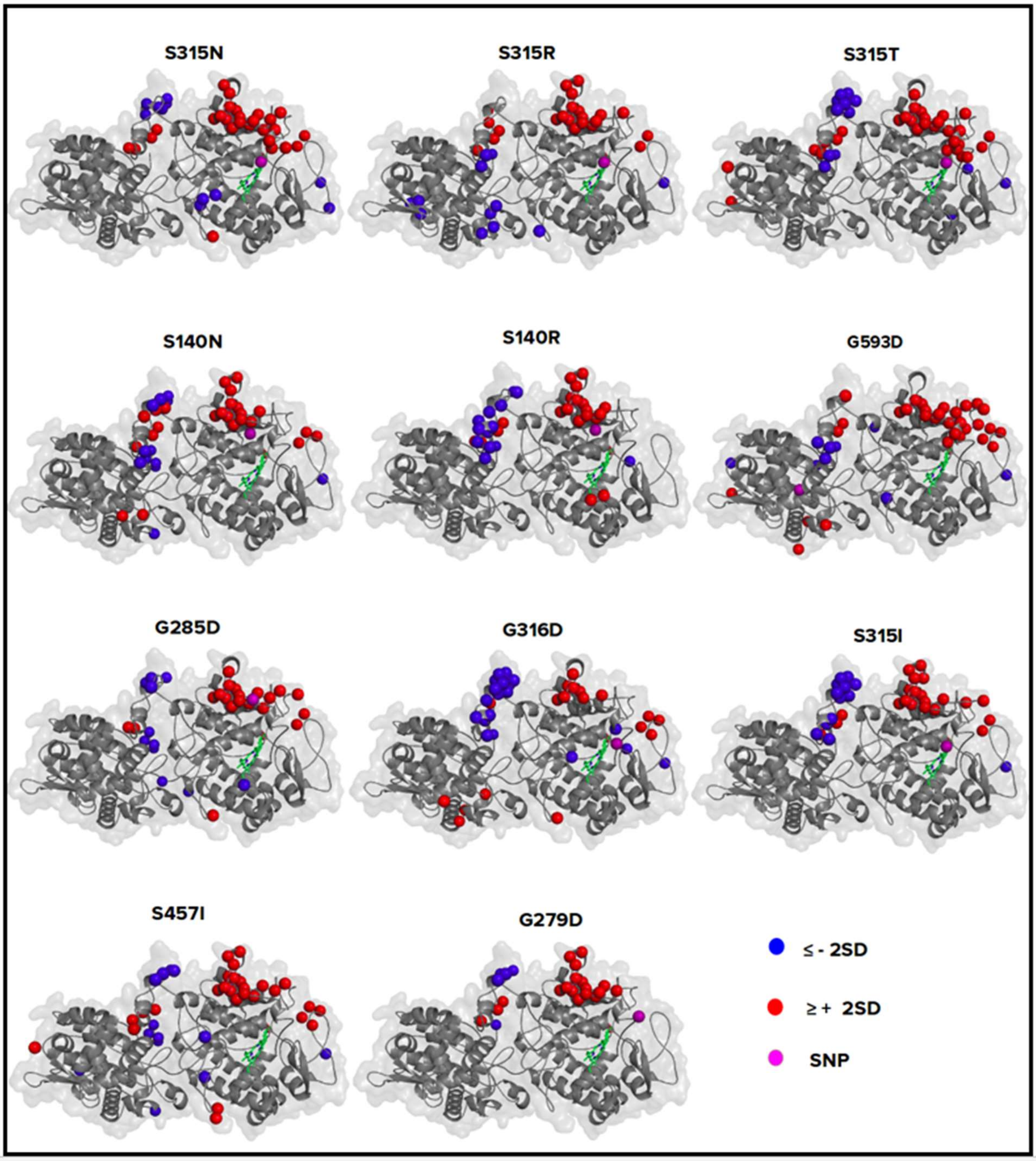


Figure 4.1: **KatG variant structures with mapped positions of residues with Δ*L* ≥ 2SD (red) and ≤ -2SD (blue).** Pink shows the mutation positions in the katG structures.

Literature suggests that the surface loop region (position 278-312) is conserved across at least three catalase-peroxidase enzymes from namely; *M. tuberculosis*, *B. pseudomallei* and *H.*

marismortui (Bertrand et al., 2004). In *B. pseudomallei* catalase peroxidase, the loop region was proposed to be the isoniazid binding site however, studies (Pierattelli et al., 2004) later presented the δ-meso edge of the heme as the favorable binding site for isoniazid in the catalase peroxidase enzymes. The proposed mode of isoniazid activation in the surface loop region was an electron transfer pathway that would have made it unique for catalase-peroxidase enzymes.

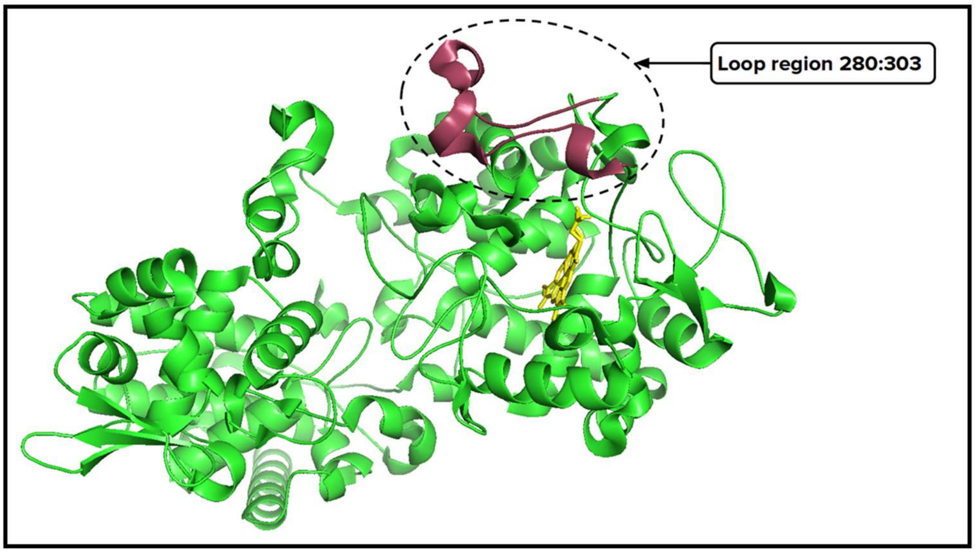


Figure 4.2: **KatG structure of the loop region with ≥ 2SD Δ*L* (residues 280 - 303: red).** The loop region showed decreased inter-residue distance across all katG variants.

An increase in inter-residue distance was noted in the region 26 - 38 across most variants (S140R, S315T, S315I, S315N, G316D, and S457I) compared to the wild type. This region forms part of the loop region (26 - 100) believed to facilitate dimerization of the katG protein (Bertrand et al., 2004).

To further investigate which residues are important in the variant's protein communication, the DRN *betweenness centrality* was calculated and discussed in the next section.

4.5.2 *Betweenness Centrality*

The *betweenness centrality* for the variants was calculated to identify which residues are important in the katG protein communication. Using the MD-TASK tool, the average *BC* for the katG wild type and for each of the protein variants was calculated and the difference between the wild type and variants average BC worked out (wild type BC minus variant BC). The residues with ΔBC of ≥ 2SD and ≤ -2SD from the mean were identified and mapped on the katG structure (Figure 4.3). In this case, the residues marked in blue had a negative ΔBC after subtracting the variants average BC from the wildtype. This meant that these residues were involved in more inter-residue shortest paths in the variants compared to the wild type communication network.

Irrespective of the location of the mutation in the katG, the residues with significant change in BC were identified (Appendix 4) especially around the interface region between the N and the C-terminal domains suggesting an allosteric effect of the mutations on the katG structure away from the active site.

The interface residues are responsible for interdomain communication between the N-terminal and the C-terminal domains in the protein structure. The mutations effect on the interface region could be impacting on the protein's communication network as the interface region is known to have a traffic of network communication.

Literature shows that besides interdomain interaction in one monomer, in the dimer *M. tuberculosis* katG protein, interdomain interactions also occur between N-terminal and C-terminal domains of different monomers (Bertrand et al., 2004).

Studies also show that the C-terminal domain is important in stabilizing the katG active site and preventing coordination of the distal histidine to the heme iron (Baker et al., 2004). Removal of the C-terminal domain from the protein structure significantly diminished the catalase and peroxidase function of the katG protein (Baker et al., 2004) implying that this domain is important in maintaining the structural architecture of the katG protein.

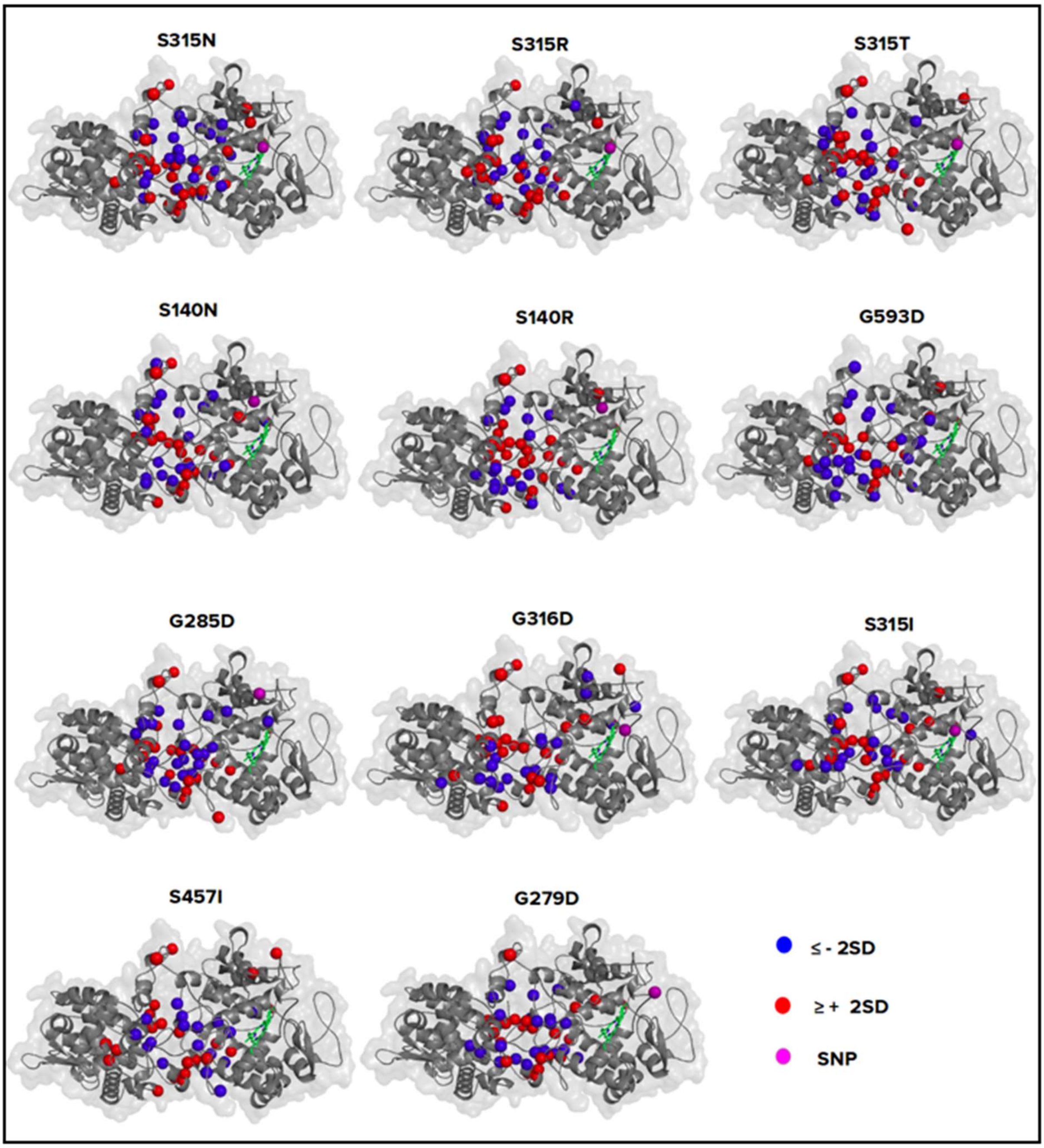


Figure 4.3: **KatG variant structures with mapped positions of residues with Δ*BC* ≥ 2SD (red) and ≤ -2SD (blue).** Pink shows the SNP position on the katG structure.

Literature suggests that there is an inverse relationship between the *BC* and *L* residues in a protein network. Consequently, the same relationship was observed between *BC* results and RMSF (Penkler et al, 2018).

Interestingly, more residues in the N-terminal domain were involved in the protein network shortest paths than those observed in the C-terminal domain across all variants (Table 4.1). The observed change in *BC* in the variants could be a copying mechanism of the katG to maintain the catalase function amidst the mutations.

Table 4.1: Distribution of Δ*BC* results (≤ -2SD) in the variant's N and C-terminal domains.

katG variant	Number of residues with ≤ -2SD *BC* in the N-terminal	Number of residues with ≤ -2SD *BC* in the C-terminal
S140R	9	9
S140N	15	6
G279D	14	7
G285D	18	8
S315T	14	7
S315R	15	7
S315N	20	4
S315I	13	7
G316D	12	8
S457I	12	5
G593D	13	12

4.6 CHAPTER SUMMARY

In this chapter, dynamic residue network analysis was used to determine residue accessibility (*reachability*) and the importance of the residues in the protein communication network (*betweenness centrality*) for both the katG wild type and variants. Molecular dynamics coupled with DRN analysis have previously shown useful in providing detailed insights into the understanding of single nucleotide polymorphism (SNPs) mechanism of action at the molecular level in addition to identifying the most important residues in the variants (Brown et al., 2017, Sanyanga et al., 2019).

Overall, across all the variants, there was an increase in accessibility for the region of residues 280 -303. This region forms part of the loop region conserved across three catalase-peroxidases that was once proposed to be the isoniazid binding site (Pierattelli et al., 2004). An increase in accessibility means reduced average *L* observed in the variants compared the wild type. A decrease in accessibility between regions 26 -38 meant increased average *L* in the katG dimer facilitating region of the variants as compared to the wild type.

From the *betweenness centrality* analysis, most change in *BC* was observed in the katG interface region between the N -terminal and C-terminal domains. Of the two domains, the N-terminal domain had the lion's share of the *BC* residues involved more in the inter-residue shortest paths. The negative change in *BC* (indicated in blue in Figure 4.3) meant increased residue importance in protein communication network of the variants as opposed to the wild type. The allosteric effect of the mutations in the katG interface region could be one of the compensatory mechanisms of the katG protein after undergoing mutation. In the next chapter, alanine scanning was used to determine which interface residues are important in the interdomain communication in a bid to zero in on the most important residues in the katG variant interface.

CHAPTER FIVE
ALANINE SCANNING

5.1 INTRODUCTION

In this chapter alanine scanning was used to identify katG N-terminal and C-terminal domain interface residues that are important in the domain binding and interdomain communication. This was in an attempt to further explain the observed change in *BC* results from Chapter 4. Alanine scanning is a computational-based approach used for identifying important residues responsible for protein - protein interaction at the protein interfaces. Since alanine scanning is useful in identifying interface residues, it not only can be used to identify residues at the protein - protein interface but also at domain - domain interfaces in a protein. In alanine scanning, residues at the protein - protein or domain - domain interface region are mutated to the alanine amino acid and a map of destabilising (important), stabilising and neutral (less important) residues generated (Kortemme et al., 2004). The effect of the deletion of an amino acid side chain beyond the C_β carbon atom on the affinity of a protein-protein complex is determined and this measure is used to quantify residue importance in the interaction (Kortemme et al., 2004). The important residues or "hot spots" as known by Clackson and Wells (Clackson & Wells, 1995) are believed to contribute the bulk of the binding energy in the protein - protein complexes. Alanine is the residue of choice for mutational scanning because it retains the beta carbon and no other side chain. Alanine also not only has a propensity to form alpha helices but can also occur in beta sheets (Kortemme et al., 2004). Alanine scanning was done using the ROBETTA webserver (Kim et al., 2004).

5.2 ROBETTA

ROBETTA is a webserver that uses free energy calculation to compute the effects of alanine mutation of the interface residues on the binding free energy of a protein-protein complex. An interface residue is one that contains one or more atoms within 4 Å radius of an atom belonging to the partner subunit (Kortemme et al., 2004). The free energy function consists of a linear combination of a Lennard-Jones potential to describe atomic packing interactions, an implicit solvation model, an orientation-dependent hydrogen-bonding potential derived from high-resolution protein structures, statistical terms approximating the backbone-dependent amino acid-type and rotamer probabilities and an estimate of unfolded reference state energies (Kortemme et al., 2004).

The input files to ROBETTA webserver are the protein-protein complex pdb structure file, the definition of the interface between the two partner (A, B) and finally an optional list of mutations. The webserver then uses the free energy function (Equation 5.1) to generate a table of predicted changes in binding free energy ($\Delta\Delta\ G_{bind}$) for all alanine mutations.

$$\Delta\ G = W_{attr}\ E_{LJattr} + W_{rep}E_{LJrep} + W_{HB(sc-bb)}E_{HB(sc-bb)} + W_{HB(sc-sc)}E_{HB(sc-sc)}$$

$$W_{sol}G_{sol} + W_{\varphi/\psi}\ E_{\varphi/\psi}\ (aa) + \Sigma_{aa}^{20}\ n_{aa}E_{aa}^{Ref}$$

E_{LJattr}: Attractive part of a Lennard-Jones potential.

E_{LJrep}: A linear distance-dependent repulsive term.

$E_{HB(sc-bb)}$:Orientation-dependent side chain-backbone hydrogen bond potential.

$E_{HB(sc-sc)}$: Orientation-dependent side chain-side chain hydrogen bond potential.

G_{sol}: Implicit solvation model.

W: Relative weights of the different energy terms.

$E_{\varphi/\psi\ (aa)}$: Amino-acid type (aa) dependent backbone torsion angle propensity.

E_{aa}^{Ref} : Amino-acid type dependent reference energy, which approximates the interactions made in the unfolded state ensemble.

Equation 5.1: **Binding free energy function used by ROBETTA**. The function calculates effects of alanine mutation on binding free energy of a protein-protein complex (Kortemme et al., 2004).

The output table consists of the following columns; column 1: number of mutated residue in the pdb file, column 2: pdb chain identifier, column 3: measure of whether a residue side chain atom is within 4 Å of an atom on the other partner, column 4: continuous residue numbering of all partners, column 5: amino acid type according to one-residue nomenclature in alphabetical order (1-A, 2-C, 3-D, 4-E ...), column 6: ΔΔG(bind), predicted change in binding free energy upon alanine mutation, column 7: ΔΔG(bind, obs) observed changes in binding free energy upon alanine mutation and finally column 8: G(partner), predicted change in protein stability of the mutated complex partner upon alanine mutation.

The hot spot residues are considered as those with predicted binding free energy ≥ 1 kcal/mol, neutral residues as those with binding energy value between -0.8 and 0.99 kcal/mol and the stabilising residues as those with binding energy value < - 0.8kcal/mol.

5.3 PROTEIN INTERACTION CALCULATOR (PIC)

In addition to determining the key residues in the N and C-terminal domain interaction, the protein interaction calculator was used to identify all the N and C-terminal interface residues involved in domains interaction irrespective of whether they were hotspot residues or not. The idea was to compare this information with the change in *BC* data and identify which *BC* residues are in the interface region and later identify which *BC* residues in the interface region are also hotspot residues.

Protein interaction calculator is a webserver that takes the protein structure coordinate file as input and determines the type of interactions and the residues involved in a protein - protein interaction (Tina et al., 2007) and in this case, the domain - domain interaction. The server is able to calculate disulphide bonds, interactions between hydrophobic residues, ionic interactions, hydrogen bonds, aromatic - aromatic interactions, aromatic - sulphur interactions and cation-π interactions within a protein or between proteins in a complex (Tina et al., 2007).

5.4 METHODOLOGY

5.4.1 Alanine scanning

In order to determine important interface residues using ROBETTA, a Python script was used to edit the katG (monomer) pdb structure file of the wild type and variants homology models, naming the N-terminal domain residues as chain A and the C-terminal domain residues as chain B. The wild type and variant coordinate structure files (pdb) were submitted to the ROBETTA webserver to generate an output table file. A Python script was used to identify the residues in the output file that had binding free energies of ≥ 1 kcal/mol (destabilising residues) and < -0.8kcal/mol (stabilizing residues). The residues with binding energy ≥ 1 kcal/mol were considered as the hot spot residues.

5.4.2 Protein interaction calculation

The model katG wild type pdb coordinate file created using the Python script in section 5.4.1 was submitted to the PIC webserver and all the interactions selected i.e. disulphide bonds interactions, hydrophobic residues interactions, ionic interactions, hydrogen bonds interactions, and aromatic–aromatic interactions. The output file was a list of all the residues involved in the respective interactions and their residue numbers.

5.5 RESULTS

The PIC server was used to identify the katG interface residues responsible for the N and C – terminal domain interaction (Table 5.1) and the ROBETTA webserver was used for alanine scanning to identify the destabilizing and stabilizing residues in katG variant interface region. From alanine scanning, only the destabilizing residues were identified together with their binding energy (Appendix 5). The change in *BC* residues that are also interface residues from the PIC results were identified to narrow down the *BC* results to only interface residues (Table 5.2 and Table 5.3)

Table 5.1: **Table of the N and C-terminal domain interface residues.**

Domain	**Interface residues**
N-terminal domain	F183, F185, G421, V423, A424, R425, R418, I115, H116, D117, G118, N41, V47, R42, Q50, Y113, R119, L43, L45, L48, A65, V68, V73, F167, F181, P422
C-terminal domain	L430, L427, D487, Y426, V431, R484, S486, Y608, N615, N701, D612, L611, V697, L704, A621, A614, M624, P429, V586, P432

Table 5.2: **Change in BC residues in the N-terminal interface region**.

Variant	**Δ*BC***	***BC* residues in the N-terminal interface residues**
S140R	≥2SD	D117, G118, R418
	≤ -2SD	**L48**, Q50, H116
S140N	≥2SD	G118, R418
	≤ -2SD	V47, **L48**, Q50, I115, H116
G279D	≥2SD	R418
	≤ -2SD	**L48**, Q50, I115, H116, A424
G285D	≥2SD	R418
	≤ -2SD	**L43**, V47, **L48**, Q50, **Y113**, I115, H116, D117, G118
S315T	≥2SD	R418
	≤ -2SD	**L43, L45, L48**, Y113, **D117**, G118
S315R	≥2SD	D117
	≤ -2SD	**L48**, Q50, H116, G118, A424
S315N	≥2SD	**D117**, R418
	≤ -2SD	**L48**, Q50, H116, G421, P422
S315I	≥2SD	D117, R418
	≤ -2SD	**L48,** Q50, G118
G316D	≥2SD	R418
	≤ -2SD	I115, H116
S457I	≥2SD	
	≤ -2SD	
G593D	≥2SD	Y113, **D117**
	≤ -2SD	Q50

The residues in red were identified as hotspot residues by alanine scanning.

Table 5.3: **Change in *BC* residues in the C-terminal interface region**.

Variant	**Δ*BC***	***BC* and C-terminal interface residues**
S140R	≥2SD	**R484**, S486, D487
	≤ -2SD	V586
S140N	≥2SD	**D487**
	≤ -2SD	V586, **N615**
G279D	≥2SD	S486, **D487**
	≤ -2SD	V586, N615
G285D	≥2SD	**D487**
	≤ -2SD	V586
S315T	≥2SD	S486, D487
	≤ -2SD	V586
S315R	≥2SD	**D487**, **Y608**
	≤ -2SD	R484, S486
S315N	≥2SD	R484, **D487**
	≤ -2SD	**S486**
S315I	≥2SD	S486, **D487**
	≤ -2SD	**R484**
G316D	≥2SD	**R484, D487**
	≤ -2SD	D612
S457I	≥2SD	**D487**
	≤ -2SD	**Y608**, N615
G593D	≥2SD	**R484, S486**
	≤ -2SD	V586

The residues in red were identified as hotspot residues by alanine scanning.

The identified residues with significant change in *BC* from the interface region were compared to the results from alanine scanning to check for any data correlation. Since *betweenness centrality* was used to identify residues with increased importance in katG variants communication network and alanine scanning used to identify hotspot residues, comparing the data helped identify the destabilizing residues in the katG protein interface that were both involved more and less in the inter-residue communication of the variants.

From the comparison, residue Leucine at position 48 (L48) was identified as a destabilizing residue in the N-terminal domain with increased importance across all variants except (G316D, S457I and G593D). Variants S457I and G593D is located in the C-terminal and while 457I did not have any significant change in *BC* residues in the N-terminal domain, G593D only had three residues with significant change in *BC* in the N -terminal interface. This suggests that variant S457I effect on the interface communication between the N and C-terminal domain could be localised to the C-terminal interface residues only. In addition, a significant change in *BC* for residues; L48, Q50, H116, D117 was identified across most variants (S140R, S140N, G279D, G285D, S315T, S315R, S315N and G316D). This suggests a common communication pattern across most of the mutants in the katG protein that involves the mentioned residues.

In the C-terminal domain, destabilizing residue Aspartic acid at position 487 had less importance in the protein communication across all variants except (S140R and S315T). Like in the N-terminal, in the C-terminal, residues; V586, N615, R484 and S486 were identified as important in residue communication across most variants compared to the wildtype. The destabilizing residues with increased importance for all the variants were plotted on the katG structure (Figure 5.1). Figure 5.1 also shows the residues of interest forming a communication path from the C-terminal to the -terminal domain emphasizing further the importance of the interface region in the katG variants.

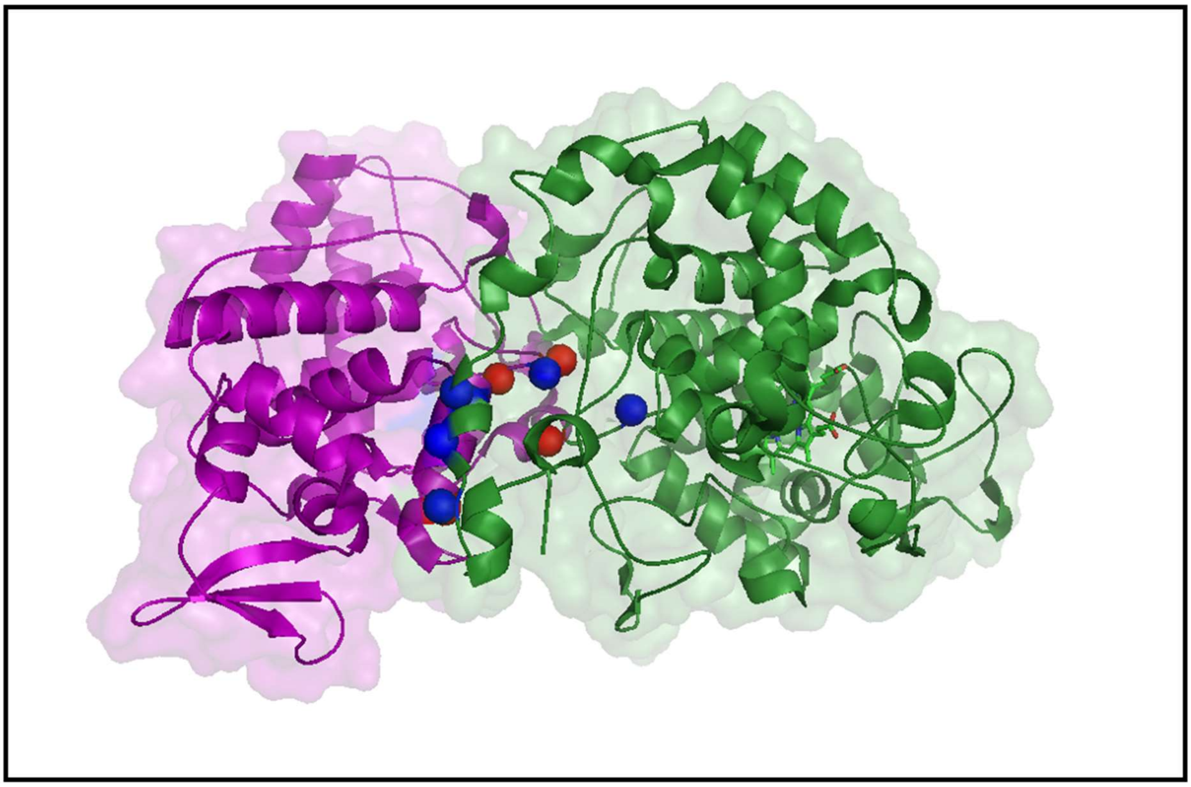


Figure 5.1: KatG structure with mapped change in *BC* residues identified as destabilising residues by alanine scanning. The N-terminal domain is green, C-terminal domain: purple, destabilizing residues with ≤ -2SD change in *BC* are blue spheres while destabilizing residues with ≥2SD change in *BC* are the red spheres.

5.5 CHAPTER SUMMARY

In this chapter, alanine scanning was used to determine the key interface residues in N and C-terminal domains of katG wild type and variants that are responsible for the bulk of the protein binding energy. From the alanine scanning residues; L43, L48, Y113, R119, D487, F737, Y608, L611, L437 and R484 were identified as the important/ key residues in the katG interdomain communication. Most of the identified residues were consistent across the different variants suggesting a common mechanism of inter-domain interaction across the variants. This observation was expected as variant homology models were used as opposed to last frame from the MD simulation trajectory which could have shown different results. Calculations using the last frame of the MD trajectory were not done in the interest of time and are to be completed in the future work. The DRN change in *BC* data was compared to the alanine scanning results and common residues across the different data sets identified for the different variants. According to alanine scanning, these residues are destabilising residues and responsible for the binding energy between the two domains. As per *BC* data, these residues showed increased importance in the residue network in the respective variants as compared to the wild type. Since the C-terminal domain is necessary for the protein’s catalase and peroxidase function, there seems to be more need for protein inter-domain communication in the variants than in the wild type.

CONCLUDING REMARKS AND FUTURE WORK

The *M. tuberculosis* catalase peroxidase (katG) protein is a 740 amino acid long homodimer containing a heme group in the N-terminal domain. The protein is responsible for protecting the bacteria from toxic radicals i.e. hydrogen peroxide present in the aerobic environment. The enzyme also activates the TB first-line pro-drug isoniazid that has anti-tubercular activity through inhibiting the synthesis of mycolic acids which are important in the bacteria cell wall synthesis.

Single nucleotide mutations in the katG have previously been reported (Lempens et al., 2018, Feuerriegel et al., 2015, Sandgren et al., 2009) to lead to isoniazid drug resistance. In this study, the high confidence mutations in the *M. tuberculosis* katG protein were investigated to gain insight in their mechanism of action.

To understand the effect of the high confidence mutations on the katG protein in terms of change in protein stability, flexibility, conformational change, interaction with isoniazid and change in commutation network, the holo and complex (katG -isoniazid) wild type and variants katG proteins were modelled and molecular dynamic simulations run on the models (Figure 6). In addition, trajectory analysis, dynamic residue analysis, and alanine scanning calculations were done to determine the structural changes, the protein communication network and to identify the key residues in the inter-domain interaction respectively.

Interestingly, from the MD simulations, the isoniazid (ligand) remained coordinated in the katG active site for only 22 ns and analysis of the drug's coordinating residues showed reduced hydrogen bond interactions in the both the template and models as the simulation progressed. Results from the 200 ns MD trajectory analysis of the holo katG and variant structures showed that the mutations had varying conformational and protein flexibility effects on the katG structure. Variants S315I, S315N, S140R, G285D, G316D and G279D had an increased backbone flexibility while S315R showed a reduced backbone flexibility when compared to the wild type. These results were further supported by the principal component analysis data which in addition implicated variants S315R, S457I, S315T, S140N and G279D as having less conformational variability compared to the wild type. Furthermore, PCA identified mutants S140R, S140N, G279D, G285D, S315I and G316D as having increased energy barriers between conformations hence increasing their energy requirements for conformational change. Literature indicates that conformational changes are essential for ligand binding and enzymatic catalysis (Kurplus & McCammon, 1983).

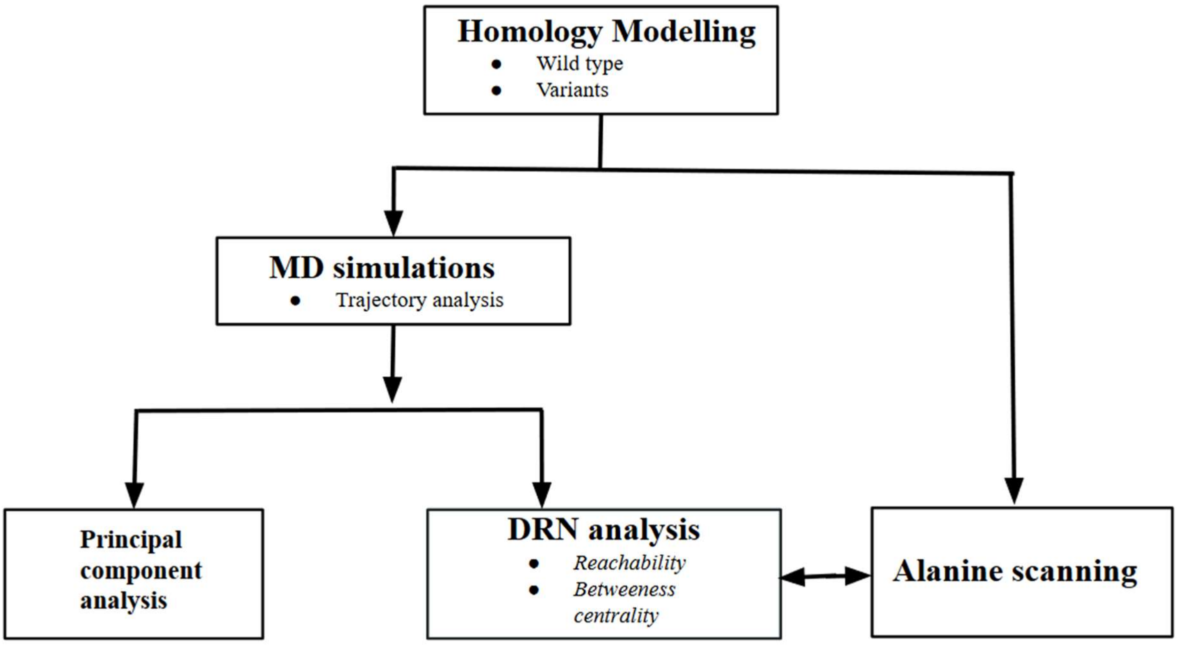


Figure 6: Flow diagram of the methodology followed in the study.

Pertaining the katG structure stability, variants S140R and G316D in the N-terminal domain showed increased structural instability compared to the wild type and the rest of the variants. Residue flexibility analysis identified the region 26-110 having less flexibility in all katG variants. This is a loop region believed to facilitate the dimerization of the katG structure in solution (Bertrand et al., 2004). This region also contains active site residues His-108, Arg-104 and Trp-107. Change in the active site residue behavior could be one of the many ways by which the mutations prevent isoniazid activation. In addition, all the variants displayed an increased structural compactness compared to the wild type. This observation is supported by literature from Marney et al, (2017) which indicated a reduction in the katG active site access channel as a result of increased compactness in the mutants.

DRN ΔL and ΔBC analysis identified the katG interface region as one containing residues most important in the mutants' communication. Alanine scanning was used to identify which residues in the interface were destabilizing (hotspots) or stabilizing. A comparison of the change in BC and alanine scanning data showed residue L48 as the most important destabilizing residue in the variants and residues L43, L48, Y113, R119, D487, F737, Y608, L611, L437 and R484 as the common destabilizing in the all variants. These structural changes observed in the holo katG and variants protein structures are indicative of the mechanisms of katG resistance to isoniazid. In summary, the high confidence mutations were noticed to have both varying and similar effects on the behavior of katG protein structure. Some of these effects included; reduction in conformational space and variability, reduction in residue fluctuation especially in the dimerization loop, increase in steric hindrance to prevent isoniazid binding, increase in structural compactness and requirement of more energy to shift between structural conformations. Finally, this study identified residues in the katG interface region as the most important in the variant's communication network unlike in the wildtype. This could be a compensatory mechanism to maintain the katG catalase activity in the mutants.

Since the katG protein functions as a dimer in solution, these findings can be used in future, to study the katG dimer resistance mechanisms while in complex with isoniazid. In addition, alanine scanning and PIC calculations will be done using the last frame from MD simulations as opposed to the homology models to clearly observe the change in protein behavior.

REFERENCES

1. World Health Organization. Global tuberculosis report 2018. Geneva; 2018

2. Churchyard, G., Kim, P., Shah, N., Rustomjee, R., Gandhi, N., & Mathema, B. et al. (2017). What We Know About Tuberculosis Transmission: An Overview. The Journal of Infectious Diseases.

3. Sotgiu, G., Centis, R., D'ambrosio, L., & Migliori, G. (2015). Tuberculosis Treatment and Drug Regimens. Cold Spring Harbor Perspectives in Medicine.

4. Nahid, P., Dorman, S., Alipanah, N., Barry, P., Brozek, J., & Cattamanchi, A. et al. (2016). Executive Summary: Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis. Clinical Infectious Diseases.

5. Treatment for TB Disease | Treatment | TB | CDC. (2019). Retrieved from https://www.cdc.gov/tb/topic/treatment/tbdisease.htm.

6. Kumar, K., & Kon, O. (2017). Diagnosis and treatment of tuberculosis: latest developments and future priorities. Annals of Research Hospitals.

7. Jeon, D. (2017). WHO Treatment Guidelines for Drug-Resistant Tuberculosis, 2016 Update: Applicability in South Korea. Tuberculosis and Respiratory Diseases.

8. Sakula, A. (1983). Carlo Forlanini, inventor of artificial pneumothorax for treatment of pulmonary tuberculosis. Thorax.

9. Goldsworthy PD, McFarlane AC. 2002. Howard Florey, Alexander Fleming and the fairy tale of penicillin. Med Journal A.

10. Crofton J. 1960. Tuberculosis undefeated. Br Med Journal.

11. Crofton J. 1969. Some principles in the chemotherapy of bacterial infections. Br Med Journal.

12. Updated TB guidelines raise upper age limit for treating latent disease. (2016). The Pharmaceutical Journal.

13. Namdar, R., & Peloquin, C. A. (2018). Drugs for tuberculosis. In Drug Interactions in Infectious Diseases: Antimicrobial Drug Interactions. Humana Press.

14. Argyrou, A., Jin, L., Siconilfi-Baez, L., Angeletti, R. H., & Blanchard, J. S. (2006). Proteome-wide profiling of isoniazid targets in Mycobacterium tuberculosis.

15. Singh, P., Kant, S., Gaur, P., Tripathi, A., & Pandey, S. (2018). Extra Pulmonary Tuberculosis: An Overview and Review of Literature. Int. J. Life. Sci. Scienti. Res.

16. Ando, H., Miyoshi-Akiyama, T., Watanabe, S., & Kirikae, T. (2014). A silent mutation in mabA confers isoniazid resistance on Mycobacterium tuberculosis. Molecular microbiology.

17. Velayati, A. A., Masjedi, M. R., Farnia, P., Tabarsi, P., Ghanavi, J., ZiaZarifi, A. H., & Hoffner, S. E. (2009). Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in Iran.

18. Zhang, F. (2017). Computational exploration of transmission and acquisition of drug-resistant tuberculosis (Doctoral dissertation, Applied Sciences: School of Computing Science).

19. Trauner, A., Liu, Q., Via, L. E., Liu, X., Ruan, X., Liang, L., & Zhang, W. (2017). The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. Genome biology.

20. Suarez, J., Ranguelova, K., Jarzecki, A., Manzerova, J., Krymov, V., & Zhao, X. et al. (2009). An Oxyferrous Heme/Protein-based Radical Intermediate Is Catalytically Competent in the Catalase Reaction of Mycobacterium tuberculosis Catalase-Peroxidase (KatG). Journal of Biological Chemistry.

21. Timmins, G., Master, S., Rusnak, F., & Deretic, V. (2004). Nitric Oxide Generated from Isoniazid Activation by KatG: Source of Nitric Oxide and Activity against Mycobacterium tuberculosis. Antimicrobial Agents and Chemotherapy.

22. Kweza, P. F., Van Schalkwyk, C., Abraham, N., Uys, M., Claassens, M. M., & Medina-Marino, A. (2018). Estimating the magnitude of pulmonary tuberculosis patients missed by primary health care clinics in South Africa. The International Journal of Tuberculosis and Lung Disease.

23. Jena, L., Deshmukh, S., Waghmare, P., Kumar, S., & Harinath, B. C. (2015). Study of mechanism of interaction of truncated isoniazid–nicotinamide adenine dinucleotide adduct against multiple enzymes of Mycobacterium tuberculosis by a computational approach. International journal of mycobacteriology.

24. Haas, W. H., Schilke, K., Brand, J., Amthor, B., Weyer, K., Fourie, P. B & Bremer, H. J. (1997). Molecular analysis of katG gene mutations in strains of Mycobacterium tuberculosis complex from Africa. Antimicrobial agents and chemotherapy.

25. Huard, R. C., de Oliveira Lazzarini, L. C., Butler, W. R., van Soolingen, D., & Ho, J. L. (2003). PCR-based method to differentiate the subspecies of the Mycobacterium tuberculosis complex on the basis of genomic deletions. Journal of clinical microbiology.

26. Chai, Q., Zhang, Y., & Liu, C. H. (2018). Mycobacterium tuberculosis: An Adaptable Pathogen Associated with Multiple Human Diseases. Frontiers in cellular and infection microbiology.

27. Holt, K. E., McAdam, P., Phan, V. K. T., Dang, T. M. H., Nguyen, N. L., Nguyen, H. L., ... & Pham, K. (2016). Genomic analysis of Mycobacterium tuberculosis reveals complex etiology of tuberculosis in Vietnam including frequent introduction and transmission of Beijing lineage and positive selection for EsxW Beijing variant. BioRxiv.

28. Barberis, I., Bragazzi, N. L., Galluzzo, L., & Martini, M. (2017). The history of tuberculosis: from the first historical records to the isolation of Koch's bacillus. Journal of preventive medicine and hygiene.

29. Public Health England. (2016). Tuberculosis in England: annual report. Retrieved from https://www.gov.uk/government/publications/tuberculosis-in-england-annual-report

30. Petruccioli, E., Scriba, T. J., Petrone, L., Hatherill, M., Cirillo, D. M., Joosten, S. A., ... & Goletti, D. (2016). Correlates of tuberculosis risk: predictive biomarkers for progression to active tuberculosis. European Respiratory Journal.

31. Agrawal, M., Bajaj, A., Bhatia, V., & Dutt, S. (2016). Comparative study of GeneXpert with ZN stain and culture in samples of suspected pulmonary tuberculosis. Journal of clinical and diagnostic research. JCDR.

32. Zumla, Alimuddin, Mario Raviglione, Richard Hafner, and C. Fordham von Reyn. "Current concepts." N Engl J Med.

33. Shi, Ruiru, Nobunori Itagaki, and Isamu Sugawara. "Overview of anti-tuberculosis (TB) drugs and their resistance mechanisms." Mini reviews in medicinal chemistry.

34. Oni, T., Youngblood, E., Boulle, A., McGrath, N., Wilkinson, R. J., & Levitt, N. S. (2015). Patterns of HIV, TB, and non-communicable disease multi-morbidity in peri-urban South Africa-a cross sectional study. BMC infectious diseases.

35. Ogwang, S., Asiimwe, B. B., Traore, H., Mumbowa, F., Okwera, A., Eisenach, K. D., ... & Ayakaka, I. (2009). Comparison of rapid tests for detection of rifampicin-resistant Mycobacterium tuberculosis in Kampala, Uganda. BMC infectious diseases.

36. Lawn, S. D., Kerkhoff, A. D., Burton, R., Schutz, C., Boulle, A., Vogt, M., ... & Meintjes, G. (2017). Diagnostic accuracy, incremental yield and prognostic value of Determine TB-LAM for routine diagnostic testing for tuberculosis in HIV-infected patients requiring acute hospital admission in South Africa: a prospective cohort. BMC medicine.

37. Stevens, W. S., Scott, L., Noble, L., Gous, N., & Dheda, K. (2017). Impact of the Gene-Xpert MTB/RIF Technology on Tuberculosis Control. Microbiology spectrum.

38. Ismail, N., Omar, S., Mvusi, L., & Madhi, S. (2018). Prevalence of drug-resistant tuberculosis in South Africa – Authors' reply. The Lancet Infectious Diseases.

39. Ismail, N. A., Mvusi, L., Nanoo, A., Dreyer, A., Omar, S. V., Babatunde, S., ... & Ihekweazu, C. (2018). Prevalence of drug-resistant tuberculosis and imputed burden in South Africa: a national and sub-national cross-sectional survey. The Lancet Infectious Diseases.

40. Hartel, L. A., Yazbeck, A. S., & Osewe, P. L. (2018). Responding to Health System Failure on Tuberculosis in Southern Africa. Health Systems & Reform.

41. Sharp, A., Donahoe, J. T., Milliken, A., Barocio, J., Charalambous, S., & McLaren, Z. M. (2018). Do Incarcerated Populations Serve as a Reservoir for Tuberculosis in South Africa? The American journal of tropical medicine and hygiene.

42. Bozzani, F. M., Mudzengi, D., Sumner, T., Gomez, G. B., Hippner, P., Cardenas, V., ... & Vassall, A. (2018). Empirical estimation of resource constraints for use in model-based economic evaluation: an example of TB services in South Africa. Cost effectiveness and resource allocation.

43. Cox, H., Dickson-Hall, L., Ndjeka, N., van't Hoog, A., Grant, A., Cobelens, F., ... & Nicol, M. (2017). Delays and loss to follow-up before treatment of drug-resistant tuberculosis following implementation of Xpert MTB/RIF in South Africa: a retrospective cohort study. PLoS medicine.

44. South African National AIDS Council. (2019). National Strategic Plan for HIV, TB and STIs (2017-2022) | SANAC. Retrieved from https://sanac.org.za/download-the-full-version-of-the-national-strategic-plan-for-hiv-tb-and-stis-2017-2022-2/.

45. Robertson, K. R., Oladeji, B., Jiang, H., Kumwenda, J., Supparatpinyo, K., Campbell, T. B., ... & Kumarasamy, N. (2018). Human Immunodeficiency Virus Type 1 and Tuberculosis Coinfection in Multinational, Resource-limited Settings: Increased Neurological Dysfunction. Clinical Infectious Diseases.

46. Skinner, Donald, and Mareli Claassens. "It's complicated: why do tuberculosis patients not initiate or stay adherent to treatment? A qualitative study from South Africa." BMC infectious diseases.

47. Ebrahimzadeh, A., Mohammadifard, M., & Naseh, G. (2014). Comparison of Chest X-Ray Findings of Smear Positive and Smear Negative Patients with Pulmonary Tuberculosis. Iranian Journal of Radiology.

48. Singh, A., Grover, S., Pandey, B., Kumari, A. and Grover, A. (2018). Wild-type catalase peroxidase vs G279D mutant type: Molecular basis of Isoniazid drug resistance in Mycobacterium tuberculosis. Gene.

49. Yu, S., Girotto, S., Lee, C. and Magliozzo, R. (2003). Reduced Affinity for Isoniazid in the S315T Mutant of Mycobacterium tuberculosis KatG Is a Key Factor in Antibiotic Resistance. Journal of Biological Chemistry.

50. Larsen, M., Vilchèze, C., Kremer, L., Besra, G., Parsons, L., Salfinger, M., Heifets, L., Hazbon, M., Alland, D., Sacchettini, J. and Jacobs, W. (2002). Overexpression of inhA, but notkasA, confers resistance to isoniazid and ethionamide in Mycobacterium smegmatis, M. bovis BCG and M. tuberculosis. Molecular Microbiology.

51. Silva, M., Senna, S., Ribeiro, M., Valim, A., Telles, M., Kritski, A., Morlock, G., Cooksey, R., Zaha, A. and Rossetti, M. (2003). Mutations in katG, inhA, and ahpC Genes of Brazilian Isoniazid-Resistant Isolates of Mycobacterium tuberculosis. Journal of Clinical Microbiology.

52. Bertrand, T., Eady, N., Jones, J., Jesmin, Nagy, J., & Jamart-Grégoire, B. et al. (2004). Crystal Structure of Mycobacterium tuberculosis Catalase-Peroxidase. Journal of Biological Chemistry.

53. Weber, W., & Hein, D. (1979). Clinical Pharmacokinetics of Isoniazid. Clinical Pharmacokinetics.

54. Fernandes, G., Salgado, H., & Santos, J. (2017). Isoniazid: A Review of Characteristics, Properties and Analytical Methods. Critical Reviews in Analytical Chemistry.

55. G. Pretorius, P. van Helden, F. Sirgel, K. Eisenach, T. Victor, Antimicrobial Agents Chemotherapy.

56. Saint-joanis, B., Souchon, H., Wilming, M., Johnsson, K., Alzari, P., & Cole, S. (1999). Use of site-directed mutagenesis to probe the structure, function and isoniazid activation of the catalase/peroxidase, KatG, from Mycobacterium tuberculosis. Biochemical Journal.

57. Mdluli, K. (1998). Inhibition of a Mycobacterium tuberculosis -Ketoacyl ACP Synthase by Isoniazid. Science.

58. Morris, S., Bai, G., Suffys, P., Portillo-Gomez, L., Fairchok, M., & Rouse, D. (1995). Molecular Mechanisms of Multiple Drug Resistance in Clinical Isolates of Mycobacterium tuberculosis. Journal of Infectious Diseases.

59. Unissa, A., Doss C, G., Kumar, T., Sukumar, S., Lakshmi, A., & Hanna, L. (2018). Significance of catalase-peroxidase (KatG) mutations in mediating isoniazid resistance in clinical strains of Mycobacterium tuberculosis. Journal of Global Antimicrobial Resistance.

60. Sali, A. (1995). Comparative protein modelling by satisfaction of spatial restraints. Molecular Medicine Today.

61. Carpena, X., Switala, J., Loprasert, S., Mongkolsuk, S., Fita, I., & Loewen, P. (2002). Crystallization and preliminary X-ray analysis of the catalase–peroxidase KatG from Burkholderia pseudomallei. Acta Crystallographica Section D Biological Crystallography.

62. Edwards, S., Nguyen Huu Xuong, Hamlin, R., & Kraut, J. (1987). Crystal structure of cytochrome c peroxidase compound I. Biochemistry.

63. Pierattelli, R., Banci, L., Eady, N., Bodiguel, J., Jones, J., & Moody, P. et al. (2004). Enzyme-catalyzed Mechanism of Isoniazid Activation in Class I and Class III Peroxidases. Journal of Biological Chemistry.

64. Glaziou, P., Floyd, K., & Raviglione, M. (2009). Global Burden and Epidemiology of Tuberculosis. Clinics in Chest Medicine.

65. Ahmed, G., Mohammed, A., Taha, A., Almatroudi, A., Allemailem, K., Babiker, A., & Alsammani, M. (2019). Comparison of the Microwave-Heated Ziehl-Neelsen Stain and Conventional Ziehl-Neelsen Method in the Detection of Acid-Fast Bacilli in Lymph Node Biopsies. Open Access Macedonian Journal Of Medical Sciences.

66. Liu, Y., Matsumoto, M., Ishida, H., Ohguro, K., Yoshitake, M., & Gupta, R. et al. (2018). Delamanid: From discovery to its use for pulmonary multidrug-resistant tuberculosis (MDR-TB). Tuberculosis.

67. Palomino, J., & Martin, A. (2014). Drug Resistance Mechanisms in Mycobacterium tuberculosis. Antibiotics.

68. Bostanabad, S. (2011). Study of Genetic Evolution in Mycobacterium tuberculosis Isolates from Patients with Active Pulmonary Tuberculosis in the Iran and Belarus. The Open Microbiology Journal.

69. Lari, N., Rindi, L., Bonanni, D., Tortoli, E., & Garzelli, C. (2006). Molecular Analysis of Clinical Isolates of Mycobacterium bovis Recovered from Humans in Italy. Journal of Clinical Microbiology.

70. Broger, T., Sossen, B., du Toit, E., Kerkhoff, A., Schutz, C., & Ivanova Reipold, E. et al. (2019). Novel lipoarabinomannan point-of-care tuberculosis test for people with HIV: a diagnostic accuracy study. The Lancet Infectious Diseases.

71. Welinder, K. (1992). Superfamily of plant, fungal and bacterial peroxidases. Current Opinion In Structural Biology, 2(3), 388-393.

72. Jilani, T., Siddiqui, A., Amma, M. and Filippava, I. (2019). Active Tuberculosis.

73. Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. Journal of Molecular Graphics.

74. Cohen, K., Manson, A., Desjardins, C., Abeel, T., & Earl, A. (2019). Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges. Genome Medicine.

75. Henzler-Wildman, K., & Kern, D. (2007). Dynamic personalities of proteins. Nature.

76. Cavasotto, C., & Phatak, S. (2009). Homology modelling in drug discovery: current trends and applications. Drug Discovery Today.

77. Xiang, Z. (2006). Advances in Homology Protein Structure Modelling. Current Protein & Peptide Science.

78. Prasad, J., Comeau, S., Vajda, S., & Camacho, C. (2003). Consensus alignment for reliable framework prediction in homology modelling. Bioinformatics.

79. Larsson, P., Wallner, B., Lindahl, E., & Elofsson, A. (2008). Using multiple templates to improve quality of homology models in automated homology modelling. Protein Science.

80. Fiser, A. (2010). Template-Based Protein Structure Modelling. Methods In Molecular Biology.

81. Salilab.org. (2019). About MODELLER.

82. Engh, R. and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallographica Section A Foundations of Crystallography.

83. Morris, A., MacArthur, M., Hutchinson, E. and Thornton, J. (1992). Stereochemical quality of protein structure coordinates. Proteins: Structure, Function, and Genetics.

84. Wiederstein, M. and Sippl, M. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research.

85. Xiong, J. (2006). Essential bioinformatics. New York: Cambridge University Press.

86. Henriksen, A., Schuller, D., Meno, K., Welinder, K., Smith, A., & Gajhede, M. (1998). Structural Interactions between Horseradish Peroxidase C and the Substrate Benzhydroxamic Acid Determined by X-ray Crystallography. Biochemistry.

87. Aitken, S., Turnbull, J., Percival, M., & English, A. (2001). Thermodynamic Analysis of the Binding of Aromatic Hydroxamic Acid Analogues to Ferric Horseradish Peroxidase. Biochemistry.

88. Fahim, A., Mukhopadhyay, R., Yandle, R., Prestegard, J. and Valafar, H. (2013). Protein Structure Validation and Identification from Unassigned Residual Dipolar Coupling Data Using 2D-PDPA. Molecules.

89. Notredame, C., Higgins, D., & Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. Journal of Molecular Biology.

90. Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search tool. Journal of Molecular Biology.

91. Soding, J., Biegert, A., & Lupas, A. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research.

92. Hatherley, R., Brown, D., Glenister, M., & Tastan Bishop, Ö. (2016). PRIMO: An Interactive Homology Modelling Pipeline. Plos one.

93. Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research.

94. Barton, G. (1992). Computer speed and sequence comparison. Science.

95. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D., & Notredame, C. (2004). 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments. Journal of Molecular Biology.

96. Shi, J., Blundell, T., & Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties11Edited by B. Honig. Journal of Molecular Biology.

97. Higgins, D., & Sharp, P. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene.

98. Edgar R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics.

99. Notredame, C., Higgins, D., & Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. Journal of Molecular Biology.

100. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., & Gumienny, R. et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research.

101. Šali, A., & Blundell, T. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology.

102. Vriend, G. (1990). WHAT IF: a molecular modelling and drug design program. Journal of molecular graphics.

103. Shen, M., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Science.

104. Hooft, R., Vriend, G., Sander, C., & Abola, E. (1996). Errors in protein structures. Nature.

105. Laskowski, R., MacArthur, M., Moss, D., & Thornton, J. (1993). PROCHECK: a program to check the stereo-chemical quality of protein structures. Journal of Applied Crystallography.

106. Sippl, M. (1993). Recognition of errors in three-dimensional structures of proteins. Proteins: Structure, Function, and Genetics.

107. Eisenberg, D., Lüthy, R., & Bowie, J. (1997). [20] VERIFY3D: Assessment of protein models with three-dimensional profiles. Methods in Enzymology.

108. Cade, C., Dlouhy, A., Medzihradszky, K., Salas-Castillo, S., & Ghiladi, R. (2010). Isoniazid-resistance conferring mutations in Mycobacterium tuberculosis KatG: Catalase, peroxidase, and INH-NADH adduct formation activities. Protein Science.

109. Özen, A., Haliloğlu, T., & Schiffer, C. (2011). Dynamics of Preferential Substrate Recognition in HIV-1 Protease: Redefining the Substrate Envelope. Journal of Molecular Biology.

110. Alonso, H., Bliznyuk, A., & Gready, J. (2006). Combining docking and molecular dynamic simulations in drug design. Medicinal Research Reviews.

111. González, M. (2011). Force fields and molecular dynamics simulations. École Thématique De La Société Française De La Neutronique.

112. Alder, B., & Wainwright, T. (1959). Studies in Molecular Dynamics. I. General Method. The Journal of Chemical Physics.

113. Stillinger, F., & Rahman, A. (1974). Improved simulation of liquid water by molecular dynamics. The Journal of Chemical Physics.

114. Gelpi, J., Hospital, A., Goñi, R., & Orozco, M. (2015). Molecular dynamics simulations: advances and applications. Advances and Applications in Bioinformatics and Chemistry.

115. Chong, L., Duan, Y., Wang, L., Massova, I., & Kollman, P. (1999). Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. Proceedings of the National Academy of Sciences.

116. Huang, D., & Caflisch, A. (2011). The Free Energy Landscape of Small Molecule Unbinding. Plos Computational Biology.

117. Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M., Smith, J., Kasson, P., van der Spoel, D., Hess, B. and Lindahl, E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics.

118. Case, D., Cheatham, T., Darden, T., Gohlke, H., Luo, R., Merz, K., Onufriev, A., Simmerling, C., Wang, B. and Woods, R. (2005). The Amber biomolecular simulation programs. Journal of Computational Chemistry.

119. Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry.

120. Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R., Kalé, L. and Schulten, K. (2005). Scalable molecular dynamics with NAMD. Journal of Computational Chemistry.

121. Abraham, M., Murtola, T., Schulz, R., Páll, S., Smith, J., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. Softwarex.

122. Malde, A., Zuo, L., Breeze, M., Stroet, M., Poger, D., & Nair, P. et al. (2011). An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. Journal of Chemical Theory and Computation.

123. Malde, A., Zuo, L., Breeze, M., Stroet, M., Poger, D., & Nair, P. et al. (2011). An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. Journal of Chemical Theory and Computation.

124. Kurplus, M., & McCammon, J. (1983). Dynamics of Proteins: Elements and Function. Annual Review of Biochemistry.

125. Ross, C., Nizami, B., Glenister, M., Sheik Amamuddy, O., Atilgan, A., Atilgan, C., & Tastan Bishop, Ö. (2018). MODE-TASK: large-scale protein motion tools. Bioinformatics.

126. Suarez, J., Ranguelova, K., Schelvis, J., & Magliozzo, R. (2009). Antibiotic Resistance in Mycobacterium tuberculosis. Journal of Biological Chemistry.

127. Ozbaykal, G., Rana Atilgan, A., & Atilgan, C. (2015). In silico mutational studies of Hsp70 disclose sites with distinct functional attributes. Proteins: Structure, Function, and Bioinformatics.

128. Brown, D., Penkler, D., Sheik Amamuddy, O., Ross, C., Atilgan, A., Atilgan, C., & Tastan Bishop, Ö. (2017). MD-TASK: a software suite for analyzing molecular dynamics trajectories. Bioinformatics.

129. MD-TASK documentation — MD-TASK 1.0.1 documentation. (2019).

130. Baker, R., Cook, C., & Goodwin, D. (2004). Properties of catalase–peroxidase lacking its C-terminal domain. Biochemical and Biophysical Research Communications.

131. Atilgan, A., Akan, P., & Baysal, C. (2004). Small-World Communication of Residues and Significance for Protein Dynamics. Biophysical Journal.

132. Brown, D., Sheik Amamuddy, O., & Tastan Bishop, Ö. (2017). Structure-Based Analysis of Single Nucleotide Variants in the Renin-Angiotensinogen Complex. Global Heart.

133. Brown, D., Penkler, D., Sheik Amamuddy, O., Ross, C., Atilgan, A., Atilgan, C., & Tastan Bishop, Ö. (2017). MD-TASK: a software suite for analyzing molecular dynamics trajectories. Bioinformatics.

134. Braun, C., Mintseris, J., Gavathiotis, E., Bird, G., Gygi, S., & Walensky, L. (2010). Photo reactive Stapled BH3 Peptides to Dissect the BCL-2 Family Interactome. Chemistry & Biology.

135. Kortemme, T., Kim, D., & Baker, D. (2004). Computational Alanine Scanning of Protein-Protein Interfaces. Science Signaling.

136. Clackson, T., & Wells, J. (1995). A hot spot of binding energy in a hormone-receptor interface. Science.

137. Tina, K., Bhadra, R., & Srinivasan, N. (2007). PIC: Protein Interactions Calculator. Nucleic Acids Research.

138. Haider, S., Parkinson, G. N., & Neidle, S. (2008). Molecular dynamics and principal components analysis of human telomeric quadruplex multimers. Biophysical Journal.

139. Normal Mode Analysis (NMA) — Mode-Task documentation. https://mode-task.readthedocs.io/en/latest/theory.html#principal-component-analysis-pca

140. Ramaswamy, S. and Musser, J. (1998). Molecular genetic basis of antimicrobial agent resistance in Mycobacterium tuberculosis: 1998 update. Tubercle and Lung Disease.

141. Morlock, G., Metchock, B., Sikes, D., Crawford, J. and Cooksey, R. (2003). ethA, inhA, and katG Loci of Ethionamide-Resistant Clinical Mycobacterium tuberculosis Isolates. Antimicrobial Agents and Chemotherapy.

142. Abe, C., Kobayashi, I., Mitarai, S., Wada, M., Kawabe, Y., Takashima, T., Suzuki, K., Sng, L., Wang, S., Htay, H. and Ogata, H. (2008). Biological and Molecular Characteristics of Mycobacterium tuberculosis Clinical Isolates with Low-Level Resistance to Isoniazid in Japan. Journal of Clinical Microbiology.

143. Haas, W., Schilke, K., Brand, J., Amthor, B., Weyer, K., Fourie, P., Bretzel, G., Sticht-Groh, V. and Bremer, H. (1997). Molecular analysis of katG gene mutations in strains of Mycobacterium tuberculosis complex from Africa. Antimicrobial Agents and Chemotherapy.

144. Lipin, M., Stepanshina, V., Shemyakin, I. and Shinnick, T. (2007). Association of specific mutations in katG, rpoB, rpsL and rrs genes with spoligotypes of multidrug-resistant Mycobacterium tuberculosis isolates in Russia. Clinical Microbiology and Infection.

145. Heym, B., Alzari, P., Honore, N. and Cole, S. (1995). Missense mutations in the catalase-peroxidase gene, katG, are associated with isoniazid resistance in Mycobacterium tuberculosis. Molecular Microbiology.

146. Marttila, H., Soini, H., Huovinen, P. and Viljanen, M. (1996). KatG mutations in isoniazid-resistant Mycobacterium tuberculosis isolates recovered from Finnish patients. Antimicrobial Agents and Chemotherapy.

147. Hazbon, M., Brimacombe, M., Bobadilla del Valle, M., Cavatore, M., Guerrero, M., Varma-Basil, M., Billman-Jacobe, H., Lavender, C., Fyfe, J., Garcia-Garcia, L., Leon, C., Bose, M., Chaves, F., Murray, M., Eisenach, K., Sifuentes-Osornio, J., Cave, M., Ponce de Leon, A. and Alland, D. (2006). Population Genetics Study of Isoniazid Resistance Mutations and Evolution of Multidrug-Resistant Mycobacterium tuberculosis. Antimicrobial Agents and Chemotherapy.

148. Bolotin, S., Alexander, D., Chedore, P., Drews, S. and Jamieson, F. (2009). Molecular characterization of drug-resistant Mycobacterium tuberculosis isolates from Ontario, Canada. Journal of Antimicrobial Chemotherapy.

149. Ssengooba, W., Meehan, C., Lukoye, D., Kasule, G., Musisi, K., Joloba, M., Cobelens, F. and de Jong, B. (2016). Whole genome sequencing to complement tuberculosis drug resistance surveys in Uganda. Infection, Genetics and Evolution.

150. Vriend, G. (1990). WHAT IF: A molecular modelling and drug design program. Journal of Molecular Graphics.

151. Clamp, M., Cuff, J., Searle, S., & Barton, G. (2004). The Jalview Java alignment editor. Bioinformatics.

152. Wallace, A., Laskowski, R., & Thornton, J. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Engineering, Design and Selection.

153. Zimmermann, L., Stephens, A., Nam, S., Rau, D., Kübler, J., & Lozajic, M. et al. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of Molecular Biology.

154. Van Aalten, D., Bywater, R., Findlay, J., Hendlich, M., Hooft, R., & Vriend, G. (1996). PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. Journal of Computer-Aided Molecular Design.

155. Sousa da Silva, A., & Vranken, W. (2012). ACPYPE - Antechamber Python Parser interface. BMC Research Notes.

156. Dheda, K., & Sharma, S. (2019). What is new in the WHO consolidated guidelines on drug-resistant tuberculosis treatment? Indian Journal of Medical Research.

157. Sandgren, A., Strong, M., Muthukrishnan, P., Weiner, B., Church, G., & Murray, M. (2009). Tuberculosis Drug Resistance Mutation Database. Plos Medicine.

158. Marney, M., Metzger, R., Hecht, D., & Valafar, F. (2017). Modelling the Structural Origins of Drug Resistance to Isoniazid via key Mutations in Mycobacterium tuberculosis Catalase-Peroxidase, KatG.

159. Cade, C., Dlouhy, A., Medzihradszky, K., Salas-Castillo, S., & Ghiladi, R. (2010). Isoniazid-resistance conferring mutations inMycobacterium tuberculosisKatG: Catalase, peroxidase, and INH-NADH adduct formation activities. Protein Science.

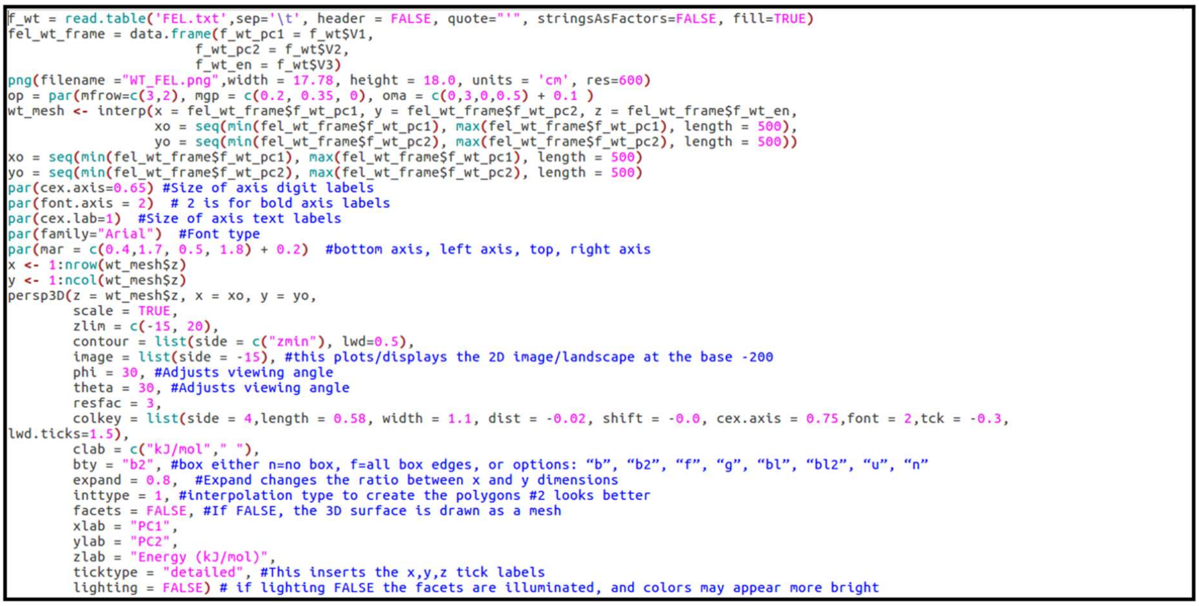
160. Penkler, D., & Tastan Bishop, Ö. (2019). Modulation of Human Hsp90α Conformational Dynamics by Allosteric Ligand Interaction at the C-Terminal Domain. Scientific Reports.

161. Kim, D., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Research.

162. Haszprunar, G. (1992). The types of homology and their significance for evolutionary biology and phylogenetics. Journal of Evolutionary Biology.

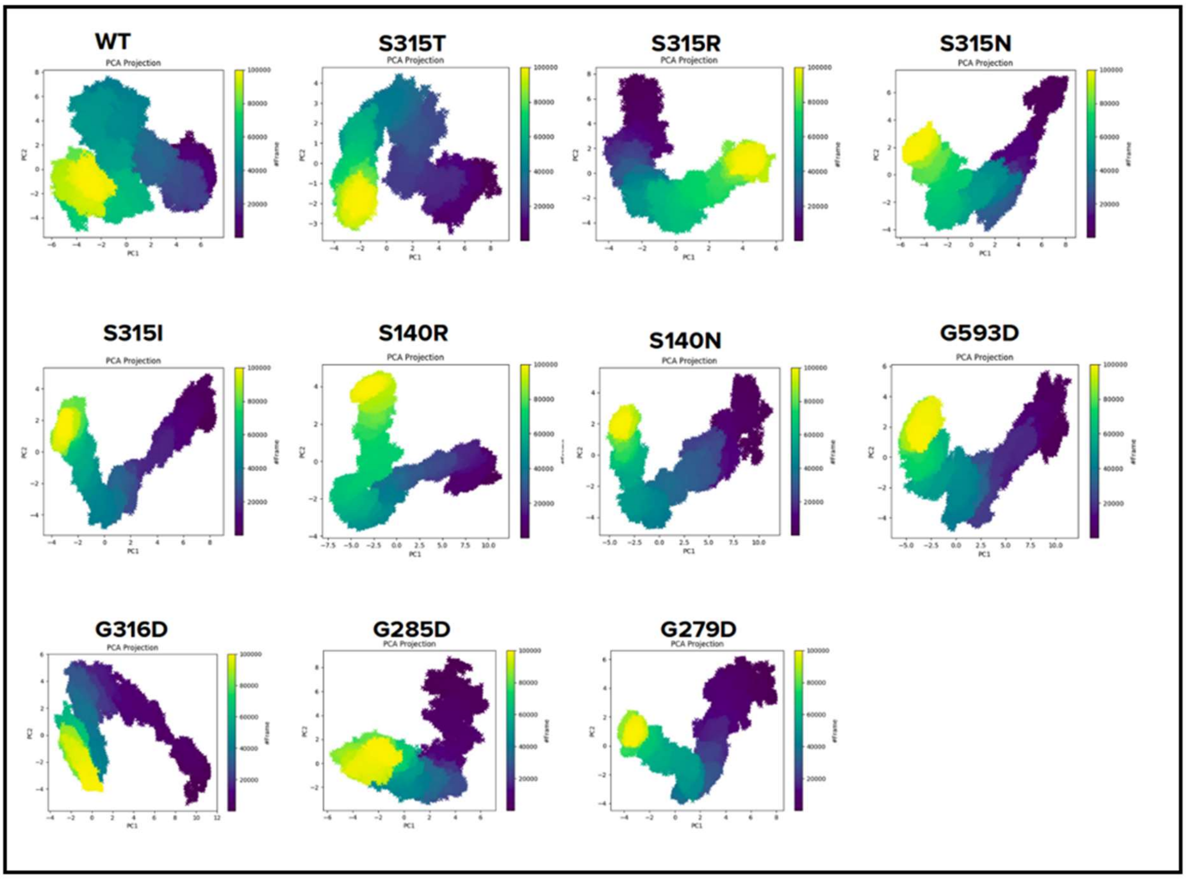
APPENDIX

Appendix 1: The R script used to plot the 3D energy heat map data from PCA.



```
f_wt = read.table('FEL.txt',sep='\t', header = FALSE, quote="'", stringsAsFactors=FALSE, fill=TRUE)
fel_wt_frame = data.frame(f_wt_pc1 = f_wt$V1,
                       f_wt_pc2 = f_wt$V2,
                       f_wt_en = f_wt$V3)
png(filename ="WT_FEL.png",width = 17.78, height = 18.0, units = 'cm', res=600)
op = par(mfrow=c(3,2), mgp = c(0.2, 0.35, 0), oma = c(0,3,0,0.5) + 0.1 )
wt_mesh <- interp(x = fel_wt_frame$f_wt_pc1, y = fel_wt_frame$f_wt_pc2, z = fel_wt_frame$f_wt_en,
                  xo = seq(min(fel_wt_frame$f_wt_pc1), max(fel_wt_frame$f_wt_pc1), length = 500),
                  yo = seq(min(fel_wt_frame$f_wt_pc2), max(fel_wt_frame$f_wt_pc2), length = 500))
xo = seq(min(fel_wt_frame$f_wt_pc1), max(fel_wt_frame$f_wt_pc1), length = 500)
yo = seq(min(fel_wt_frame$f_wt_pc2), max(fel_wt_frame$f_wt_pc2), length = 500)
par(cex.axis=0.65) #Size of axis digit labels
par(font.axis = 2)  # 2 is for bold axis labels
par(cex.lab=1)  #Size of axis text labels
par(family="Arial")  #Font type
par(mar = c(0.4,1.7, 0.5, 1.8) + 0.2)  #bottom axis, left axis, top, right axis
x <- 1:nrow(wt_mesh$z)
y <- 1:ncol(wt_mesh$z)
persp3D(z = wt_mesh$z, x = xo, y = yo,
        scale = TRUE,
        zlim = c(-15, 20),
        contour = list(side = c("zmin"), lwd=0.5),
        image = list(side = -15), #this plots/displays the 2D image/landscape at the base -200
        phi = 30, #Adjusts viewing angle
        theta = 30, #Adjusts viewing angle
        resfac = 3,
        colkey = list(side = 4,length = 0.58, width = 1.1, dist = -0.02, shift = -0.0, cex.axis = 0.75,font = 2,tck = -0.3,
lwd.ticks=1.5),
        clab = c("kJ/mol"," "),
        bty = "b2", #box either n=no box, f=all box edges, or options: “b”, “b2”, “f”, “g”, “bl”, “bl2”, “u”, “n”
        expand = 0.8,  #Expand changes the ratio between x and y dimensions
        inttype = 1, #interpolation type to create the polygons #2 looks better
        facets = FALSE, #If FALSE, the 3D surface is drawn as a mesh
        xlab = "PC1",
        ylab = "PC2",
        zlab = "Energy (kJ/mol)",
        ticktype = "detailed", #This inserts the x,y,z tick labels
        lighting = FALSE) # if lighting FALSE the facets are illuminated, and colors may appear more bright
```

Appendix 2: 2D PCA graphs for the katG wild type and variants.



Appendix 3: Table of katG residues with ≥ 2SD and ≤ -2SD Δ*L* in the katG variants.

S315N	≥2SD	L45, L48 ,H49 ,Q50 ,N51 ,R214 ,P280 ,A281 ,D282 ,L283 ,V284 ,G285 ,P286 ,E287 ,P288 ,E289 ,A290 ,A291 ,P292 ,L293 ,L298 ,G299 ,W300 ,K301 ,S302 ,S303 ,P367 ,F368
	≤ -2SD	A53 , V54 ,A55 ,D56 ,P57 ,M58 ,D207 ,E208 ,R209 ,A359 ,T363
S140R	≥2SD	L4 ,L48 ,H49 ,Q50 ,N51 ,N236 ,G237 ,G285 ,P286 ,E287 ,E289 ,A290 ,A291 ,P292 ,L293 ,E294 ,G297 ,L298 ,G299 ,W300 ,K301
	≤ -2SD	E3 , G33 ,G34 ,N35 ,Q36 ,D37 ,W38 ,W39 ,P40 ,N41 ,P57 ,P325 ,R740
S140N	≥2SD	G32, G33, G34 ,L48 ,H49 ,Q50 ,N51 ,P286 ,E287 ,P288 ,E289 ,A290 ,A291 ,P292 ,L293 ,E294 ,L298 ,G299 ,W300 ,K301 ,P367 ,F368 ,G369 ,K590 ,L598
	≤ -2SD	M2, K27 ,P40 ,N41 ,P57 ,M58 ,T363 ,P501 ,F737 ,D738 ,V739 ,R740
G593D	≥2SD	K27, H49, Q50 ,N51 ,D282 ,L283 ,V284 ,G285 ,P286 ,E289 ,A290 ,A291 ,L298 ,G299 ,W300 ,K301 ,S302 ,S303 ,G305 ,T306 ,D311 ,A312 ,I313 ,T314 ,D366 ,P367 ,F368 ,G369 ,N508 ,D511 ,K600 ,A649
	≤ -2SD	W39, P40, N41 ,A69 ,R209 ,T363 ,G560 ,N733 ,R740
S315R	≥2SD	L48, H49, Q50 ,N51 ,P52 ,A60 ,G285 ,P286 ,E287 ,P288 ,E289 ,A290 ,A291 ,P292 ,L293 ,E294 ,L298 ,G299 ,W300 ,K301 ,S303 ,P367 ,F368
	≤ -2SD	P40, N41 ,S211 ,T363 ,K600 ,G601 ,N602 ,P603 ,L643 ,G644 ,V645 ,R740
S315T	≥2SD	N44,L45,K46,L48,H49,Q50,N51,H276,G277,A281,D282,L283,V284,G285,P286,E287,P288,E289,A290,A291,P292,L298,G299,W300,K301,S303,I313,T314,P367,F368,A649,M664
	≤ -2SD	M26, K27,Y28,P29,V30,E31,G32,P40,N41,P325,P363,P401,R740
G285D	≥2SD	L45,L48,R214,V284,D285,P286,E287,P288,E289,A290,A291,P292,L298,G299,W300,K301,S302,S303,G305,P367,F368
	≤ -2SD	M26,P29,V54,P57,N236,L437,V442,F737,D738,V739,R740
G316D	≥2SD	P52,R214,E289,A290,A291,P292,G297,L298,G299,K301,D366,P367,F368,G369,N508,D509,P589,G676
	≤ -2SD	M26,K27,Y28,P29,V30,E31,G32,G33,G34,Q36,D37,W38,P40,P57,M58,L205,A281,P325, T363,R740
S315I	≥2SD	L48,H49,Q50,N51,P286,P288,E289,A290,A291,P292,L293,E294,Q295,G297,L298,G299, W300,K301,G305,G307,P367,F368

	≤ -2SD	M26,K27,Y28,P29,V30,E31,G32,G33,Q36,D37,W39,P40,T363
S457I	≥2SD	R42,L45,K46,Q50,K213,R214,P286,E287,P288,E289,A290,A291,P292,L293,E294,G297,L298,G299,W300,K301,S302,S303,D366,P367,F368,G369,G665
	≤ -2SD	M26,K27,E31,P57,D202,R209,T363,Q500,L634,F737,D738,V739,R740
G279D	≥2SD	L48,H49,Q50,N51,G285,P286,E287,P288,E289,A290,A291,P292,L293,E294,L298,G299, W300,K301,S302
	≤ -2SD	M26,K27,P57,G59,A60,A61,R740

Appendix 4: Table of katG residues with ≥ 2SD and ≤ -2SD Δ *BC* in the katG variants.

S315N	≥2SD	M26, P40, P57, H108, D117, S140, L143, R209, N218, A411, K414, R418, W438, D440, R484, D487, L488, T579, D580, S583, F584, G593, E602, K613, L616, N733, L734, 740
	≤ -2SD	L48, Q50, W107, R114, H116, G125, L132, S134, V166, Q190, P193, E195, V196, Y197, W198, V223, R254, L415, G421, P422, F483, S486, P589, L604
S140R	≥2SD	M26,P57,D117,G118,G121,R140,K143,A411,F414,R418,W438,V442,R484,G485,S486,D487,G491,P501,D580,S583,K613,L616,L734,V739,R740
	≤ -2SD	L48,Q50 ,N51 ,H116 ,M126 ,D194 ,V196 ,R253 ,N258 ,S482 ,E582 ,V586 ,L587 ,E588 ,P589 ,K590 ,D592 ,L604
S140N	≥2SD	M2, P40 ,P57 ,M105 ,G118 ,N138 ,N140 ,N218 ,A411 ,K414 ,R418 ,W438 ,Q439 ,G485 ,D487 ,K488 ,P501 ,T579 ,D580 ,S583 ,F584 ,L616 ,N733 ,L734 ,V739 ,R740
	≤ -2SD	P29,V47 ,L48 ,Q50 ,N51 ,P100 ,I115 ,H116 ,G125 ,R128 ,D194 ,V223 ,A256 ,M257 ,M420 ,G490 ,V586 ,L587 ,P589 ,K590 ,N615
G593D	≥2SD	P40 ,M105 ,Y113 ,D117 ,K143 ,N218 ,A256 ,K414 ,W438 ,R484 ,S486 ,D580 ,S583 ,D593 ,K613 ,L616 ,N733 ,L734
	≤ -2SD	K2, Q25, N26, T87, M101, P111, D169, E170, P194, A197, L202, I203, N233, R464, G465, N468, E557, V561, K565, A566, D567, K575, P578, L579, P580
S315R	≥2SD	P40 ,P57 ,D117 ,S140 ,E217 ,N218 ,T251 ,K414 ,Y418 ,W438 ,D487 ,G491 ,T579 ,D580 ,S583 ,N602 ,E607 ,Y608 ,M609 ,L616 ,L734 ,R740

	≤ -2SD	N44 ,L48 ,H49 ,Q50 ,N51 ,R114 ,H116 ,G118 ,G121 ,G123 ,G125 ,D194 ,M255 ,P288 ,A424 ,A480 ,F483 ,R484 ,S486 ,E588 ,P589 ,T618
S315T	≥2SD	M26 ,P40 ,P57 ,W91 ,G121 ,K213 ,N218 ,M257 ,D411 ,T414 ,R418 ,W438 ,V442 ,G485 ,S486 ,D487 ,K488 ,D580 ,S583 ,G593 ,N602 ,K613 ,L616 ,N733 ,L734 ,V739 ,R740
	≤ -2SD	L43 ,N44 ,L45 ,K46 ,L48 ,H49 ,N51 ,Y113 ,D117 ,G118 ,S134 ,E195 ,V196 ,N258 ,A480 ,F483 ,E582 ,A585 ,V586 ,K590
G285D	≥2SD	M26 ,P57 ,K213 ,N218 ,A411 ,K414 ,R418 ,D419 ,W438 ,V442 ,D487 ,K488 ,D580 ,S583 ,F584 ,G593 ,K613 ,L616 ,N733 ,L734 ,R740
	≤ -2SD	L43 , N44 ,V47 ,L48 ,Q50 ,T112 ,Y113 ,R114 ,I115 ,H116 ,D117 ,G118 ,P131 ,S134 ,E195 ,Y197 ,G199 ,T314 ,S482 ,F483 ,G491 ,V586 ,L587
G316D	≥2SD	M26 ,P40 ,P57 ,W90 ,M105 ,A109 ,N138 ,E217 ,N218 ,K414 ,R418 ,R484 ,G485 ,D487 ,K488 ,P501 ,S583 ,R595 ,K613 ,L616 ,N733 ,L734 ,R740
	≤ -2SD	D94, P100 ,T112 ,I115 ,H116 ,G120 ,D142 ,W198 ,L216 ,R254 ,A256 ,L298 ,F483 ,L587 ,E588 ,P589 ,A591 ,L604 ,D612 ,N637
S315I	≥2SD	M26, P40,P57 ,M105 ,A109 ,D117 ,K143 ,N218 ,A411 ,K414 ,R418 ,W438 ,G485 ,S486 ,D487 ,G491 ,D580 ,S583 ,G593 ,M609 ,K613 ,L616 ,L734
	≤ -2SD	L48, Q50 ,R114 ,G118 ,M126 ,L132 ,V188 ,Q190 ,Y197 ,G199 ,A221 ,K274 ,T435 ,A480 ,F483 ,R484 ,L604 ,P605 ,T618 ,R693
S457I	≥2SD	M26, P57 ,W90 ,K143 ,R209 ,N218 ,A411 ,K414 ,W438 ,G485 ,D487 ,P501 ,T579 ,D580 ,S583 ,G593 ,R595 ,K583 ,L616 ,V694 ,V739 ,R740
	≤ -2SD	I103, A110 ,R114 ,E195 ,V196 ,Y197 ,W198 ,R214 ,L216 ,I228 ,F252 ,M255 ,F483 ,G491 ,L587 ,Y608 ,N615
G279D	≥2SD	M26, M105 ,A109 ,N218 ,A256 ,K414 ,R418 ,W438 ,G485 ,S486 ,D487 ,K488 ,G491 ,T579 ,D580 ,S583 ,N602 ,K613 ,L616 ,N733 ,L734
	≤ -2SD	L48 ,Q50 ,T112 ,I115 ,H116 ,G120 ,G125 ,D194 ,Y197 ,W198 ,F252 ,M255 ,A424 ,L436 ,V586 ,L587 ,P589 ,D592 ,G593 ,L604 ,N615

Appendix 5: Table of destabilizing residues in the katG variant interface.

Variant	Residue	Binding energy (kcal/mol)
S140N	L43	2.07
	L45	1.07
	L48	2.21
	Y113	4.08
	R484	1.76
	D487	4.94
	R489	1.19
	Y608	1.74
	L611	1.78
	N615	1.07
S140R	L43	1.96
	L48	2.28
	Y113	1.78
	R119	1.28
	L437	1.14
	R484	2.31
	D487	1.93
	Y608	2.19
	L611	1.21
G279D	L43	2.16
	L45	1.30
	L48	1.92
	Y113	1.85
	R119	1.40
	L437	1.28
	D440	4.04
	D487	1.79
	R489	3.49
	Y608	2.23
	F737	1.39
G285D	L43	2.15
	L45	1.26
	L48	1.90
	Y113	1.94
	R119	1.33
	L437	1.31
	D440	4.16
	R484	2.24
	D487	1.63
	R489	3.61
	L611	1.86
	N615	1.64
	F737	1.41

S315I	L43	1.95
	L45	1.05
	L48	2.30
	Y113	2.85
	R119	1.19
	L437	1.23
	D440	1.11
	R484	1.66
	D487	4.35
	R489	1.21
	Y608	1.98
	L611	1.49
	D612	1.12
	F737	1.45
S315N	L43	1.88
	L45	1.06
	L48	2.18
	Y113	1.36
	R119	2.27
	L437	1.70
	D440	1.26
	R484	1.29
	S486	1.27
	D487	1.37
	R489	1.81
	Y608	1.95
	L611	1.41
	D612	1.10
	F737	1.19
S315R	L43	1.96
	L48	2.24
	Y113	1.88
	R119	1.29
	L437	1.25
	D487	2.11
	Y608	2.23
	L611	1.14
	F737	1.16
S315T	L43	1.39
	L45	1.21
	L48	2.29
	D117	2.07
	R484	1.56
	Y608	1.73
	L611	1.53

	F737	1.32
G316D	L43	1.42
	Y113	2.11
	R119	1.36
	L437	1.23
	D440	3.29
	R484	3.02
	S486	1.02
	D487	1.53
	R489	2.72
	Y608	1.84
	F737	1.07
S457I	L43	1.61
	L48	2.43
	Y113	2.56
	R114	1.08
	D117	1.44
	R119	1.78
	L437	1.21
	R484	2.07
	S486	1.27
	D487	3.65
	R496	1.50
	Y608	1.89
	L611	1.77
	D612	1.19
	F737	1.43
G593D	L43	1.96
	L45	1.05
	L48	2.39
	D117	1.44
	R119	2.50
	L437	1.16
	D440	3.47
	R484	2.18
	S486	1.35
	R489	2.76
	Y608	2.12
	L611	1.75
	F737	1.14