IDENTIFICATION OF SELECTIVE NOVEL HITS AGAINST *Mycobacterium tuberculosis* KasA POTENTIAL ALLOSTERIC SITES 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 mini-thesis)

of

RHODES UNIVERSITY, SOUTH AFRICA

Research Unit in Bioinformatics (RUBi)

DEPARTMENT OF BIOCHEMISTRY AND MICROBIOLOGY

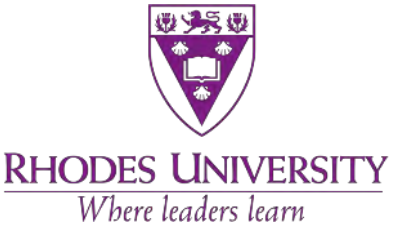
Faculty of Science

by

Fadzayi Faith Hare

G17H2028

ORCiD: 0000-0002-7763-5309



ABSTRACT

Tuberculosis (TB) is a global health threat that has led to approximately 1.5 million deaths annually. According to the World Health Organization (WHO), TB is among the top ten deadly diseases and is the leading cause of death due to a single infectious agent. The main challenge in the effective treatment and control of TB is the ongoing emergence of resistant strains of *Mycobacterium tuberculosis* (*Mtb*) which lead to multi-drug resistant (MDR) and extensive-drug resistant (XDR) TB. Hence, the identification and characterization of novel drug targets and drugs that modulate the activity of the pathogen are an urgent priority. The current situation even necessitates the reengineering or repurposing of drugs in order to achieve effective control. The β-ketoacyl-acyl carrier protein synthase I (KasA) of *Mycobacterium tuberculosis* is an essential enzyme in the mycobacterial fatty acid synthesis (FAS-II) pathway and is believed to be a promising target for drug discovery in TB. It is one of the five main proteins of the FAS-II pathway and catalyzes a key condensation reaction in the synthesis of meromycolate chains, the precursors of mycolic acids involved in cell wall formation. Although this protein has been extensively studied, little research has been devoted to the allosteric inhibition of potential drug compounds. The main aim of this research was to identify the allosteric sites on the protein that could be involved in the inhibition of substrate binding activities and novel drug compounds that bind to these sites by use of *in-silico* approaches. The bioinformatics approaches used in this study were divided into four main objectives namely identification of KasA homolog sequences, sequence analysis and protein characterization, allosteric site search and lastly virtual screening of DrugBank compounds via molecular docking. Fifteen homolog sequences were identified from the BLASTP analysis and were derived from bacteria, fungi and mammals. In order to discover important residues and regions within the KasA proteins, sequence alignment, motif analysis and phylogenetic studies were performed using *Mtb* KasA as a reference. Sequence alignment revealed conserved residues in all KasA proteins that have functional importance such as the catalytic triad residues (Cys171, His311 and His345). Motif analysis identified 18 highly conserved motifs within the KasA proteins with structural and functional roles. In addition, motifs unique to the *Mtb* KasA protein were also identified and explored for inhibitor drug design purposes. Phylogenetic analysis of the homolog sequences showed a distinct clustering of prokaryotes and eukaryotes.

A distinctive clustering was also observed for species belonging to the same genus. Since the mechanism of action of most drugs involves the active site, allosteric site search was conducted on *Mtb* KasA and the human homolog protein using a combination of pocket detection algorithms with the aim of identifying sites that could be utilized in allosteric modulator drug discovery. This was followed by the virtual screening of 2089 FDA approved DrugBank compounds against the entire protein surfaces of *Mtb* KasA and *Hsmt* KasA, performed via molecular docking using AutoDock Vina. Screening of the compounds was based on the binding energies, with more focus on identifying ligands that bound exclusively to the acyl-binding tunnel of *Mtb* KasA. This reduced the data set to 27 promising drug compounds with a relatively high binding affinity for *Mtb* KasA, however, further experiments need to be performed to validate this result. Among these compounds were DB08889, DB06755, DB09270, DB11226, DB00392, DB12278, DB08936, DB00781, DB13720 and DB00392, which displayed relatively low binding energies for *Mtb* KasA when compared to the human homolog protein.

DECLARATION

I, **Fadzayi Faith Hare**, hereby declare that this thesis submitted to Rhodes University for the purpose of the fulfilment of the one-year MSc in Bioinformatics and Computational Molecular Biology has been composed solely by myself, except where it states otherwise by referencing and acknowledgements. I also declare that it has not been submitted, in whole or in part, in any previous application for a degree.

…… 15/02/22 ……

Signature **Date**

ACKNOWLEDGEMENTS

Firstly, I would like to extend my sincere gratitude and appreciation to my main supervisor, Professor Özlem Tastan Bishop for awarding me the opportunity to work on this study and for her immense guidance and support in making this study a success. Thank you for seeing potential in me to partake this study. Your continuous encouragement and support have enabled me to accomplish the objectives of this study, for which I am truly grateful.

I would also like to extend my heartfelt appreciation to Dr. Taremekedzwa Allan Sanyanga, whose role as my co-supervisor played a great part in making this research a success. Thank you for the insights you provided and your tireless efforts throughout the course of the study. I will forever be indebted to your patience in mentoring me and your dedication towards the completion of this thesis.

I would also like to offer my invaluable gratitude to the RUBi research group, with a special mention to Victor Barozi for his guidance and assistance during the course of this study. I would also like to thank Dr. Fisayo Olotu and Afrah Khairallah for providing assistance with my research objectives.

Special thanks to my family and friends for their constant support and encouragement and above all, the Lord Almighty for giving me strength to work on this research project.

Computational resources for this research were provided by the Centre for High Performance Computing (CHPC), South Africa.

Funding acknowledgement

This work was supported through the Grand Challenges Africa programme [GCA/DD/rnd3/023]. Grand challenges of Africa is a programme 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 views expressed herein are those of the author and not necessarily those of the AAS and its partners.

DEDICATION

This thesis is dedicated to my wonderful parents, Mariane Mutamangira and Panganayi Hare. Thank you for your love, encouragement and support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY FIGURES

LIST OF SUPPLEMENTARY TABLES

LIST OF ABBREVIATIONS

2D	Two dimensional
3D	Three dimensional
ACP	Acyl carrier protein
BLAST	Basic Local Alignment Search Tool
CPU	Central Processing Unit
CHPC	Center for High Performance Computing
DNA	Deoxyribonucleic Acid
FAS	Fatty Acid Synthesis
FDA	Food and Drug Administration
HIV	Human Immuno-deficiency Virus
Hsmt	*Homo Sapiens mitochondrial*
KAS	Ketoacyl-acyl carrier protein synthase
LTBI	Latent tuberculosis infection
MDR	Multi-drug resistance
MSA	Multiple sequence alignment
Mtb	*Mycobacterium tuberculosis*
NMR	Nuclear Magnetic Resonance
RNA	Ribonucleic acid
RMSD	Root Square Mean Deviation
RR-TB	Rifampicin-resistant tuberculosis
SBDD	Structure based drug design
SVM	Support Vector Machine
TB	Tuberculosis
WHO	World Health Organization
XDR	Extensive drug resistance

LIST OF AMINO ACIDS

Single Letter Code	Three Letter Code	Amino Acid
A	ALA	Alanine
C	CYS	Cysteine
D	ASP	Asparagine
E	GLU	Glutamic acid
F	PHE	Phenylalanine
G	GLY	Glycine
H	HIS	Histidine
I	ILE	Isoleucine
K	LYS	Lysine
L	LEU	Leucine
M	MET	Methionine
N	ASN	Asparagine
P	PRO	Proline
Q	GLN	Glutamine
R	ARG	Arginine
S	SER	Serine
T	THR	Threonine
V	VAL	Valine
W	TRP	Tryptophan
Y	TYR	Tyrosine

LIST OF WEBSERVERS

Webserver	Link
BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
CAVITYPLUS	http://www.pkumdl.cn/cavityplus/
CLUSTAL OMEGA	https://www.ebi.ac.uk/Tools/msa/clustalo/
MAFFT	https://www.ebi.ac.uk/Tools/msa/mafft/
MAST	https://meme-suite.org/meme/doc/mast.html
MEME	https://meme-suite.org/meme/tools/meme
PROTEIN PLUS DOGSITESCORER	http://dogsite.zbh.uni-hamburg.de
T-COFFEE	http://tcoffee.crg.cat/tcs

THESIS OVERVIEW

The purpose of this research was to use *in silico* (computer-based) approaches to virtually screen DrugBank compounds to evaluate their potential as inhibitory compounds against *Mycobacterium tuberculosis* (*Mtb*) β-ketoacyl-acyl carrier protein synthase I (KasA) for drug discovery and design. This thesis consists of a total of five chapters that detail the approaches used in this study.

CHAPTER ONE

This chapter sets forth an introduction into the Tuberculosis disease (TB), one of the leading causes of morbidity and mortality worldwide. The main focus of this chapter is the *Mtb* KasA protein, which plays an essential role in the mycolic acid biosynthesis pathway. The structure and mechanism of action of the protein are detailed in this chapter.

CHAPTER TWO

This chapter details sequence and structural analysis of the *Mtb* KasA protein and its homologs, in order to identify conserved regions that are structurally and functionally important in the protein family. Different *in silico* approaches were used to analyze the similarities and differences in sequence, structure, and evolution.

CHAPTER THREE

This chapter introduces allosteric site search on the *Mtb* KasA and human homolog protein for allosteric modulation. A number of allosteric site search tools with different algorithms were used in the identification of potential allosteric sites. A consensus was drawn from the sites identified and these were further explored in drug compound screening.

CHAPTER FOUR

Chapter 4 describes the use of AutoDock v4.2, a molecular docking software that assists in the analysis of the conformation and orientation of the ligand molecule into the binding sites of the target protein. The chapter details the virtual screening of the identified drug compounds via blind docking to identify potential *Mtb* KasA inhibitors.

CHAPTER FIVE

In this chapter, a summary of the findings in the thesis are reported and the possible future aspects are presented.

CHAPTER ONE

LITERATURE REVIEW

1.1 INTRODUCTION

Tuberculosis (TB) continues to reign as one of the leading causes of morbidity and mortality on a global scale. It is an airborne, communicable disease caused by the bacillus *Mycobacterium tuberculosis* (*Mtb*) [1]. Infection with *Mtb* can develop from a dormant state (latent TB), in which the host retains viable *Mtb* within their lungs and is asymptomatic, and in some cases can progress into a contagious state (active TB) in which a person exhibits clinical manifestations [2]. The disease typically affects the lungs (pulmonary TB) but can also affect other parts of the body such as the brain, kidneys, intestines as well as the spine. Clinical features associated with active pulmonary TB include severe weight loss, fever, night sweats, fatigue, chronic cough as well as chest pains [3].

1.2 PREVALENCE OF TB

1.2.1 Global and National TB implications

Approximately one-third of the world's population is estimated to be infected with *Mtb* [4], with about 5-10% of the infected persons at a risk of developing active TB in their lifetime [5]. According to the World Health Organization [6], an estimated 10 million (range, 8.9 – 11.0 million) people contracted TB in 2019 worldwide. Among these 10 million individuals, 56% of the cases were reported from men, 32% from women and children accounted for 12%. A majority of the approximated number of cases recorded in 2019 were in the WHO regions of South-East Asia, Africa and the Western Pacific, with a smaller proportion occurring in the Eastern Mediterranean Region, the Americas and Europe [7]. About 87% of the universal cases recorded in 2019 were reported in 30 high TB burden countries [8], and eight of these countries account for two-thirds of the total. These include India (26%), Indonesia (8.5%), China (8.4%), the Philippines (6.0%), Pakistan (5.7%), Nigeria (4.4%), Bangladesh (3.6%) and lastly South Africa (3.6%) [9].

Drug resistance in TB treatment is a public health dilemma and health security threat. Globally, approximately half a million people developed rifampicin-resistant TB (RR-TB) [10], with 78% of the population having developed multi-drug resistant TB (MDR-TB). The countries that contributed the largest to this global burden in 2019 were India (27%), China (14%) and the Russian Federation (8%). Figure 1.1 shows the estimated TB incidence rates as reported by the World Health

Organization in 2019. The annual number of incident TB cases relative to the population size varied greatly among the different countries. A low incidence of TB (< 10 cases per 100 000 population per year) was noted mostly in the American, European, Eastern Mediterranean and the Western Pacific regions [6]. On the other hand, more than 500 cases were discovered in the Central African Republic, the Democratic People's Republic of Korea, the Philippines as well as South Africa.

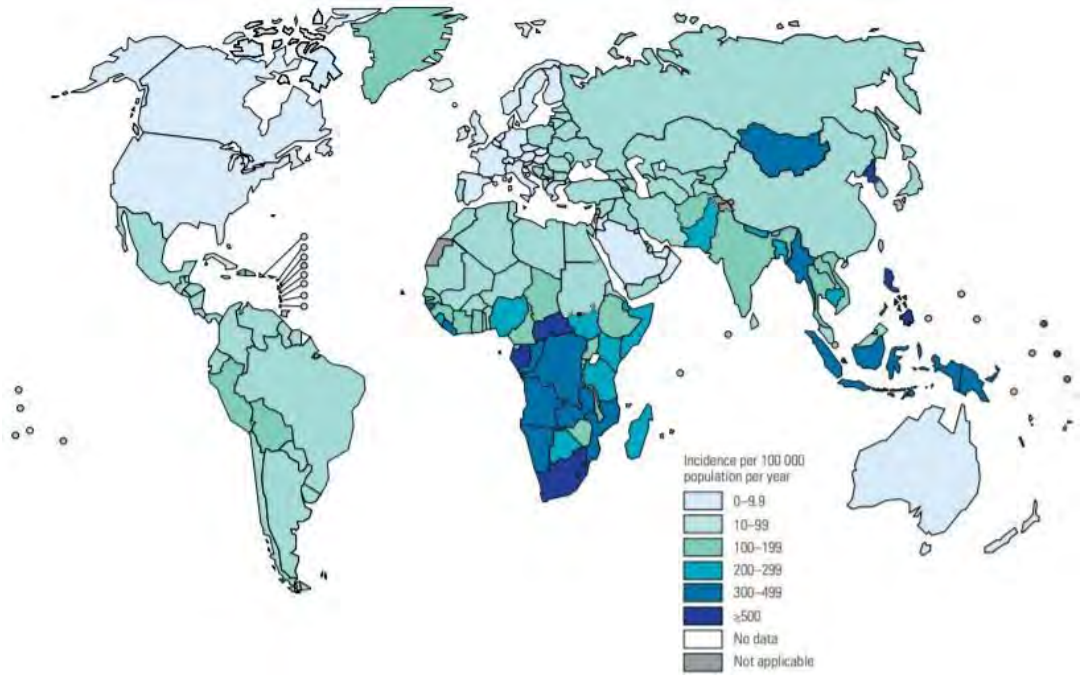


Figure 1.1: Estimated TB incidence rates, 2019 (Adapted from WHO Global TB Report 2020).

There has been a significant decline in the annual number of TB-related deaths globally over the years. In 2019, a total of 1.4 million people succumbed to TB, and this included 208 000 deaths reported among people with HIV [11]. This is in comparison with 1.7 million deaths reported in 2000, with 678 000 of these cases recorded among HIV-positive individuals [12]. In South Africa, an estimated incidence of 360 000 cases of active TB were reported in 2019, with 58% of the population being HIV-positive. The high rates of TB have been observed from the early 1990s, largely owing to the HIV epidemic that negatively impacted the control of TB in the country. It is believed that about 80% of the population of South Africa is infected with *Mtb*, with a vast majority with latent TB rather than the active TB disease.

The first national TB prevalence survey in South Africa was conducted by a collaboration of the South African National Department of Health TB Program, the Human Sciences National Council (HSRC) and the National Institute for Communicable Diseases (NICD), between 2017 and 2019 [13]. This survey was performed on a sample size of 55 000 people aged 15 years and older in 110 randomly determined clusters, with a total of 35 000 participants. The results of this survey revealed that TB is more common among males than females, reaching its peak in individuals aged 35 – 44 years and those above 65 years [14]. This study also revealed that over two-thirds of the HIV-negative patients showing classic symptoms of TB had not sought treatment. A greater percentage of HIV-negative persons were asymptomatic compared to HIV-positive individuals. This discovery shows that there is need to encourage people of all age groups to test for TB as previous studies have only focused on symptomatic and HIV-positive persons.

In an attempt to exterminate the burden of TB on humanity, the 67th World Health assembly adopted a policy known as the End TB Strategy, which targets to end the global TB outbreak by 2035 [15]. This strategy provides a comprehensive approach to overcome the issues and challenges associated with the epidemic by aiming to reduce TB-related deaths by 95% and scale down new cases of TB by 90% between 2015 and 2035 [16]. Over the decades, TB medications have been used to combat the burden of this infectious disease. However, strains that are resistant to one or more of the medications have been documented and this has negatively impacted the current efforts to eradicate TB. While TB is present in every country, the majority of TB sufferers reside in low-and middle-income countries, thus it is regarded as a disease of poverty [17]. Other factors leading to high prevalence of TB include the difficulty in accessing healthcare and TB tests, failure of patients to complete the recommended treatment regimens, inadequate contact tracing as well as high drug abuse.

Despite free treatment services of TB in public clinics, the control of TB in South Africa is still a challenge and patients incur considerable direct and indirect costs prior to starting treatment. According to joint research conducted between South Africa and Britain in 2015, it was found that the average cost of a newly diagnosed TB case is approximately US$ 210 (equivalent to R2500), whilst an MDR-TB case costs up to US$ 9500 (equivalent to R115 000) [18]. However, the cost of treatment is dependent upon a number of factors which include socioeconomic status, health systems structure, TB service delivery model, insurance coverage and hospitalization costs [12, 19]. It has been estimated that around R5 billion annually is needed to reach the target to eliminate TB by 90%, in addition to the current expenditure. This poses a very high economic burden and hence

the need to develop novel therapeutic ways of combating TB infection and managing treatment.

1.3 PATHOGENESIS OF TB

Tuberculosis is an ancient disease, whose causative agent was discovered by the German physician, Robert Koch in 1882 [20]. This disease, formerly known as consumption, can present itself in many forms including skeletal deformities (Potts disease) that have been found in remains of ancient Egyptians. A greater proportion of TB cases are attributed to *Mtb* whilst other closely related organisms that make up the *Mycobacterium tuberculosis* complex (MTC) such as *Mycobacterium africanum*, *Mycobacterium bovis* and *Mycobacterium caprae* only contribute to a smaller number of cases recorded. Tuberculosis may occur in 3 stages namely primary, latent and active infection. Primary infection by *Mtb* is acquired through inhalation of contagious aerosol particles (droplet nuclei). Mycobacteria-loaded droplet nuclei are released when a patient with active pulmonary TB coughs or sneezes, and they remain suspended in room air currents for hours which increases the chances of transmission. The tubercle bacilli develop infection as a result of the deposition of the droplet nuclei (1 – 5 μm in diameter) in the terminal airspaces of the lungs. The bacteria multiplies within the alveoli sacs for a period of 2 – 12 weeks until they reach a number that is adequate to evoke an immune response in the host. The infection that develops as a result of *Mtb* infection can lead to various outcomes: (1) clearance of the organism by the host's immune system, (2) suppression into an inactive form called latent TB, (3) development of progressive TB disease and (4) reactivation of the disease several years later.

The vital cells involved in the innate defense against *Mtb* are the alveolar macrophages and the dendritic cells [21]. In the event that the host's defense system fails to eliminate the infection, the bacteria multiplies within the alveolar macrophages and eventually leads to cell death. The survival of infected macrophages results in the production of cytokines and chemokines that attract monocytes, other alveolar macrophages and neutrophils to form a granulomatous structure known as a tubercle [22]. When replication of the tubercle is uncontrolled, the structure enlarges, and the bacilli eventually lead into the lymph nodes and cause lymphadenopathy. This is a characteristic clinical representation of primary TB.

Latent TB infection occurs mostly after a primary infection. This is a result of an effective immune response in the lungs that results in the successful inhibition in the growth of the bacteria, leading to the bacteria becoming dormant. This suppression of bacterial replication usually occurs before signs and symptoms develop and infected individuals are not capable of transmitting it to others

[3, 24, 25]. Tubercle bacilli can survive in the host for years and the microbial virulence condition whether the infection remains dormant, resolves without medical intervention or develops into the active TB infection. TB is known to affect any part of the body, however, 85% of the patients present with a pulmonary infection. Extrapulmonary TB occurs outside the lungs, as a result of hematogenous dissemination. The most common sites of extrapulmonary TB disease include mediastinal lymph nodes, vertebral bodies, adrenals, meninges and the gastrointestinal tract.

Reactivation TB, also known as active TB, results from a proliferation of previously dormant seeded bacteria from the time of the primary infection. There is a 5-10% lifetime risk of developing active TB disease in healthy people with no underlying conditions, although this can be affected by age and other risk factors. It is believed that immunosuppression is linked to reactivation TB as the secreted protein resembling resuscitation-promoting factor (Rpf) is activated in this state and the bacteria is reverted from a suppressed state to an active state [25]. Clinical manifestation of TB begins when it switches from a latent to an active state, and this is dependent on the proliferation rate of the bacteria. Any organ can become a site for reactivation, however, in most cases, reactivation occurs in the lung apices due to the favorable conditions such as high oxygen tension. Conditions that impair immune function facilitate reactivation TB. These include diabetes, HIV infection, gastrectomy, cancer as well as chronic kidney disease. Reactivation also occurs as a result of a reinfection rather than an activation of latency in some patients.

1.4 SCREENING AND DIAGNOSIS OF TB

An essential step in ensuring the effective treatment and cure of TB is rapid and accurate testing. Although there is no gold standard when it comes to TB diagnosis, various tools and technologies have been used to diagnose TB accurately over the years. Tuberculosis should be suspected when a patient presents with at least one of the four symptoms: (1) cough lasting longer than two weeks, (2) long-lasting fever, (3) night sweats and (4) weight loss. In addition to the above-mentioned criteria, epidemiological characteristics such as history of contact with a person with active pulmonary disease as well as other risk factors for TB reactivation ought to be considered. The diagnosis of TB is confirmed by the presence of the causative agent, *Mtb* in the biological specimen. The choice of a diagnostic test is dependent upon whether it is used to detect latent tuberculosis infection (LTBI), active TB disease or drug resistance [26].

The primary screening methods for LTBI are the Mantoux tuberculosis skin test (TST) and the Interferon gamma-release blood assays (IGRA). The TST was developed by Robert Koch more

than a 100 years ago and is also known as the 'old tuberculin' test. It is performed using the Mantoux technique, which involves injecting intradermal purified protein derivative (PPD) of 5 tuberculin units (5 TU) into the forearm of the patient. A pre-test of the probability of an *Mtb* infection is indicated by an induration of about 15 mm, after 48 – 72 hours of taking the test. This test is based on a delayed-type hypersensitivity (DTH) to reactivation of the skin to PPD. In spite of the fact that the TST has many advantages, including low reagent and equipment costs as well limited skill requirements, it has two serious limitations. Firstly, the specificity of the test is affected by late or repeated Bacille Calmette-Guerin (BCG) vaccination boosters in children, and to a lesser extent, subjection to non-tuberculosis mycobacteria species [27, 28]. Furthermore, it has a restricted prediction value as it does not usually distinguish latent and active TB and most patients with a positive TST result do not usually develop active TB.

In an attempt to improve TB screening and replace TSTs, the IGRAs were introduced about two decades ago. The IGRA is a blood test that is dependent on the release of interferon-gamma by white blood cells exposure to antigens specific to TB in-vitro and is used in the diagnosis of *Mtb*. The Food and Drug Administration (FDA) approved IGRAs are the QuantiFERON®-TB Gold-In Tube test (QFT-GIT) and the T-SPOT®.TB test (T-Spot). The benefits of this test are that results are available within 24 hours, it requires a single patient visit, and a prior BCG vaccination does not cause a false-positive IGRA result. The limitations of this test are that the blood samples need to be processed within 8-30 hours when the white blood cells are still viable and like the TST test, it fails to accurately distinguish the type of TB infection.

In order to detect active TB disease, the following technologies are used: imaging approaches (Chest X-ray and PET-CT), microscopy (sputum smears), culture-based methods and molecular tests. A chest radiography is used to search for a multinodular infiltrate above or behind the clavicle as this is the most predictable feature of active TB. Since X-rays lack specificity, it is imperative that an abnormal chest X-ray is followed up by microbiological tests. Sputum smear microscopy continues to be the widely used diagnostic test despite its numerous limitations. The first step normally involves microscopic examination to test for acid-fast bacilli (AFB). When a positive test is detected, a nucleic acid amplification test (NAAT) is used to further confirm the diagnosis. The two types of NAATs available are the Xpert MTB/RIF and the Line probe assay. Xpert MTB/RIF identifies *Mtb* DNA in sputum sample and detects resistance to rifampicin whilst the Line probe assay detects *Mtb* resistance to rifampicin and isoniazid.

1.5 TB TREATMENT AND MANAGEMENT

Drug therapy in TB was initiated in the 1940s with the use of an antimicrobial agent, streptomycin. However, due to the rapid development in drug resistance and elevated failure rates, there was a need to implement combination therapy by using at least two potent drugs to combat drug resistance [28]. This led to the use of streptomycin and para-aminosalicylic acid in the early 1950s. Due to increased bacterial resistance and mortality, isoniazid was introduced as a third agent to the existing treatment regimen, thus creating an initial treatment regimen consisting of isoniazid, streptomycin and para-aminosalicylic acid, followed by the administration of isoniazid and para-aminosalicylic acid in successive months [29]. The administration of this regimen was for a total of 18 months. In the mid-1970s, TB treatment took an astounding turn when a short treatment regimen of 6 months was introduced, consisting of isoniazid (H), rifampicin (R) and pyrazinamide (Z) [30]. In 2009, ethambutol (E) was added to a fixed dose regimen of RHZ, due to increased resistance to isoniazid. This new regimen was then divided into two phases; an initiation phase of 2 months followed by a continuation phase of 6 months. Currently, patients with monoresistant TB and polyresistant TB are treated to a 2-month administration of RHEZ, followed by a 4-month administration of RH.

Drugs used to treat Tuberculosis are divided into 5 groups in accordance with the WHO guidelines. The first group contains first-line anti-tuberculosis drugs whilst subsequent groups consist of second line agents. Group 1 contains isoniazid, rifampicin/rifabutin and pyrazinamide as core drugs whilst ethambutol is added as a companion drug. Streptomycin is not commonly used. If rifampicin resistance is detected, rifabutin is added to the regimen [31]. Treatment of patients with MDR-TB and XDR-TB involves the use of second line agents. Group 2 includes parenteral anti- TB drugs such as amikacin, kanamycin, streptomycin and capreomycin. These drugs have bactericidal activity only compared to Group 3 agents which have both bactericidal and sterilizing activity, and hence their safety profile is low. Compounds in this category are structurally similar therefore only one compound can be used. However, streptomycin is not recommended for treatment of MDR/XDR-TB even in the event that a drug susceptible isolate is found because the test is not always accurate [32]. Capreomycin is usually the first choice of drug in treatment, followed by kanamycin and lastly amikacin. Fluoroquinolones are broad-spectrum antibacterial agents that are used to treat TB and make up Group 3. Drugs in this group include levofloxacin, ciprofloxacin, ofloxacin, gatifloxacin and moxifloxacin. Properties such as excellent oral bioavailability, bactericidal activity as well as lack of cross resistance make fluoroquinolones favorable and safe to administer. Levofloxacin has the first preference in this group, succeeded by moxifloxacin,

gatifloxacin and ofloxacin. Ciprofloxacin is the least effective anti-TB agent in this group therefore it is avoided.

The fourth group includes drug compounds from three classes of drugs: thioamides (ethionamide and protionamide), cycloserine/terizidone and an aminosalicylic acid. Since drugs in this group have different mechanisms of action, the use of more than one drug is warranted. Thioamides have first preference over cycloserine and aminosalicylic acid because of the bactericidal activity as well as low toxicity profile. Cycloserine and aminosalicylic acid are added last and this order of administration is based on the effectiveness, side effects and cost. Group 5 drugs are very diverse and consists of agents with little known clinical evidence in humans and high toxicity. These drugs are considered as adjuvant agents and are counted as only half of one of the four basic drugs used in the treatment of MDR/XDR-TB [33, 34]. Drugs in this category include linezolid, clofazimine, thioacetazone, co-amoxiclav, clarithromycin, delamanid, imipenem, bedaquiline and a high-dose isoniazid. These drugs are only used when all other drugs have failed. Table 1.1 gives a summary of the drugs used to treat Tuberculosis according to the WHO guidelines and recommendations.

Table 1.1: Summary of classifications of anti-TB drugs according to the WHO guidelines.

Groups	Drugs
1	Ethambutol, Isoniazid, Pyrazinamide, Rifampicin
2	Amikacin, Capreomycin, Kanamycin, Streptomycin
3	Gatifloxacin, Levofloxacin, Motifloxacin, Ofloxacin
4	Cycloserine, Ethionamide, p-amino salicylic acid, Thioacetone
5	Bedaquiline, Clarithromycin, Clofazimine, Co-amoxiclav, Delamanid, Imipenem/cilastatin, Isoniazid (high dose), Linezolid

1.6 TB DRUG RESISTANCE

1.6.1 Intrinsic Resistance in *Mycobacterium tuberculosis*

Various mechanisms lead to the development of intrinsic drug resistance in bacteria. These include the presence of a thick, waxy, hydrophobic cell wall, activation of the efflux pump on the surface of the bacterial cells, modification of the drug target, inactivation of a drug by use of enzymes as well as reducing drug uptake. The cell wall of *Mtb* is thicker and more hydrophobic compared to other gram-positive bacteria. It consists of three key structural components: (1) a network of peptidoglycan (PG), (2) the arabinogalactan (AG) polysaccharide, and (3) long chain mycolic acids (MA) [34]. The peptidoglycan consists of *N*-glycolylmuramic acid rather than the usual *N*-acetylmuramic acid found in most bacterial cells. The lipid nature of the cell wall renders it hydrophobic, thus preventing the permeation of hydrophilic compounds. However, it is believed that small hydrophilic compounds active against *Mtb* can traverse the cell wall via water- filled porins. The diffusion of many hydrophobic antibiotics such as rifampicin, tetracyclines and fluoroquinolones is dependent on the molecule hydrophobicity; the more hydrophobic the molecule, the more readily it passes through the cell wall [35].

Efflux pumps are important components of bacterial and eukaryotic systems. The genome of *Mtb* encodes various putative efflux systems which are divided into five superfamilies, based on sequence homology. These include the ATP-binding cassette (ABC), the major facilitator superfamily (MFS), the multi-drug and toxic compound extrusion superfamily (MATE), the resistance-nodulation-cell-division superfamily (RND), and the small multi-resistance superfamily (SMR) [36]. Approximately 2.5% of the *Mtb* genome consists of genes encoding the ABC superfamily transporters. The increased transcription of Rv0194, Rv1819c (BacA) and Rv2936/Rv2937/Rv2938 (DrrABC), has led to increased multidrug resistance involving the extrusion of substrates such as chloramphenicol, macrolides and tetracyclines. According to a study conducted by Li et al., 19 MFS-type transporters encoded in the *Mtb* genome are correlated with drug resistance [37]. The MFS pump Rv1258c (Tap) confers resistance to a vast array of substrates such as clofazimine, ethambutol, erythromycin, ethidium bromide, fluoroquinolones, isoniazid, rifampicin and tetracyclines. The upregulation of the RND transporter, MmpL5 and its periplasmic accessory protein MmpS5 has been linked to increased resistance to bedaquiline and clofazimine. In clinical strains of *Mtb*, overexpression of efflux pumps has been attributed to antibiotic stress. Therefore, understanding the mechanisms controlling the overexpression is essential to the search and design of novel therapeutics to combat drug resistance [36].

The most common mechanism of resistance in *Mtb* is drug target alteration. Modifications in the drug-drug target interaction sites can inhibit successful binding of the drug and therefore confer resistance [38]. Non-synonymous mutations in drug target encoding genes and nucleotide substitutions in the operon encoding the ribosomal RNA confer resistance in *Mtb*. For instance, mutations in the active sites of DNA-dependent RNA polymerase confer resistance to rifampicin by lowering the affinity of the drug for the target. Drug target modification can also occur as a result of enzymatic action, in which the bacteria express resistance to antibiotics by producing enzymes that prevent the binding of the drugs.

Overexpression of the drug target may overthrow inhibition due to an abundance of the target. Drug target overexpression leads to low level resistance, as is the case with isoniazid and cyclosporine. This can be overcome by increasing the frequency of dosing of the drugs. However, an increase in drug dose can result in severe adverse effects and this contradicts overcoming resistance due to drug target overexpression.

1.6.2 Acquired drug resistance in *Mycobacterium tuberculosis*

The World Health Organization has categorized drug resistance in TB into 5 main classes namely monoresistance, polydrug resistance, multi-drug resistance, rifampicin resistance and extensive drug resistance [6]. Mono-resistant TB arises as a result of *Mtb* strains that are resistant to only one first line anti-TB drug. Poly-resistant TB refers to resistance to at least one of the first line antitubercular agents, but not to both isoniazid and rifampicin. When *Mtb* strains become resistant to at least two of the potent frontline TB drugs, isoniazid and rifampicin, this is called multidrug resistant TB. Phenotypic and genotypic methods can be used to detect rifampin resistance with or without resistance to other anti-TB drugs and this is termed rifampicin resistance [39]. *Mtb* strains that are resistant to isoniazid and rifampicin, with the addition of any fluoroquinolone and any one of the three second-line injectables (i.e., amikacin, kanamycin and capreomycin) gives rise to extensive drug resistance. In recent studies, it has been discovered that there is a new case of *Mtb* strains that are resistant to all antibiotics that have been tried and tested, and this is known as totally drug resistant TB (TDR) [40].

Two types of drug resistance are associated with *Mtb*, namely phenotypic and genetic resistance. Genetic drug resistance arises as a result of chromosomal mutations in proliferating bacteria whilst phenotypic resistance is due to changes in gene expression that leads to tolerance to drugs in slowly

growing bacteria [41]. Drug resistance can happen in two ways, i.e., primary and secondary resistance. Primary drug resistance occurs when a person who has no history of first-line anti-TB treatment develops MDR-TB, usually via exposure to an already resistant drug strain whereas the latter, also known as acquired drug resistance, develops as a result of poor adherence to treatment, drug malabsorption as well as insufficient regimens [43, 44]. Most cases of MDR-TB and XDR-TB develop as a result of acquired resistance, although a portion of some cases have been caused by nosocomial infections.

Despite the advances in the control of TB in South Africa and worldwide, rifampicin-resistant and multidrug resistant cases of TB are still on the rise. According to the WHO, 558 000 cases of RR-TB and MDR-TB were recorded globally in 2017 [15]. In spite of this large number, only 25% of the cases were detected and started on treatment. The End TB strategy focuses on the diagnosis and treatment of any susceptible forms of drug resistant TB. Universal access to drug susceptibility testing (DST) is one of the recommended guidelines used in the initiation and successful delivery of treatment. Table 1.2 summarizes the existing anti-tuberculosis agents, genes known to infer resistance as well as their mechanism of action.

Table 1.2: Common genes involved in acquired resistance in *Mycobacterium tuberculosis* and associated anti-TB drugs. Modified from Dookie *et al*. 2018 [44].

Drug involved	Resistance Genes	Gene function	Mechanism of action
Isoniazid	*katG* *inhA* *kasA*	Catalase peroxide Enoyl ACP reductase β-keto acyl ACP synthase	Inhibition of mycolic acid biosynthesis
Rifampicin	*rpoB*	B-subunit of RNA polymerase	Inhibition of RNA synthesis
Ethambutol	*embCAB* *embR*	Arabinosyltransferases *embCAB* transcription regulator	Inhibition of cell wall arabinogalactan synthesis
Pyrazinamide	*pncA* *rpsA*	Pyrazinamidase S1 ribosomal protein	Inhibition of trans-translation, inhibition of pantothenate and coenzyme A synthesis
Streptomycin	*rpsL* *rrs* *gidB*	S12 ribosomal protein 16S rRNA 16S rRNA methyltransferase	Inhibition of protein synthesis translation.
Amikacin, Kanamycin and Capreomycin	*rrs* *eis* *tylA*	16S rRNA Acetyltransferase rRNA methyltransferase	Inhibition of protein synthesis translation
Ethionamide	*ethA* *inhA* *mshA* *ndh*	Flavin monooxygenase Enoyl-ACP reductase Glycosyltransferase	Inhibition of cell wall mycolic acid synthesis
Para-amino salicylic acid	*ethR* *thyA* *folC* *ribD*	NADH dehydrogenase II *ethA* transcription repressor Thymidylate synthase A	Inhibition of folic acid and thymine nucleotide metabolism
Fluoroquinolones	*gyrA* *gyrB*	DNA gyrase subunit A DNA gyrase subunit B	Inhibition of DNA synthesis
Bedaquiline/ Clofazimine	*pepQ*	Hydrolase	Inhibition of mycobacterial ATP synthesis
Linezolid	*rplC* *rrl*	Ribonucleoprotein 23S rRNA	Inhibition of protein synthesis

The emergence of drug resistant tuberculosis has aggravated the TB public health burden and in order to curb the issue of resistance, new drugs and drug targets are constantly being discovered. Three primary approaches have been used been used in the development of therapeutic agents and

these include: (1) modification of the existing agent classes, (2) inference with resistance mechanisms and (3) searching for agents with novel mechanisms of action. The antibiotic development pipeline that began in the early 1940s reveals that all the drug compounds that are currently used to treat TB are either chemical modifications of existing structures or repurposed drugs. The model of mycobacterial cell wall structure which has emerged from studies conducted by Brennan Patrick et al., revealed that the fatty acid synthesis pathway presents several potential drug targets [45]. This existing pathway has been explored for potential drug targets over the years and is detailed in the section below.

1.7 FATTY ACID SYNTHESIS PATHWAY AND KAS ENZYMES

Fatty acid synthesis (FAS) is a critical anabolic pathway essential for the formation of membranes and viability of cells. There are two basic types of FAS pathways present in *M. tuberculosis*, namely the eukaryotic, multifunctional FAS I and the prokaryotic and discrete FAS II. FAS I is responsible for the *de novo* synthesis of C_{16} - C_{24} fatty acid chains, which are then passed to the FAS II pathway for elongation and formation of mycolic acids [46]. The multifunctional protein domains of the FAS I pathway include acyltransferase, enoyl reductase, dehydratase, malonyl/palmitoyl transferase, acyl carrier protein, ketoacyl reductase and ketoacyl synthase. FAS II of *Mtb* includes β-keto-acyl carrier protein synthase I (KasA), β-keto-acyl carrier protein synthase II (KasB), β-keto-acyl ACP reductase (MabA), β-hydroxyacyl ACP dehydrase (HadABC) and enoyl-ACP reductase (InhA). FAS II is a dissociated system, whereby protein is encoded for by a separate gene and catalyzes one step in the pathway. This pathway has also been partially characterized in eukaryotic mitochondria. The FAS I complex of mycobacteria is an essential enzymatic complex, which catalyzes the de novo synthesis of C_{16} and C_{18} acyl-CoAs from acetyl-CoA using malonyl-CoA [47]. The acyl-CoAs can be used in the synthesis of membrane phospholipids or can be further elongated by the FAS I pathway into C_{24} and C_{26} fatty acid chains. FAS I is a single polypeptide that usually produces only palmitate, whereas FAS II has a diverse number of intermediates, owing to the presence of the acyl carrier protein (ACP) intermediates, which can be diverted into other biosynthetic pathways.

The elongation of the fatty acids in the FAS II pathway is initiated by *fadD,* a gene that encodes malonyl-CoA:AcpM that catalyzes the formation of malonyl-AcpM, the two-carbon substrate which forms the basis of the synthesis of mycolic acids. FabD catalyzes the transacylation of malonate from malonyl-CoA to phosphopantothenylated *holo*-AcpM [48]. AcpM shuttles the

growing acyl chains between the discrete monofunctional enzymes that catalyze single-step reactions in the FAS II pathway. The elongation process is carried out by β-keto-acyl-AcpM synthase enzymes encoded by *kasA* (Rv2245) and *kasB* (Rv2246). KasA and KasB catalyze a condensation reaction of acyl-AcpM and malonyl-AcpM, which results in the elongation of the growing meromycolic acid chain by two carbons. Meromycolic acids are the precursors of the final product, mycolic acids. When condensation is complete, the β-keto-acyl-AcpM product undergoes a series of reactions catalyzed by β-keto-acyl-AcpM reductase (MabA), β-hydroxyacyl- AcpM dehydratase complex (HadABC) and the enoyl-AcpM reductase (InhA). Figure 1.2 shows the steps involved in the biosynthesis of mycolic acids in *Mtb*.

KasA and KasB are members of the thiolase superfamily, established on the three-dimensional fold characterized from *Saccharomyces cerevisiae* [49]. These enzymes are known to catalyze Claisen condensation reactions in the FAS II pathway. KasA and KasB are both specific to the elongation of acyl-AcpM primers. The crystal structures of these enzymes have been solved and polyethylene glycol was used to mimic a long (C_{40}) fatty acid chain to characterize the hydrophobic acyl-binding tunnel of KasA, in order to accommodate the growing fatty acid chains [50]. KasA and KasB have similar roles, however, KasA has been found to be involved in the initial elongation of mycolate chains whereas KasB is involved in the full-length extension, to give rise to meromycolic acids (up to C_{56} chains). The *Mtb* KasA enzyme is as such a promising drug target which plays a role in the mycobacterial cell wall development and has contributed to the resistance of this pathogen against antibiotics. The structure and mechanism of action of this enzyme are detailed below.

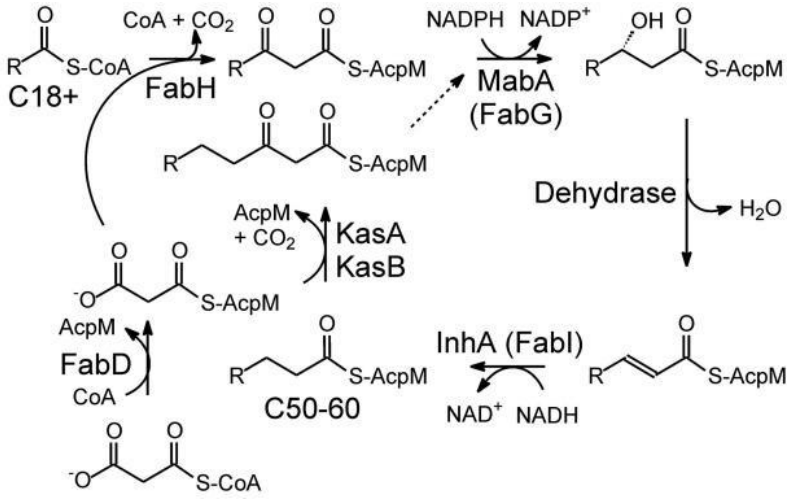


Figure 1.2: Schematic diagram showing the intermediates and products involved in the synthesis of fatty acids by the FAS II pathway. Adapted from Luckner *et al*. 2009 [50].

1.7.1 Structure of *M. tuberculosis* KasA

KasA, the mycobacterial β-ketoacyl-AcpM synthase, is a homodimeric assembly in its crystal structure. Each monomer is composed of two core domains and a capping region, with each core domain consisting of a mixed five-stranded β-sheet covered on each face by α-helices. The two core domains (N-terminal and C-terminal) are arranged into a five-layered αβαβα structure, characteristic of the thiolase superfamily [51, 52]. All members of this family have at least one catalytic cysteine in the catalytic triad, which is subjected to covalent modification during catalysis. In *Mtb* KasA, the catalytic triad is composed of Cys171, His311 and His345 (CHH). The cysteine residue is located in the N-terminal domain and lies at the N-terminus of an α-helix, whereas the other catalytic residues are located in the C-terminal domain. The N-terminal half consists of residues 2 – 259, whilst the C-terminal domain consists of residues 260 – 416. The hypothetical gate segment (GS), also known as the helix-turn-helix (HTH) region consists of residues 115 – 147, essential for the opening of the acyl cavity for substrate binding (Figure 1.3).

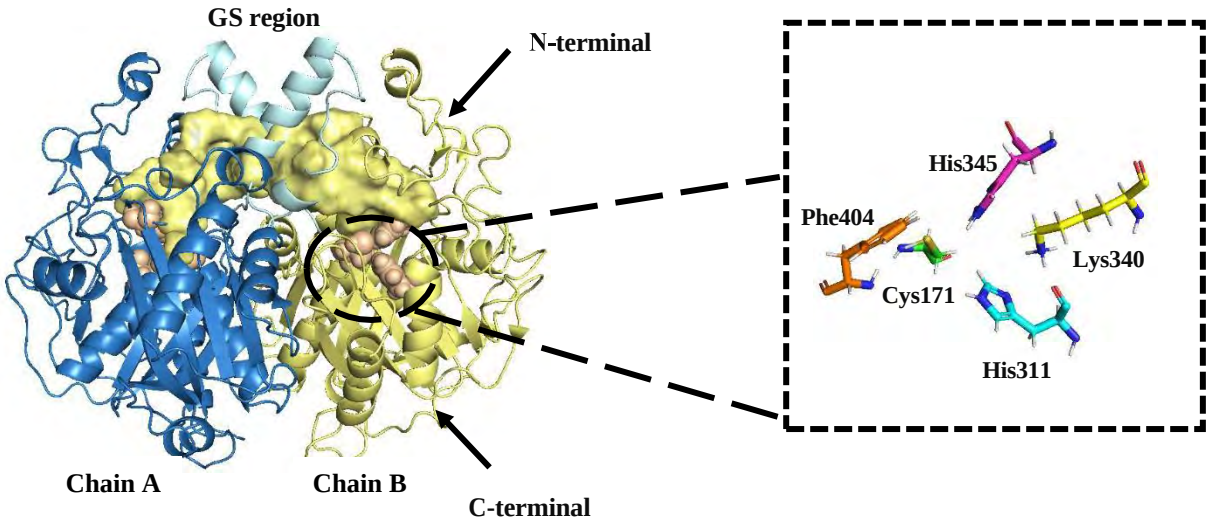


Figure 1.3: Overall structure of dimeric *Mtb* KasA (PDB ID: 6P9K). Each monomer of KasA is represented as a cartoon colored in blue and yellow, respectively. The active site residues are shown as a space filling representation on both monomers in wheat. The gate segment (GS) region is shown in pale cyan and the acyl-binding tunnel is represented as a closed surface in pale-yellow.

The acyl-binding channel of KasA is accessible through two openings: the malonyl binding pocket and the opening of the acyl channel at the surface of the protein [50]. This hydrophobic acyl binding

cavity is lined with hydrophobic amino acids that accommodate the growing fatty acid chain. The short hydrophobic side chains of Ala209, Ile122 and Ala119 facilitate fatty acid binding. The acyl binding tunnel poses two challenges in the mechanism of action of KasA: (1) migration of acyl chains through the malonyl binding pocket and past the hydrophilic and charged catalytic triad residues is energetically unfavorable, and (2) movement of the pantetheine group via the hydrophobic environment of the tunnel to the active site Cys171 residue causes steric hindrance. However, the gate segment of the protein is highly flexible and upon substrate binding, the scissor-like motion of residues contained in the gate segment (115 – 147) provide direct access to the acyl-binding tunnel by opening the channel. In addition, the acyl carrier protein (ACP) also facilitates the binding of the substrate and its interaction with the tunnel. The acyl binding tunnel thus serves as a shuttle to transport substrates to the active site for catalysis and moving products out of the active site and transferring them to other enzymes in the fatty acid synthesis pathway.

The lipophilic pocket comprises of the loop residues Asp273-Pro280 and two water molecules that are present in all KAS enzymes, that are under scrutiny for the role they play in the decarboxylation process. An important active site residue, Phe404 is presumed to act as a gatekeeper and facilitates the widening of the malonyl-binding pocket upon binding of an inhibitor. The second gate keeper, Lys340, in its protonated form is responsible for the formation of a salt bridge with adjacent deprotonated Glu345 residues. Its role as a functionally important residue was proven by mutational studies that revealed that the substitution of Lys340 resulted in diminished catalytic activity. Thus, this residue is believed to play an important role in maintaining the structure of the active site [53, 54].

1.7.2 Mechanism of action of *Mtb* KasA

KasA catalyzes the formation of mycolic acids via a ping-pong reaction mechanism. This reaction pathway has three main steps, namely acyl transfer, decarboxylation and condensation. In the first step, the active site cysteine is acylated by either an acyl-ACP or an acetyl-CoA molecule. The nucleophilic attack of the active site cysteine is facilitated by the dipole of the active site helix and an oxyanion hole composed of two backbone NH group atoms [51]. This is followed by a decarboxylation reaction, in which the active site residues His311 and His345 function in the decarboxylation of malonyl-AcpM. In addition to decarboxylation, the two catalytic histidine residues also play a role in stabilization of the acetyl-AcpM carbanion that is formed as a result of

decarboxylation. The condensation reaction results in the formation of a C-C bond between malonyl-AcpM and the acyl enzyme carbonyl group. Lastly, the acetyl-AcpM carbanion reacts with the acetylated enzyme thioester, leading to the formation of β-ketoacyl-ACP as a product. The hypothesis behind the mechanism of action of the KasA enzyme is supported by the fact that mutation of the active site cysteine residue leads to an increased decarboxylation of the malonyl-AcpM substrate by the two histidines [55, 56]. The reaction mechanism steps are shown in Figure 1.4.

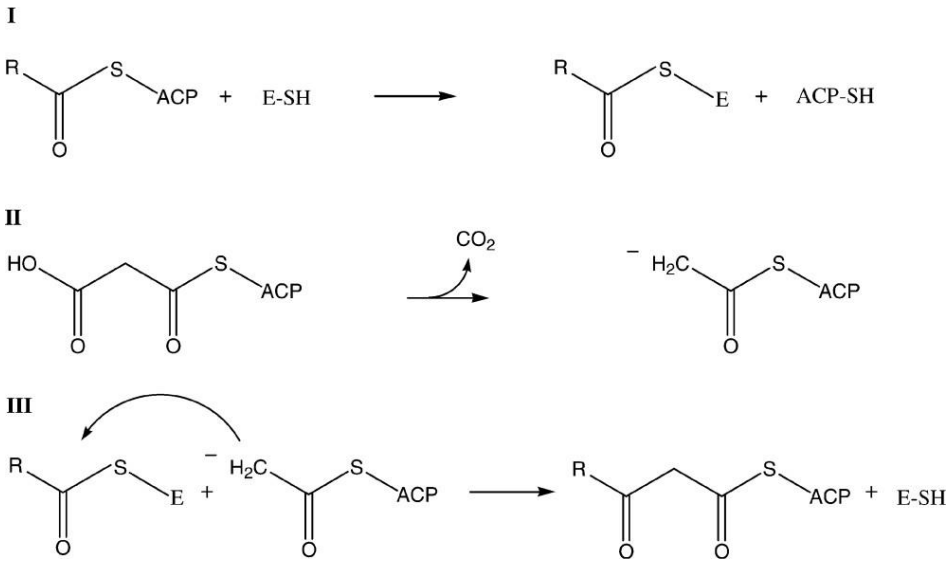


Figure 1.4: β-ketoacyl-ACP synthase I mechanism. (I) Acyl transfer of pantetheine-bound acyl primer to cysteine residue in the active site, (II) Decarboxylation of malonyl-AcpM to yield an acetyl-AcpM carbanion and (III) condensation reaction and two-carbon bond formation to give β-ketoacyl-ACP. Adapted from Bhatt *et al.* 2007 [56].

1.7.3 Inhibitors of *Mtb* KasA

Inhibitors of KasA have been reported in literature and these include platensimycin, thiolactomycin and its derivatives and cerulenin (Figure 1.5) [58–60]. Platensimycin is a natural product that has been isolated from *Streptomyces platensis,* and is known to inhibit KasA and KasB by binding to the acetylated form of the enzyme. It targets fatty acid synthesis by inhibiting FabH, the enzyme that performs the initial elongation of fatty acid chains from the FAS-I system.. Platensimycin targets the active site of the KasA enzyme, however, its mechanism of action is unknown. Alkyl-based substituents such as dihydroxybenzoate are known to interact with the catalytic histidines and occupy the area that the malonyl-AcpM substrate binds to. Binding of dihydroxybenzoate is believed to affect the decarboxylation stage of the mechanism of action of *Mtb* KasA. Thiolactomycin (TLM)

is also known to bind to the malonyl-binding pocket, thus acting as a competitive inhibitor of malonyl-AcpM [50]. Due to its favorable physicochemical properties, TLM is more sensitive to KasA than KasB [60]. TLM is known to form hydrogen bond interactions with Thr315 and His345 of the active site, thus inhibiting the decarboxylation stage of mycolic acid synthesis by *Mtb* KasA [61]. A mycotoxin-based inhibitor, cerulenin is produced by the fungus *Cephalosporium caerulens*. This molecule covalently interacts with and alkylates the active site cysteine residue via an epoxide opening of the ring. Structural analysis of KAS-cerulenin complexes reveal that cerulenin mimics the transition state of the condensation reaction by inhibiting the acyl-transfer stage of the mechanism of action of *Mtb* KasA [62].

However, further research has shown that the currently known inhibitors of *Mtb* KasA perform poorly *in-vitro*. In addition, a recent study revealed an overexpression of the *kasA* gene in treatment cases involving TLM in combination with a first-line anti-TB drug, isoniazid. Although cerulenin causes a growth inhibitory effect in *Mtb* KasA, it lacks stability and thus cannot be used for successful inhibition. Culture inhibition studies and *in-vitro* assay data indicate that platensimycin targets *Mtb* KasA and *Mtb* KasB, with a preferential targeting of the *Mtb* KasB protein [59]. This presents a research gap in the identification of compounds that can successfully inhibit *Mtb* KasA and produce the desired therapeutic outcome.

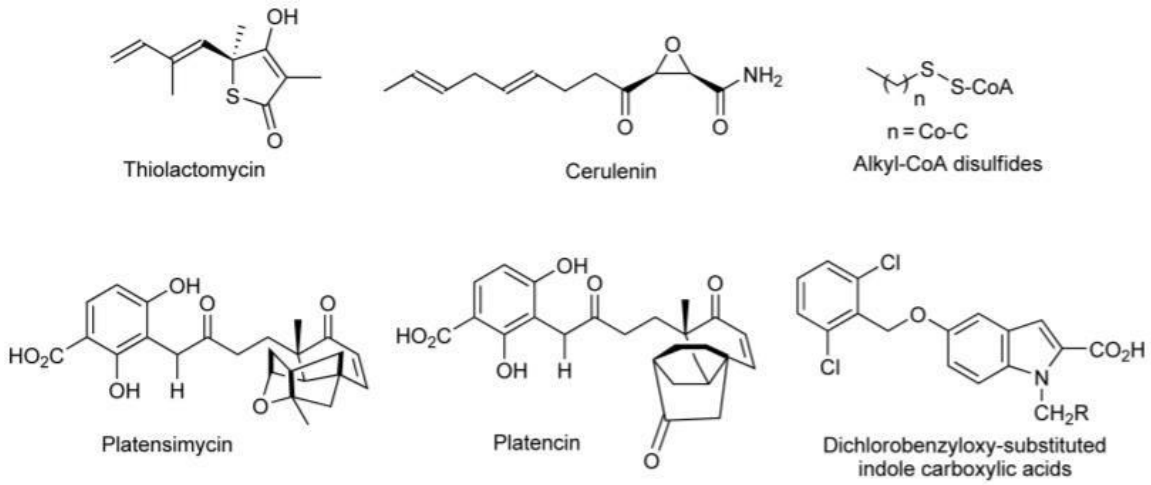


Figure 1.5: KAS inhibitors. Adapted from Zhang *et al.* 2010 [51].

1.8 DRUG REPURPOSING

Due to the costly and lengthy process of drug discovery and development, repurposing and the revival of drugs has become an attractive alternative strategy in TB treatment. Drug repurposing, also known as drug reprofiling, is a technique used to identify new uses for approved or investigational drugs apart from the already known indication of the drug [63]. This strategy offers numerous advantages over developing a new drug entirely. Firstly, the risk of failure is lower because the drug has already been approved for use and has been safe for use in humans after sufficient clinical trials. Furthermore, the time taken in drug development is sufficiently reduced as most of the time-consuming stages of pre-testing, formulation development and safety assessment have already been done. Thirdly, a minimal investment will be required in approving the drug for an alternative use, however, this is dependent on the stage of development of the drug. It has been noted that approximately 30% of the US Food and Drug Administration (FDA) approved drugs and vaccines are repositioned drugs [64]. Drug repurposing consists of three main steps: (1) identification of a candidate molecule for a particular indication, (2) assessment of the effects of the drug in preclinical models, and (3) evaluation of efficacy in phase I and II clinical trials. The most important step of this pipeline is the identification of the right drug for an indication of interest, which utilizes both *in-silico* and experiment-based methods.

Repurposing is not new to the treatment of TB. In the early 1930s, sulfonamides and sulphanilamide were used as anti-TB drugs but were discontinued due to the reduced efficacy compared to the first-line streptomycin and isoniazid. However, the revival of sulfamethoxazole (SMX) showed its efficacy in HIV-TB coinfections. Clofazimine is a repurposed molecule used in the management of MDR-TB and was initially used as an anti-leprosy drug in the early 1950s. It is recommended as a second-line agent used in combination with other anti-TB drugs. Linezolid, an antibiotic used in the treatment of gram-positive bacterial infections has now been repurposed for the treatment of MDR-TB and XDR-TB. Other repurposed drugs used in TB treatment include antibiotics such as biapenem and minocycline, antifungals such as artemisinin and chloroquine as well as antivirals such as isoprinosine [65]. As repurposing has been successfully performed in the past, this approach could assist with the identification of alternative drugs to treat TB and hence combat drug resistance.

1.9 PROBLEM STATEMENT

Most research in *Mtb* KasA drug discovery has reported on potential inhibitors that disrupt protein function by targeting the active site. Although recent studies have explored the inhibition of the

protein at sites distinct from the active site, there has been a loss of efficacy noted against *Mtb* strains with KasA mutations [66]. The current antibacterial drugs used in the treatment of TB target macromolecular synthesis; involving cell wall, protein and nucleic acid synthesis. Of these targets, cell wall biosynthesizing enzymes have historically proven to be effective because they are unique to bacteria, thereby limiting toxicity to mammalian cells. However, most of the promising inhibitors of *Mtb* KasA target more than one protein in the fatty acid synthesis pathway, thus lowering enzyme specificity.

This presents three major research gaps in relation to the identification of novel therapeutics that can be used to treat TB and alleviate the issue of drug resistance. Firstly, there is need to discover drug compounds that can successfully inhibit *Mtb* KasA with high enzyme specificity. In addition, the currently known inhibitors of *Mtb* KasA are not FDA-approved drug compounds and are still undergoing clinical evaluation. To combat this challenge, drug repurposing of FDA-approved compounds can be used to identify potential inhibitors and reduce the time and costs involved in developing new antimycobacterial agents. Lastly, exploring allosteric inhibition of the *Mtb* KasA protein would be beneficial to the field of research as it has been less explored and could provide valuable insights and aid in the identification of novel compounds that can be used to treat TB. Allosteric drugs are highly specific and do not bind to the active site, which is highly conserved in protein families. Furthermore, they have improved selectivity and a lower potential for side effects, as seen with orthosteric drugs.

In order to accomplish the aims of the End TB strategy and improve the quality of life, extensive research and the development of new drugs is essential. This study focuses on the identification of allosteric sites on the *Mtb* KasA protein and selective compounds displaying preferential binding for these sites for inhibitor drug design purposes.

1.10 HYPOTHESIS

This study hypothesizes that the mycolic acid synthesis pathway of the KasA enzyme in *M. tuberculosis* provides an excellent target for inhibitory compounds that could potentially be used as anti-mycobacterial agents against MDR-TB.

1.11 AIM OF THE STUDY

The aim of this research is to virtually screen DrugBank compounds against the allosteric sites of *Mtb* KasA via molecular docking and identifying selective hits of interest using the binding energies as well as the protein-ligand interactions.

1.12 STUDY OBJECTIVES

1. Protein characterization via sequence analysis methods to understand the structural and functional relationships between *Mtb* KasA and homologs derived from diverse species.
2. Allosteric sites identification on the *Mtb* KasA and human homolog protein (*Hsmt* KasA) by using various computational approaches.
3. Identification of potential DrugBank compounds that bind to the allosteric sites of *Mtb* KasA by performing molecular docking studies using AutoDock Vina.
4. Screening of docking analysis outcome according to the binding affinities.

CHAPTER TWO

SEQUENCE AND STRUCTURAL ANALYSIS

The complexity of the mycobacterial cell wall has enabled it to be one of the most successful targets of *Mtb* drug discovery. The unusually complex, lipid rich coat of the cell wall has been proposed to be critical for its pathogenicity and it is believed to provide inherent resistance to many antibacterial agents. Mycolic acids are a distinctive feature of the *Mtb* cell wall and are characterized by long chain α-alkyl (C_{60} - C_{90}), β-hydroxy fatty acids that are responsible for the low-permeability, acid-fastness and virulence of *Mtb* [68, 69]. Enzymes involved in the biosynthesis of mycolic acids thus represent fascinating targets for the novel development of anti-tubercular agents.

KasA, the mycobacterial β-ketoacyl- [acyl-carrier protein] synthase is an essential enzyme in the fatty acid synthesis II (FAS-II) pathway responsible for the biosynthesis of mycolic acids. Inhibition of the mycolic acid synthetic pathway thus serves as a starting point in the development of potential inhibitory compounds in drug discovery. This chapter is aimed at analyzing homolog protein sequences of KasA, derived from different species by use of various bioinformatics search tools and methods. Fifteen sequences of KasA were obtained in total, eight of these obtained from bacteria, three from fungi and four from mammals. A better understanding of the sequence and structure of the KasA enzymes, particularly *Mtb* KasA, is essential in extracting knowledge about the biological function of the enzyme and this also paves a way for the development of therapeutic agents. This chapter also focuses on the evaluation of the similarities and differences between *Mtb* KasA and the human homolog in terms of sequence and structure in order to identify regions that are unique to *Mtb* KasA that could serve as a drug target.

2.1 INTRODUCTION

Biological sequence data has displayed an exponential growth over the years. The availability of expansive databases on DNA, RNA and protein sequences as well as computational tools for sequence analysis has positively impacted the field of research in that data retrieval and analysis can be conducted for a wide range of scientific projects. Protein sequence analysis involves subjecting an amino acid sequence to *in-silico* methods in order to study its function, structure and evolution. Analytical methods employed in this process include sequence alignment, motif discovery, phylogenetic analysis and other methods. This is particularly important as it permits for

the recognition of conserved residues and residue groups that are structurally and functionally important. Substitution of amino acids with similar physicochemical properties also allows for the preservation of the protein structure within the protein families. Sequence alignment plays a vital role in identifying regions of similarity that could reflect biological relationships among the input sequences [69]. In addition, alignment of protein sequences also assists in the development of homology models as well as in the construction of phylogenetic models. Motif analysis entails identifying conserved patterns within protein families that are important for structure and function. These patterns are represented either as sequence logos or as regular expressions [70]. These identified patterns have structural and functional properties that are used in the characterization of a protein of interest. Phylogenetics involves the study of evolutionary ties among biological species to gain an understanding of the relationship between the ancestral sequence and the descendants that arose from evolution. Detailed protein sequence analysis and the specific programs used for the aforementioned processes are outlined below.

2.1.1 Protein Sequence Alignment and Algorithms

Sequence alignment plays a vital role in the analysis of newly determined DNA and protein sequences. There are fundamentally two types of alignments, namely global and local alignment. A global alignment is an end-to-end alignment of sequences performed to find the best alignment across the entire length of the sequences, whereas local alignments focus on aligning local regions with the highest level of similarity between the sequences [72, 73]. The optimal alignment of two sequences is a computationally exhaustive task that incorporates a technique called dynamic programming. This is an algorithmic approach that involves matching two sequences in search of all possible pairs of characters and produces a scoring matrix that accounts for matches, mismatches and gaps. The alignment scores generated from the scoring matrices also take into consideration the gap penalties as well as the pairwise substitution scores obtained from the matched residue types. The commonly used substitution matrices are the PAM (Point Accepted Mutations) and the BLOSUM (BLOck SUbstitution Matrix).

The PAM matrices were first developed by Margaret Dayhoff and colleagues in the early 1970s using a set of closely related proteins with a minimum 85% sequence identity [73]. Dayhoff calculated the probability of an amino acid being replaced by another at a set evolutionary distance. A PAM unit is 1% amino acid change per 100 residues. The increasing PAM units represent an increase in the evolutionary distances as denoted by the Markov model in which the PAM1 matrix

can be multiplied by itself *N* times. Thus, a PAM250 matrix represents 250 mutations per 100 residues. As such, the PAM250 matrix is used for divergent sequences whereas the lower PAM matrices are used for short alignments that are highly similar. The BLOSUM matrices were developed by Henikoff from an inspection of every possible substitution of amino acids in multiple sequence alignments. Ungapped alignments (blocks) of less than 60 amino acids, derived from divergent sequences were used to calculate the frequencies of residue substitutions in order to produce a block substitution matrix [75, 76]. The alignment blocks represent conserved regions within related proteins and are used to compute the log odd (LOD) scores. BLOSUM62 is the most commonly used matrix and it illustrates that the sequences used for the construction of the matrix share an identity of 62%. Higher BLOSUM matrices are used for highly similar, short alignments whilst lower matrices are applied to divergent sequences. Other commonly used matrices include the VTML matrix developed by Vigron and Mueller using alignments from highly divergent sequences [76].

2.1.2 Database Similarity Search

Sequence similarity search is predominant in the analysis of biological data in bioinformatics. It is used primarily to identify homologous sequences in order to provide information on the protein structure, function and phylogeny. In addition, the information derived from the homologous sequences can be used to characterize and annotate the query sequence. In order to understand how this approach works, knowledge on sequence similarity and homology is required. Sequence similarity refers to the measure of degree in which two sequences are alike or similar. It is a quantitative measure that is represented as score or percentage. On the other hand, sequence homology refers to the inference of an evolutionary relationship between two sequences based on similarity and is used to deduce a common ancestral relationship. Database similarity searching involves the submission of a query sequence and conducting a pairwise comparison of the query sequence with all the sequences in the database . In order to implement the algorithms used in database similarity search, a stringent criterion is followed. Firstly, a sensitivity search method is performed to identify as many correct hits as possible and to avoid missing distant homologues. This is followed by a specificity test, to exclude any incorrect hits or ‘false-positives’. Another vital requirement in database similarity search is the speed at which the results from the computation are displayed. Databases may also contain very large families of repeated sequences or motifs thus it is important to conduct similarity search against non-redundant databases to avoid bias and misleading

scores.

Specialized programs have been developed to perform database similarity searching. These include BLAST [77], PSI-BLAST [78], FASTA [79], SSEARCH [71]and HHpred [80]. These programs produce statistical estimates which are used to infer homology to sequences that share significant similarity. Basic Local Alignment Search Tool (BLAST) is a heuristic that attempts to find short matches from two sequences and aligns the identified matching regions. Several variations of the BLAST program are used to search for sequence similarity in protein and DNA databases. These include BLASTP, BLASTN, BLASTX, TBLASTN and TBLASTX. BLASTP compares protein queries to protein databases and is a prototype of the BLAST family. This program requires time relative to the lengths of the sequences of the query sequence and the database.

Position Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) is an iterative form of the algorithm used in BLAST that begins with a regular BLASTP search then builds up a position specific scoring matrix (PSSM) from the best hits. This approach is highly sensitive and is used to detect distant homologs in protein databases [79, 82]. HHpred applies a pairwise comparison of profile Hidden Markov Models (HMM) for remote protein homology and structure prediction. Statistical measures such as E-value, p-value, percentage identity and sequence coverage are used to identify the homologous sequences from database similarity searching.

2.2 MULTIPLE SEQUENCE ALIGNMENT

Multiple Sequence Alignment (MSA) involves the comparison of more than two homologous sequences in order to perform phylogenetic reconstruction, structural and functional prediction analysis. The accuracy of MSA results is essential in the analysis of biological data as this process is a prerequisite of other bioinformatics approaches used in the identification of conserved regions of functional importance. A multiple sequence alignment arranges protein sequences in such a way that there is a maximum number of matches and integrates this into a scoring function based on the sum-of-pairs (SP). When calculating the SP scores, factors such as pairwise matches, mismatches and gaps are included. Gaps in an MSA represent insertions and deletions of biological information within sequences as a result of evolution. Most MSA techniques use a global alignment, however, when there is a large difference in the lengths of the sequences being compared, a local alignment is used instead. The most accepted heuristic used in MSA is a progressive alignment. This approach uses a pairwise alignment from algorithms such as Needleman-Wunsch and Smith-Waterman to

solve a complex MSA task [82]. The sequences are clustered together and similarity scores are calculated from pairwise comparisons. Guide trees are built based on the similarity scores using methods such as Neighbor-Joining (NJ) [83] and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [84]. The drawback of this heuristic is that it only focuses on two sequences at a time, therefore an optimal result is not guaranteed. In addition, any mistakes in the initial pairwise alignment cannot be rectified in the later stages and remain fixed. An improved version of the progressive alignment called the iterative method was developed to enhance the accuracy of the MSA. This method performs post-processing of the progressive alignment by modifying the guide tree construction. The benefit of this approach is that any errors made initially can be corrected, thus improving the quality of the alignment. The algorithms that employ the progressive alignment technique include Clustal Omega [85], MAFFT (Multiple Alignment using Fast Fourier Transform) [87, 88], PROMALS3D (PROfile Multiple Alignment with Local Structures and 3D constraints) [88] and MUSCLE (MUltiple Sequence Comparison by Log Expectation) [89], among others. In addition, all the aforementioned algorithms with the exception of PROMALS3D also employ an iterative alignment technique.

The quality of the alignment produced can be determined in various ways. One way is to utilize refined repositories such as BAliBASE [90] and SABmark [91] as benchmarks upon which simulated alignments can be compared to gauge performance of the alignment programs and determine accuracy. The disadvantage, however, of using this approach is that benchmarks do not cover the full range of scenarios of protein evolution and require high level expertise to generate them. To overcome this problem, several programs and webservers have been developed to evaluate the alignment accuracy using different scoring techniques, with the assumption that the alignment with the highest score outperforms the others. The Transitive Consistency Score (TCS) webserver is part of the T-Coffee web platform and uses a pre-computed MSA to estimate the reliability of every pair of aligned residues (*PairTCS*), every MSA column (*ColumnTCS*), every MSA sequence (*SequenceTCS*) and the whole alignment (*AlignmentTCS*) [92]. These metrics can be used to identify aligned positions most likely to contain structurally analogous residues that can be used for homology modelling and phylogenetic reconstruction purposes.

2.3 MOTIF ANALYSIS

Motifs are defined as a short, conserved sequence patterns that are used to identify vital structural and functional regions within a group of closely related proteins. These regions are generally more

conserved than other regions of a protein and tend to evolve as a unit [93]. Motifs are used to predict sites that are essential for the functioning of the protein such as active sites, ligand binding sites, post-translational modification sites as well as cleavage sites. The identification of motifs immensely relies on multiple sequence alignment (MSA) as well as profile hidden Markov models. Two models have been used to facilitate the identification of unknown patterns in protein sequences and these are deterministic [94] and probabilistic [95] models.

Deterministic models use consensus sequence patterns, also known as *regular expressions,* to find the number of matches against a sequence of interest and determine a potential functional site. A regular expression is a string of characters comprising of exact and ambiguous symbols as well as flexible gaps. Two mechanisms are used to match regular expressions with a query sequence: exact and fuzzy matching. Exact matching uses a strict matching of sequence patterns and no variations are permitted. Fuzzy matching is an approximate which allows more flexible matching of residues with similar physicochemical properties. Probabilistic models employ a position-specific scoring matrix (PSSM), whereby each entry (*i, a)* is the probability of finding the amino acid *a*, at its *i*th position in a sequence motif [96]. Information from the PSSM can be represented as a sequence logo; a stack of letters at each position in the motif, where the size of the letters indicates their frequency in the sequence.

Various databases are used in the characterization of proteins to help identify motifs and domains. Commonly used motif databases include PROSITE [97] and Pfam [98]. The MEME suite tools [99] are widely used for the discovery of protein motifs. The Multiple Expectation Maximization for Motif Elicitation (MEME) discovers ungapped motifs in sequences by using a probabilistic algorithm. It uses statistical modelling techniques to select the best width, number of occurrences and description of each motif [100]. The Motif Alignment Search Tool (MAST) determines the best match in the sequence based on the MEME results. The scores for the best matches are combined into E-values, and motifs with values below 0.001 are considered. Motifs with MAST pairwise correlation values greater than 0.6 are disregarded [101].

2.4 PHYLOGENETIC ANALYSIS

Phylogenetics is a process used to estimate the evolutionary relationships by examining biological data such as DNA, protein sequences or morphological data from taxa [102]. This process was initially studied in order to evaluate the historical relationships between protein sequences but has

diversified into the study of mechanisms involved in microbial outbreaks, prioritizing the conservation of endangered species as well as inferring the function of genes that have not been studied experimentally [101, 102]. A phylogenetic tree is a diagrammatic representation with tree-like resemblance built in four main steps, namely: (1) identifying a set of homologous DNA/protein sequences of interest, (2) performing a multiple sequence alignment, (3) selecting a phylogenetic reconstruction method and (4) identifying and evaluating the best tree. Phylogenetic trees can be constructed via distance-based or character-based methods [105]. When using the distance-based approach, pairwise alignments are converted to distant values that are used to generate a distance matrix. The distance matrix is incorporated into the tree building approaches such as UGPMA, weighted pair group methods with arithmetic mean (WGPMA), neighbor-joining (NJ), least squares (LS) and minimum evolution (ME). As for the character-based method, approaches such as maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference have been used in the tree construction process. The Molecular Evolutionary Genetics Analysis (MEGA) software is widely used in the statistical analysis of molecular evolution and in the construction of phylogenetic trees[106]. It utilizes both distance-based and character-based methods in the tree construction process and allows for visualization of the constructed tree in a tree explorer. The evaluation of the reliability of the constructed tree is performed using the bootstrap test and the standard error test. Other commonly used programs used for phylogenetic analysis include MrBayes [107], Randomized Axelerated Maximum Likelihood (RAxML) [108] and Bayesian Evolutionary Analysis Sampling Trees (BEAST) [109].

2.5 METHODOLOGY

The flow diagram of the web-based bioinformatics tools and databases used in the analysis of the *Mycobacterium tuberculosis* KasA protein at sequence and structural level is presented in Figure 2.1.

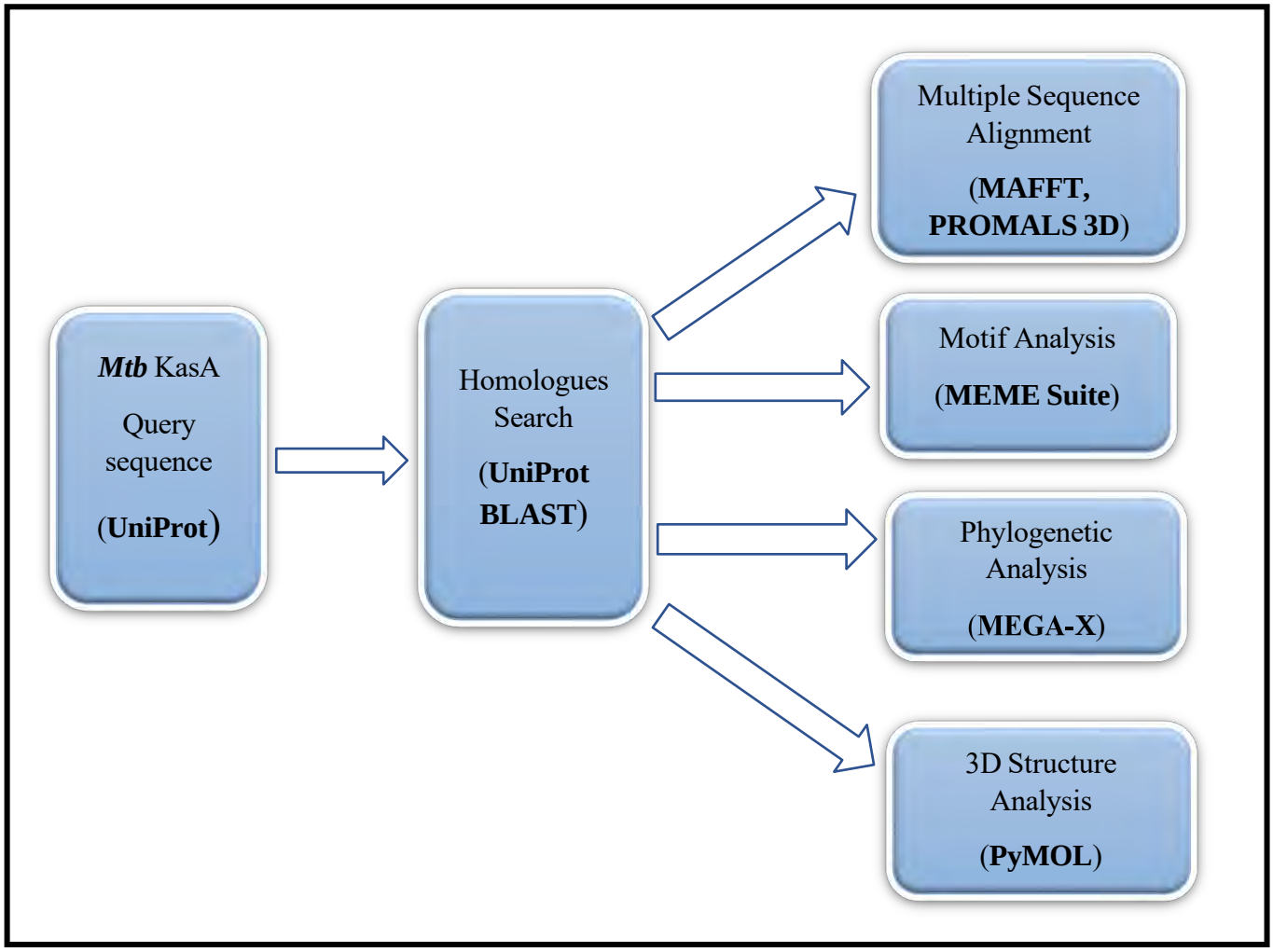


Figure 2.1: Summary workflow of analytical approaches used in the identification of KasA protein homologues.

2.5.1 Data retrieval

2.5.1.1 Protein sequence retrieval

Mtb KasA protein sequence (accession number : P9WQD9) was obtained from the UniProtKB database by searching using the key words: *M. tuberculosis* and KasA [110]. Using *Mtb* KasA as a query sequence, homologous sequences from various organisms such as bacteria, fungi and mammals were derived using the BLASTP algorithm in UniProt [111], together with a BLOSUM62 matrix and E-threshold value of 1000. A gap penalty of 11 and an extension penalty of 1 was also used. All the retrieved protein sequences were obtained from 8 bacterial, 3 fungal and 4 mammalian species, inclusive of the query sequence. The statistical measure, E-value, was vital to selection of sequences. Only sequences having an E-value below 1.0 e-5 were considered as significant. In

addition to the criteria used for sequence selection, query coverage was also considered. The BLAST details used on the data set used in the analysis are shown in Table 2.1.

Table 2.1: KasA homologues sequence details. The sequence identities are given relative to *M. tuberculosis*.

Source Organisms	E-value	Sequence Identity (%)	Accession code (UniProt)	Residue Count
M.bovis	0.0	100	P63455	416
M.tuberculosis	0.0	100	P9WQD9	416
Arthrobacter sp.	0.0	85.6	A0A542NQV0	416
Rhodococcus sp.	0.0	71.6	A0A5A7SCX5	416
Norcadia cyriacigeorgica	0.0	70.6	A0A5R8NXG2	416
Gordonia paraffinivorans	0.0	67.7	A0A2YIZNN7	422
Willimsia limnetica	0.0	66.8	A0A318RJO0	421
Segniliparus rotundus	0.0	64.4	D6ZB28	419
Verticillium dahliae	1.2e-85	39.8	A0A366NG543	424
Epicoccum nigrum	2.3e-84	37.9	A0A1Y2LXX4	430
Mus musculus	2.8e-67	36.5	Q9D404	459
Botytis porri	9.3e-77	35.9	A0A4Z1K683	430
Homo sapiens	1.3e-66	35.8	Q9NWU1	459
Bos taurus	1.7e-65	34.9	F1MXW5	460
Rhinolopus ferrumequinum	5.4e-67	34.7	A0A671ENN5	459

2.5.1.2 3-D protein structure retrieval

The crystal structure of *Mtb* KasA (UniProt accession code: P9WQD9), strain ATCC 25618/H37Rv has been solved in 10 structures and the *Homo sapiens* mitochondrial KasA (*Hsmt* KasA) (UniProt accession code : Q9NWU1) has been solved in 3 structures. Co-ordinate files for the crystal

structures of *Mtb* Kas A (PDB ID: 6P9K) and *Hsmt* KasA (PDB ID: 2IWY) were retrieved from the Protein Data Bank (PDB) [112]. Both structures were determined via X-ray diffraction and the resolutions are 1.70Å and 1.65Å respectively. These structures were visualized in PyMOL [113].

2.5.2 Multiple Sequence Alignment (MSA)

Multiple sequence alignment of *Mtb* KasA and the retrieved homolog sequences was performed using three alignment tools: PROMALS3D, MAFFT and Clustal Omega. Crystallographic secondary structure information obtained from the PDB file of *Mtb* KasA was added to PROMALS3D to enhance alignment quality, producing a consensus alignment with sequence and structural information of the input protein and its homologs. The sequence alignment parameters used for the MAFFT alignment included a BLOSUM62 matrix as well as a gap opening and extension penalty of 1.53 and 0.123 respectively. Default parameters were used for the Clustal Omega alignment. Visualization and editing of the alignments was performed using the Jalview vs. 2.11.1.4 software [114]. The TCS webserver was used to evaluate the resultant alignments from the different tools and determine the best performing alignment from the calculated scores.

2.5.3 Motif analysis

In order to identify the distribution of motifs within the KasA proteins in the dataset, the online MEME suite vs. 5.3.3 was used to conduct motif discovery. The MEME parameters used included a minimum and maximum motif width of 3-20 amino acid residues and a zero-order model of sequences [115]. A total of 100 motifs was set for discovery. Overlapping motifs were detected using the MAST tool [116]. Validation of the discovered motifs was performed using the MAST file and E-values in order to meet the criteria required for inclusion or omission. Motifs with pairwise interaction greater than 0.6 were discarded from the MAST results and only motifs that had an E-value less than 0.001 were retained. The final data set was reduced to 44 motifs. In-house python scripts were used to calculate motif conservation heatmaps and map motifs onto their respective 3D structures using PyMOL.

2.5.4 Phylogenetic analysis

The evolutionary relationships of living organisms are generally inferred through molecular phylogenetic analysis, which employs various mathematical methods. The MEGA X software was used to study the evolutionary relationships within the KasA protein sequences. The MAFFT

alignment was used as an input file and the evolutionary models were calculated using three gap deletion options (90%, 95% and 100%). The best three models for each deletion option were selected based on the lowest Bayesian Information Criterion (BIC) scores. The Maximum Likelihood (ML) statistical method was used to construct phylogenetic trees for the top three models for each gap deletion option. The Le and Gascuel model with gamma distribution (LG+G) [117], Le and Gascuel model with gamma distribution and invariable sites (LG+G+I) and the Whelan and Goldman model (WAG+G) [118] were used for phylogenetic tree construction. A total of 9 trees (3 x 3 x 3) were generated. Bio-NJ and Neighbour Join algorithms were used for a matrix of pairwise distances to obtain the initial trees for the heuristic search and the topology with the highest log likelihood. A strong branch swap filter and 1000 bootstrap replicates were used in each tree construction. Lastly, the comparison of the trees generated to the bootstrap consensus tree was performed in order to determine the reliability of the construction process and to ensure that the branching patterns were accurate [119]. The best model and gap deletion was then selected for further analysis.

2.6 RESULTS AND DISCUSSION

2.6.1 Sequence Analysis

A total of fifteen homologous sequences were retrieved from the UniProt database using *Mtb* KasA as a query sequence. These homolog sequences belong to the bacterial, fungal and mammalian classification of organisms. Eight of these sequences were bacterial in nature, three were fungal and the rest were mammalian. The BLAST search results displayed the query sequence as the first hit and this had a 100% sequence identity, together with *M. bovis* which is a member of the same genus (Table 2.1). The bacterial sequences had the highest sequence identity when compared to the query sequence (> 60%), followed by fungal sequences and lastly mammalian sequences. Both the fungal and mammalian sequences had a sequence identity below 40%. From the sequence identities, it can be deduced that the sequences obtained using *M. tuberculosis* KasA as a query sequence are homologous. Most of the bacterial sequences had similar lengths, however, the fungal and mammalian sequences were longer, and this also explains the low sequence identities and divergent nature as regions of local similarity decrease with sequences of varying lengths.

2.6.2 Multiple Sequence Alignment

Multiple sequence alignment of the homologous sequences was carried out to identify highly

conserved sequence regions, postulated to have structural and functional significance within the KasA family, as well as unique sequence regions in the *Mtb* KasA protein that can be exploited for inhibitor drug design purposes. In order to evaluate the best alignment, a comparison of the outcomes produced by the different alignment tools was performed on the TCS webserver. Generally, all the alignment programs produced a high overall alignment score, suggesting their accuracy in identifying conserved regions of structural and functional importance. The N-terminal region was less conserved in the PROMALS3D and Clustal Omega alignments, as shown in Figure S1 and Figure S2 respectively. The inserts in the alignments, largely concentrated in the N-terminal region, were observed to be occurring in the loop regions when mapped to the crystal structure of *Mtb* KasA. The MAFFT alignment produced the highest alignment score and was selected as the most suitable alignment for subsequent analyses. This alignment is shown in Figure 2.2.

2.6.2.1 Areas of conservation and insertions

The multiple sequence alignment generated showed a conservation of residues among the homolog sequences, but with some areas of non-conservation as well (Figure 2.2). The columns are colored per percentage identity. The MSA of *Mtb* KasA showed a high conservation of the active site residues, particularly the key catalytic active site residues, namely Cys171, His311 and His345. In addition, the gate keeper residues, Phe404 and Lys340 were also conserved. These residues are believed to be involved in maintaining the integrity of the active site and previous studies have revealed that mutations involving any of these residues result in diminished catalytic activity [53]. Other highly conserved residues include the hydrophobic residues which are located in the acyl-binding tunnel and interact with the substrate, namely Phe237 and Phe239. Phe237 is highly conserved in all the KasA sequences, whilst Phe239 is highly conserved in the bacterial KasA sequences but is substituted by a Val or Ile residue in fungal KasA sequences and a Met residue in mammalian sequences. In addition, the acyl-binding cavity forming residues Ile347 and Ala349 are also highly conserved in KasA sequences, with Ile347 showing high conservation in bacterial KasA sequences but is substituted by a Lys residue in fungal and mammalian sequences. The residues Thr313 and Thr315 involved the ACP-substrate binding were also highly conserved in all sequences. Another highly conserved residue is Ser41, which is interacts with the growing fatty acid chains in the acyl-binding cavity. This residue is highly conserved in bacterial and mammalian KasA sequences but is substituted by a Cys residue in fungal sequences. All in all, the active site residues as well as residues involved in key interactions with the substrate of KasA were highly conserved, thus

denoting their importance in protein function.

Clearly visible from the MSA is the N-terminal insert **I**, made up of ≈ 40 residues at the start of the alignment, as a result of the mammalian sequences. The *Hsmt* KasA protein has an N-terminal mitochondrial target peptide that includes 38 amino acid residues [116, 117]. This feature is believed to be only unique to mitochondrial KAS enzymes. Previous studies on *Hsmt* KasA revealed that the target peptide has an overall positive charge resulting from the relatively high proportion basic amino acids and a lack of negatively charged residues. Other inserts that were identified across the alignment are shown in red boxes. The second largest insert, **II** made up of 7 residues is located between residues Pro93- Phe102 in mammalian sequences. According to previous studies on the protein, this is a loop region that connects the first and fourth strands of the N-terminal β-sheet and is only found in mitochondrial KAS enzymes, as shown in Figure S4 [121]. Relatively small inserts in the alignment were located in the C-terminal region. These inserts of about 3 residues were due to both mammalian and fungal sequences.

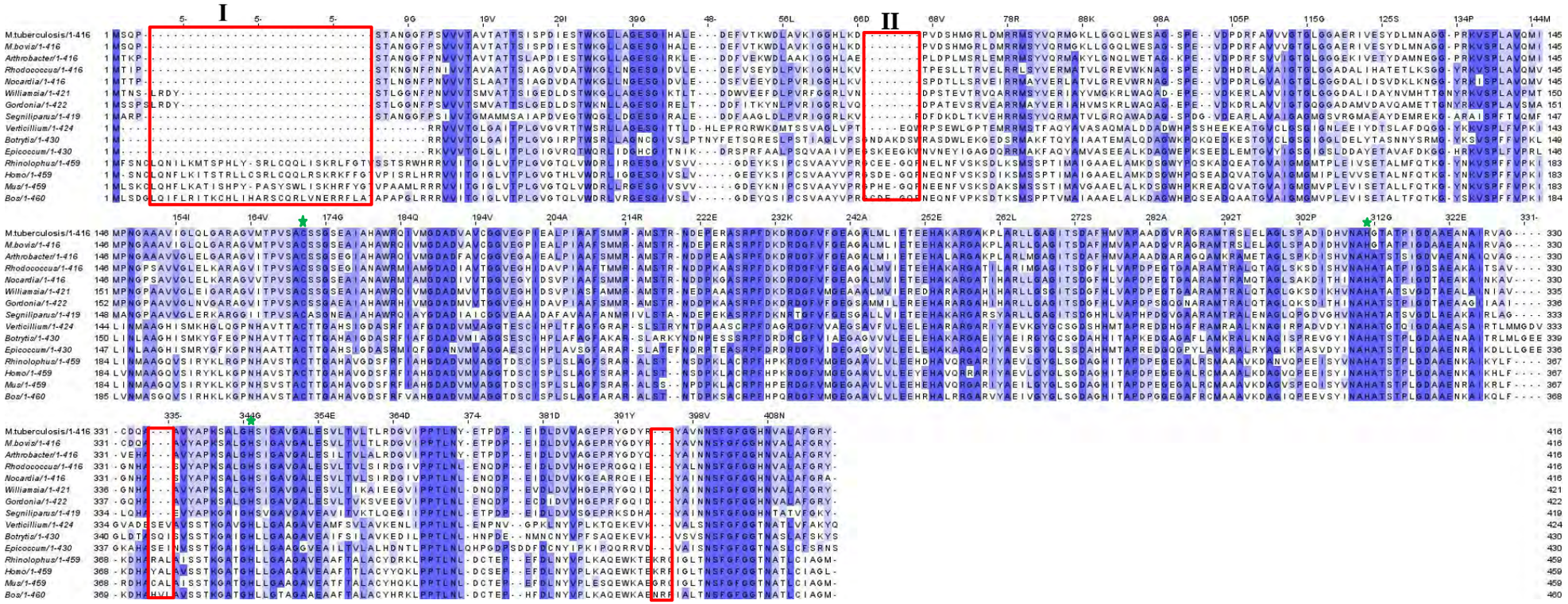


Figure 2.2: MAFFT multiple sequence alignment of *Mtb* KasA and its homologs. Residue numbering is given for *M. tuberculosis*. Residues of the catalytic triad are depicted by green asterisks. The inserts due to the mammalian sequences are shown in red boxes. These residues were conserved across the homologs of KasA. The alignment is colored by percentage identity.

2.6.3 Motif Analysis

Protein sequence motifs are signatures of protein families used as tools for the prediction of protein function. The MEME software is widely used to identify motifs in DNA or protein sequences. The MEME parameters were set to find motifs of 3-20 residues in length in order to cater for motifs that are longer than the short linear average of 3-11 residues [101]. MEME identified a total of 18 highly conserved motifs across all the sequences. These results were displayed as a heatmap using an in- house python script as shown in Figure 2.3. MEME was set to discover 100 motifs, however, the pairwise correlation analysis in MAST reduced the final data set to 39 motifs. Motif conservation in the KasA homologs is represented as the number of sites per total number of protein sequences. The motifs identified were numbered according to the MEME output. A value of zero on the heat map indicates the absence of the motif in any of the protein sequences whereas a value of 1 indicates a 100% conservation of the motif in all sequences.

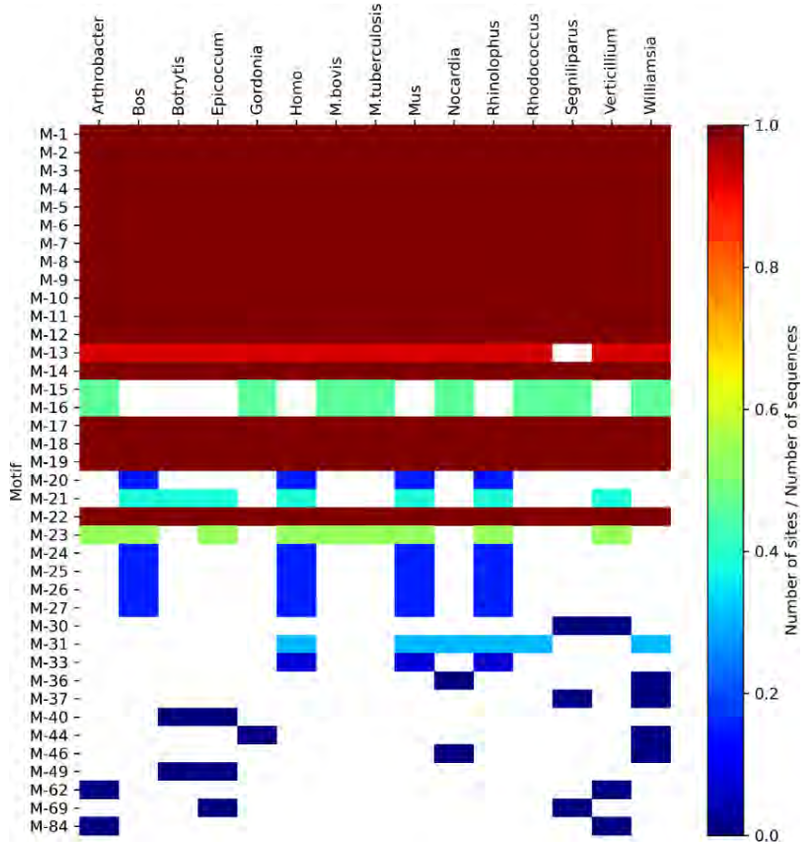


Figure 2.3: Motifs identified in KasA homologs presented as a heatmap. The colours represent the conservation of motifs identified. Conservation increases from blue to red, while the absence of motifs is shown by a white colour.

The E-values of the highly conserved motifs (M1-M14,M17-19 and M22) are shown in Table 2.2. From the table, it was observed that all the highly conserved motifs have an E-value less than 0.001. These motifs were selected for further analysis. Motif 1 contains residues involved in the binding of acyl-fatty acid products of catalysis. This motif contains hydrophobic residues such as Phe237 and Phe239, which are positioned to interact with the substrate or inhibitor, as is the case with Thiolactomycin [50]. Motifs 2, 8 and 4 are involved in the catalytic function of the enzyme. The active site Cys171 is located in motif 4 whilst the other catalytic triad residues, His311 and His345 are located in motifs 2 and 3 respectively. The histidines play an important role in the decarboxylation of the malonyl-AcpM substrate as well as in the stabilization of the acyl-carbanion intermediate. In addition, nitrogen atoms of the histidines are involved in hydrogen bonding, which is believed to help stabilize the loop region between Asp273-Pro280. In addition to the catalytic function, Lys340 in motif 8 plays an important role as the 'second gatekeeper' residue to the malonyl-binding pocket. This residue helps to maintain the integrity of the active site thus aiding in enzyme stability. Residues 347 and 349 of this motif are also involved in the formation of the acyl binding cavity. This cavity connects the active site and the pocket at the surface of the protein and is responsible for accommodating the growing fatty acid chain.

Thr313 and Thr315 of motif 2 are highly conserved residues in the KAS family and are proposed to be involved in ACP-substrate binding [52]. Motif 7 contains the lipophilic pocket residues (Asp273-Pro280), and Motif 9 contains a Ser41 residue that interacts with the growing fatty acid chains in the acyl-binding tunnel. Arg135, Lys136, Val142 and Met146 in motif 6 are positioned to interact with the malonyl-AcpM substrate. In this orientation, the bound acyl-carrier protein permits the entrance of substrates via the phosphopantetheine tunnel opening. Another key interesting observation is that the HTH region of the KAS proteins is associated with this motif. This region is responsible for opening the acyl-binding tunnel during catalysis in order to shuttle substrates and products in and out of the active site. Motif 10 and 17 contain residues that make up the acyl-binding tunnel, which acts as a shuttle between KasA and other enzymes in the fatty acid synthesis pathway. The arginine residues in motif 14 (R74, R78 and R79) are essential for ACP-substrate binding. Some conserved motifs such as M5,M11-13,M18-19 and M22 had no related known function from literature. However, it would be interesting to further investigate the functionality of these motifs. Mapping of the highly conserved motifs onto the 3D structure of *Mtb* KasA is shown in Figure S5.

Table 2.2: Highly conserved motifs in KasA homologs. The starting and ending positions of the motifs in *Mtb* KasA, constituent residues, values and contribution to function is displayed. Residues with important functionality in the motif are underlined and highlighted in bold.

Motif	Residue Range	Residues	E-Values	Residue Count	Function
1	220-240	NDDPKAASRPFDKDRDG**F**V**F**	7.4E-169	20	Hydrophobic pocket residues
2	307-327	HVNA**H**A**T**S**T**PLGDAAEAKAI	4.4e-151	20	Catalytic mechanism and ACP-substrate binding.
3	397-417	AINNS**FGFGG**HNAALAFGRY	5.0e-150	20	Active site pocket
4	166-186	TPVSA**C**SSGAEAIADAWRFI	9.5e-132	20	Catalytic mechanism
5	367-382	LPPTLNLENPDPEID	1.2e-108	15	-
6	132-152	GGY**RK**VSPLA**V**PMI**M**PNGAA	1.5e-137	20	ACP-substrate binding
7	261-281	IYARLLGAG**LTSDAFHLVAP**	3.3e-125	20	Lipophilic pocket residues
8	336-356	VYAP**K**GALG**H**S**L**G**A**VGAVEA	2.7e-123	20	Catalytic function, enzyme stability and acyl-binding cavity
9	24-44	PLGVDVESTWKGLLAGE**S**GI	7.8e-114	20	Binding of acyl-fatty acids
10	187-207	MGDADVMVAGGVESCIDPL**P**	6.9e-113	20	Acyl-binding tunnel
11	240-260	GEGAALMVLEEEEHAKARGA	1.2e-112	20	-
12	286-306	GAARAMTRALKTAGLSPEDI	2.7e-088	20	-
13	95-115	WEPAGEPEVDPTGVAVGIGT	1.4e-070	20	-
14	69-89	DNFVS**R**VEM**RR**MSYVERMAI	4.0e-070	20	ACP-substrate binding
17	207-219	LA**AF**S**M**M**R**ALST	1.0e-051	12	Acyl-binding tunnel
18	382-394	LBVVPGEPREGK	1.6e-046	12	-
19	153-165	VVGLRYGARGGV	4.3e-032	12	-
22	356-365	VLTVLALRD	4.9e-013	9	-

Mapping of highly conserved motifs as well as motifs unique to a particular species group onto the multiple sequence alignment was also done and this is shown in Figure S3. Differences at residue level were observed upon mapping of the motifs despite the high level of motif conservation across the KasA homologs. Prokaryotic specific motifs such as motif 15 and 16 were identified and motif 21 was unique to eukaryotes. In addition, motifs 20, 24, 25, 26 and 27 were found to be unique only to mammalian species. The functional role of these motifs has not been revealed in literature, however, an investigation into these motifs could also provide new insights in understanding the mechanism of KasA proteins. In an attempt to gain a better understanding of the function of the identified unique motifs, mapping of these motifs onto the respective structures (*Mtb* KasA and *Hsmt* KasA) was done using an in-house python script in PyMOL (Figure 2.4). The individual motifs are colour-coded onto the corresponding protein structures. It can be noted however that even though the motifs are unique to a particular species at sequence level, the 3D structures of these proteins are highly similar, and this forms a basis to which more insights of drug discovery can be explored, and this is discussed in Chapter 3. All in all, motif analysis accurately identified regions of functional and structural importance in the KAS protein family such as the active site and substrate binding sites and this is consistent with previous literature.

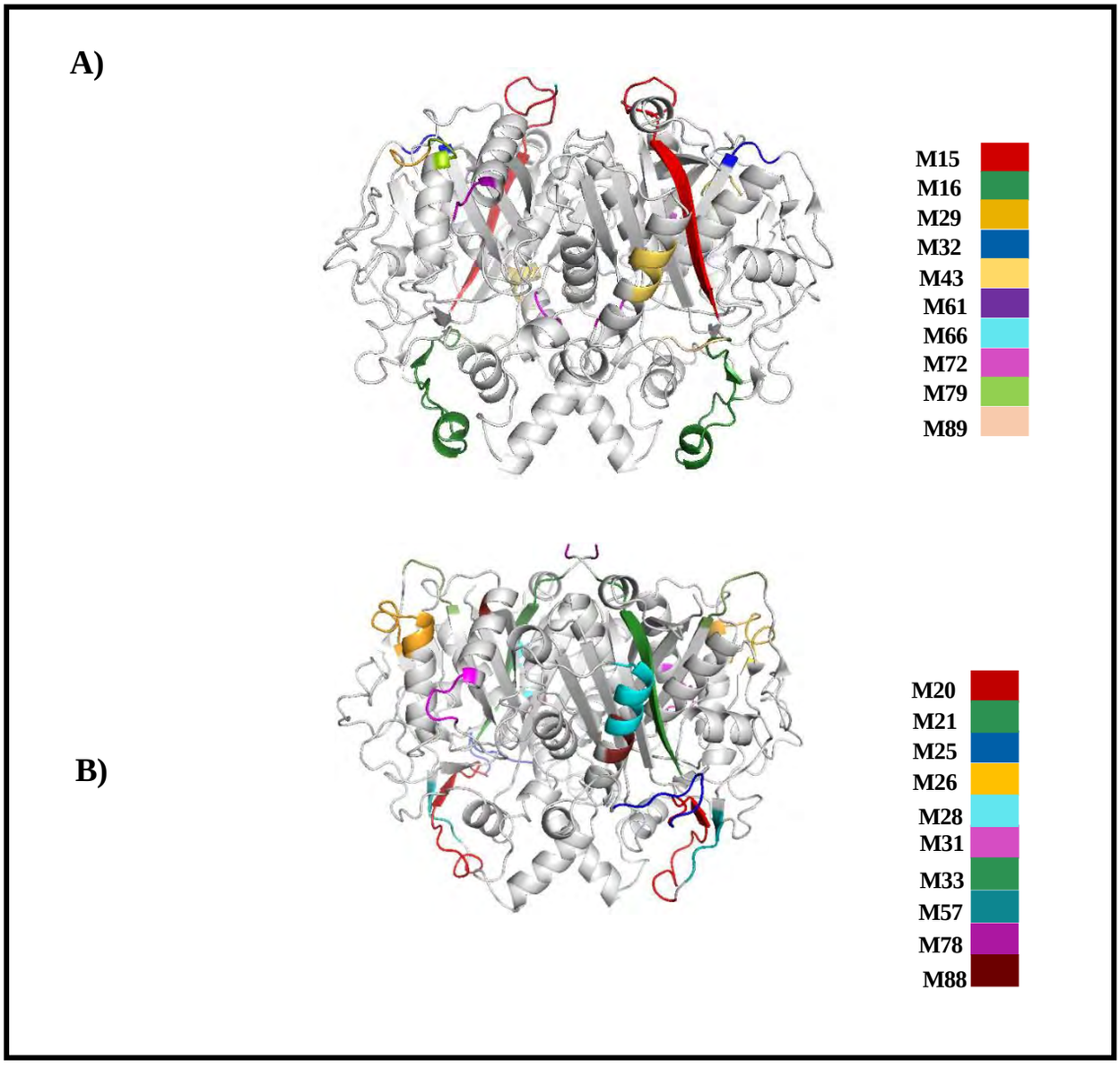


Figure 2.4: Crystallographic structures of *Mtb* KasA (PDB ID:6P9K) and *Hsmt* KasA (PDB ID:2IWY) and the identified unique motifs mapped to the respective 3D structures. A) The location of identified unique motifs on *Mtb* KasA and B) Unique motifs of *Hsmt* KasA.

2.6.4 Phylogenetic Analysis

In order to evaluate the deeper phylogenetic relationships between *Mtb* KasA and its homologs, the MEGA X software was used to infer evolutionary differences among the sequences. This was done in order to unveil the differences among the sequences which could provide insights on designing inhibitors that only target *Mtb* KasA and not the human homolog, for inhibitor selectivity purposes [122].

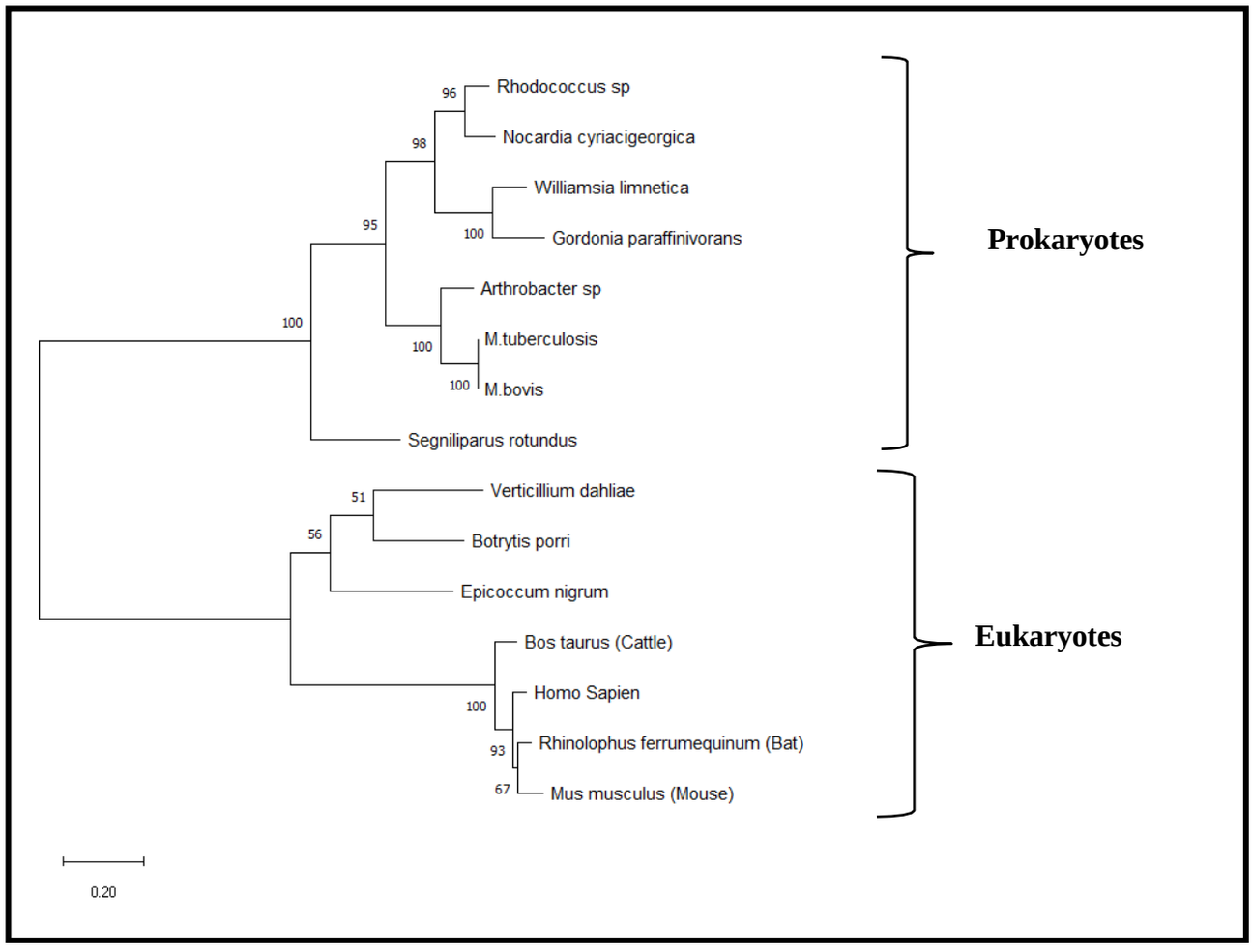


Figure 2.5: Phylogenetic analysis for 15 KasA protein sequences constructed using the MEGA X software. The Maximum Likelihood method was used to infer evolutionary relationship using the Le and Gascuel 2008 model at 100% site coverage. Initial trees for the heuristic search were obtained using the BioNJ and Neighbour-Join algorithms to a matrix of pairwise distances calculated using the JTT model.

The LG+G, LG+G+I, and WAG+G models were selected for the construction of the phylogenetic trees as these models had the lowest BIC scores. The LG+G model at 90% deletion was selected as the best tree as it had the highest log likelihood (-6180.23) and is shown in Figure 2.5. The most notable observation from the phylogenetic tree is the distinct clustering of the KasA protein sequences into two groups namely prokaryotes and eukaryotes. The bacterial species cluster together while the fungal and mammalian species also cluster together. This distinct clustering is consistent with low sequence identities observed between *Mtb* KasA and the fungal and mammalian sequence homologs. In addition, the phylogenetic tree also shows a distinct clustering of the rodent KasA (*Mus musculus*) sequence from the human analogue despite the similarities at sequence level. This clustering is vital as most in-vivo TB drug testing is done on rodent specimens. KasA proteins belonging to the same genus were observed to be clustered together, as is the case with *M. tuberculosis* and *M. bovis*. This observation is supported by the sequence alignment as they share a 100% sequence identity.

2.6.5 All-versus-all sequence identities

All vs all sequence identity calculations were assessed in order to investigate the extent of sequence conservation in all KasA sequences. This calculation was done using an in-house python script and the results were presented as a heatmap. The clustering of sequence groups observed in Figure 2.6 and the behavior of the sequences was compared to the sequence alignment and phylogeny results. The clustering of the sequences in the heatmap was consistent with the phylogenetic tree results. The magnitude of identity between sequences increases from 0, shown by a white colour to 1, indicated by a red colour. The heatmap shows high sequence identities in KasA proteins of the same species. Bacterial species share a high sequence identity (> 60%), followed by fungal species (40-45%) and lastly mammals (32-38%). In addition, it is important to note that proteins belonging to the same genus/family were clustered together, as is the case with the *Mycobacterium* sequences. These sequences share a 100% sequence identity and the heatmap visibly shows the clustering of these sequences separately from the other bacterial species from different families. A pairwise sequence comparison of a sequence to itself gave a sequence identity value of 1.

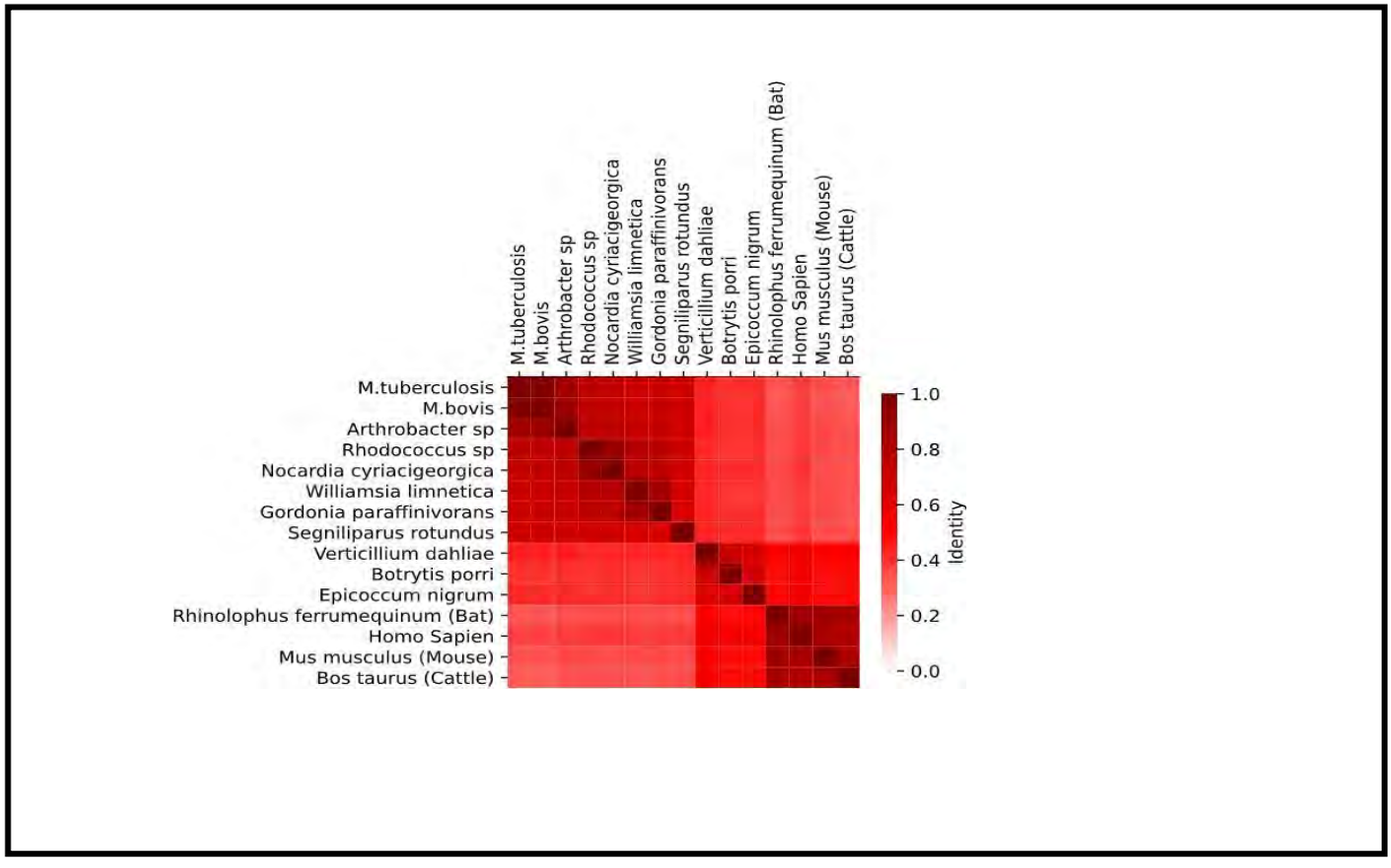


Figure 2.6: The pairwise sequence identity heatmap of the KasA proteins. The heatmap shows the pairwise sequence identity scores of the MSA as a colour-coded matrix. Identity scores increases from white to red (least to most conserved sequences) in the heatmap.

2.7 CHAPTER CONCLUSION

This chapter provides an in-depth sequence and structural analysis of the KAS (Keto-acyl-ACP synthase) proteins. Fifteen protein sequences were retrieved from the UniProt database from a diverse range of species namely bacteria, fungi and mammals. The sequences identities and coverage of the homolog sequences with respect to the query sequence (*Mtb* KasA) revealed that the sequences were true homologs (sequence identity > 30%). Multiple sequence alignment of these protein homologs revealed the high conservation of residues that infer functionality of the protein such as active site residues. MSA also identified an N-terminal insert in the mammalian sequences, with particular reference to *Hsmt* KasA, that has characteristics of a mitochondrial transit peptide. Motif analysis was employed in order to identify conserved sequence patterns within proteins of the KAS family as well as to investigate whether these motifs played an important role in the structure and function of the proteins. Eighteen highly conserved motifs were identified across the homologs and these were explored at residue level for protein characterization. In addition, the HTH region of the KAS proteins was also identified and is associated with Motif 6. Motifs unique to *Mtb* KasA were also identified and these could serve potential as targets for allostery and drug repurposing. Phylogenetic analysis highlights the evolutionary relationships and distinct clustering of the prokaryotic-based sequences from the eukaryotes. The results of this clustering are also consistent with the all-versus-all sequence identity heatmap observations.

In the next chapter we explore the identification of potential allosteric sites on the *Mtb* KasA protein by use of various allosteric site search tools in order to further explore these sites for structure-based drug design as well as to further validate the suitability of *Mtb* KasA as a potential drug target.

CHAPTER THREE

ALLOSTERIC SITE IDENTIFICATION

The mechanism of action used by most drugs involves altering the action of enzymes, receptors and transporter molecules by binding directly to the orthosteric sites, commonly known as the active sites. These drugs have been designed to mimic the body's natural substrates and binding of these molecules can either result in an activation or deactivation of the receptor [123]. However, an insight into the binding of drug molecules to sites away from the orthosteric sites is believed to be a step that can potentially revolutionize drug discovery. Allostery is an approach that involves the binding of a potential drug molecule (termed allosteric modulator) to an allosteric binding site, altering the conformation of the active site of the protein [124]. The use of allosteric modulators in drug discovery have advantages that include increased target specificity, selectivity as well as the enabling the design of therapeutic agents with fewer side effects [125, 126].

In this study, the 3D structures of *Mtb* KasA and *Hsmt* KasA were explored for potential allosteric sites using various allosteric site identification tools. The tools used included CavityPlus [127], AutoLigand [128], Protein Plus DoGSiteScorer [129] and SiteMap [130, 131]. The above-mentioned tools employ different detection algorithms in the identification of probable allosteric sites, including but not limited to the following: structural geometry, machine learning, grid-based and energy-based methods. This chapter focuses on the discovery of potential allosteric sites on the *Mtb* KasA protein and its homolog, *Hsmt* KasA by utilizing several pocket detection tools as well as exploring these sites for inhibitory drug design purposes.

3.1 INTRODUCTION

3.1.1 Overview of Allostery

Allostery is a key regulatory cellular process that involves modifying the nature of a biological target (DNA or protein). In relation to proteins, this process involves altering the conformation of the protein at the active site or other sites as a result of distant perturbation of the protein [132]. The term *allosteric* originated from the Greek word *allos*, meaning other, and *steric*, which refers to the spatial arrangement of atoms in a molecule [1 3 3] . Simply put, allostery is a change in shape. Allosteric perturbation is not only limited to the binding of a ligand molecule but can also arise as a result of other non-covalent modifications such as a change in pH, temperature and concentration [134]. In addition, covalent alterations such as phosphorylation, glycosylation, ubiquitination [135] and point mutations also result in allostery [132, 136]. Allosteric modulators can either function as a positive allosteric modulator (PAM), a negative allosteric modulator (NAM) or as a neutral allosteric modulator (NAL). PAMs increase the binding affinity of the ligand to the orthosteric site, which in turn results in boosted activity. NAMs result in a decreased affinity for the agonist to the receptor, thus leading to a decrease in activity. NALs work by binding to a receptor's allosteric site but do not introduce any change to the behavior of the receptor or the orthosteric ligand [123, 137]. It must be noted that the NAL group has a very small number of molecules and hence it is not commonly used.

The allosteric behavior of multimeric proteins is best explained by two main models: the concerted Monod-Wyman-Changeux (MWC) model and the sequential Koshland-Nemethy-Filmer (KNF) model [138]. The MWC model hypothesizes that allosteric oligomeric proteins exist in two interchangeable states, *T* or *R*, that are in thermal equilibrium, with all the subunits in the protein either adopting the *T* state or the *R* state. The R state, which is also known as the active/relaxed state has a higher free energy and in the presence of an allosteric ligand, the free energy is lowered, and this enables the ligand to bind tightly to the protein [134]. In addition, the ligand affinities for the orthosteric and allosteric sites vary between the two states, thus permitting preferential binding. The KNF model on the other hand postulates that in the absence of a ligand, the protein is found in a single state, usually the *T* state. Ligand binding prompts a conformational change to the subunit that it binds to, which in turn induces a conformational change in the neighboring subunits [139]. This model represents the theory of induced fit, by which binding of substrates to the active site result in a conformational change in the residues constituting the active site, which are essential

for the protein's functioning. Other models that have been used to substantiate the hypotheses of the two models described above include the Population Shift Model, Morpheein Model as well as the Dynamically Driven Model [140].

3.1.2 Properties of Allosteric Proteins

Over the decades, various experimental techniques have been used to characterize allosteric proteins. These include X-ray crystallography, Nuclear Magnetic Resonance (NMR), Fluorescence and Hydrogen-deuterium exchange mass spectrometry [134]. These techniques have given rise to high resolution structures which have been used to determine the structural characteristics of allosteric proteins. According to a study conducted by Li et al., it was reported that allosteric proteins exist as monomers and even-numbered multimers such as dimers, tetramers, hexamers etc. [141]. Moreover, predicated on the sequence compositions, it was found that allosteric sites are more hydrophobic than active sites due to the proportion of hydrophobic residues that make up the sites. Charged residues such as lysine, histidine, glutamic acid and aspartic acid are usually found in the active sites whereas residues such as proline, tryptophan, leucine, isoleucine, valine and methionine are largely concentrated in the allosteric sites. Hydrophobic residues provide a binding pocket for the ligand; hence these residues usually make up the allosteric sites. Hydrophilic residues on the other hand are usually involved in hydrogen bond formation to facilitate bond formation or breakage [142]. Another characteristic feature of allosteric proteins is that ligand binding sites, both active and regulatory are located at subunit interfaces. This enables communication between subunits as sites at the interface of subunits respond to the alterations in the interactions of the subunits, which is a key step in allosteric regulation. Ligand binding sites can also be located between two domains in the same subunit. This is characteristic of the active site in allosteric proteins, which can change the size and shape of the protein due to induced conformational changes.

Yang et al., discovered that allosteric sites are less conserved than active sites [143]. This is because allosteric sites allow for ligand binding but are not involved in the catalytic conversion of the ligand. Active sites on the other hand are highly conserved as they are involved in the binding and conversion of substrates, therefore mutations of the active site residues usually result in the loss of catalytic function of the enzyme. Allosteric binding sites have been found to be located in the low-stability regions of the protein. To support this notion, it has been noted that binding of a ligand to these

regions stabilizes the protein, with respect to salt bridges and hydrogen bonds formed at the subunit interfaces.

3.1.3 Benefits of Allosteric Drugs

Allosteric drugs have several advantages over orthosteric drugs that target the functional site of the protein. Firstly, they are highly specific as they do not bind to the active site, which is highly conserved in protein families, but tend to bind to sites away from the active site. This is particularly advantageous in drug design as this lowers the chances of side effects that most orthosteric drugs are associated with. Most drugs that bind to the orthosteric sites of one protein are also capable of binding to the orthosteric sites of the homologous protein family. Secondly, allosteric drugs can activate a target protein directly or indirectly. For instance, the binding of a ligand to one receptor molecule's subunit can allosterically modulate the response of another subunit to a ligand, thereby creating a mechanism of specificity [132]. In addition, allosteric drugs allow for the modulation of a protein's activity without completely terminating it. This does not hold true for orthosteric drugs that stop the protein's activity entirely [144]. Allosteric modulators do not compete with the endogenous ligand; therefore, they can impact their influence even in the presence of an endogenous ligand that is bound to a different site on the same target protein, hence acting like co-factors [145]. However, due to the challenges posed by drug resistance, a combination of allosteric and orthosteric drugs can be beneficial.

3.1.4 Identification and Characterization of Allosteric Inhibitory Sites

The identification of allosteric sites in a protein is the first and important step in allosteric drug discovery. A vast number of allosteric sites have been identified using various biochemical experiments such as high throughput screening, disulfide trapping, X-ray crystallography as well as fragment-based screening [146]. Although these methods have successfully identified allosteric sites, the vast increase in the number of allosteric drug targets has made it tedious to detect potential allosteric sites. In an effort to identify allosteric sites more efficiently and effectively, a number of *in-silico* methods have been used to provide platforms to identify these sites based on sequence, structure and dynamics. In this study, allosteric site prediction was achieved by use of a combination of prediction tools that employ different algorithms in the identification of potential allosteric sites based on druggability of the sites. CavityPlus uses the protein's 3D structural information to detect potential binding sites on the surface of the protein. AutoLigand, a site detection tool of the AutoDock Tools package uses a grid-based representation to identify fill

points, which are the proposed binding sites on the protein. Protein Plus DoGSiteScorer uses a support vector machine (SVM) technique to detect the grid points used to characterize the binding site. SiteMap, a package of the Schrödinger Suite, identifies ligand binding sites using the Goodford's GRID algorithm [147]. In order to improve accuracy in the allosteric site search, the consensus regions identified by the combination of the different algorithms were used as a basis to state that these regions were most likely allosteric sites.

3.2 METHODOLOGY

Figure 3.1 shows the steps used in allosteric site prediction in *Mtb* KasA and *Hsmt* KasA.

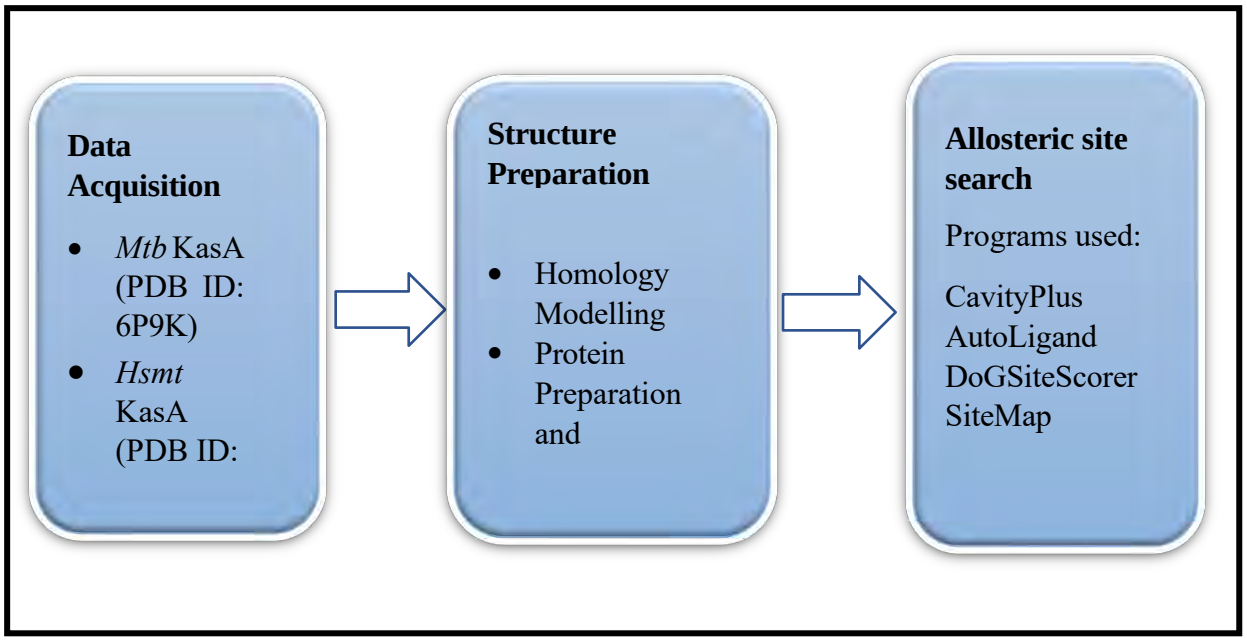


Figure 3.1: Detailed workflow of the analytic approaches used in this study. A flowchart showing the sequence of steps used in this analysis, starting with retrieval of protein structures through protein preparation and lastly performing allosteric site search using various pocket detection tools.

3.2.1 Data Acquisition

Binding cavities on a protein's surface are essential for protein function because these are the regions at which ligands bind to the protein, either to the allosteric site or active sites. Therefore, a target protein of interest is required in order to identify these sites. The 3D structures of *Mtb* KasA and *Hsmt* KasA were retrieved from the Protein Data Bank using the PDB IDs: 6P9K and 2IWY respectively. The 3D structure of the wild type *Mtb* KasA protein has a resolution of 1.70 Å, a co-crystallized sulfonamide inhibitor and is made up of two chains (A and B) with 414 residues in each chain. Glycerol, isopropyl alcohol and sodium ions were crystallized with the protein. The wild type *Hsmt* KasA crystal structure has a resolution of 2.06 Å and consists of two chains of 438 residues each. In addition, ammonium ions were also crystallized with the protein. This structure had no co-crystallized ligand and had missing side chain atoms. Both structures were prepared using X-ray crystallography.

3.2.2 Structure Preparation

3.2.2.1. Homology Modelling

In order to fill in the missing side chain atoms in the *Hsmt* KasA crystal structure, homology modelling was performed using Schrödinger Prime [148, 149]. To initiate the process, homology modelling was performed using the protein's 3D crystal structure as a template. ClustalW was selected as an alignment tool to align the query and template sequences as it is suitable for instances where there is a high sequence identity between the query and the template. After alignment, the unaligned first 16 residues were trimmed off as the template did not cover these residues.

Before the model building process began, all ligands, cofactors and waters were deleted. A knowledge-based model building method was selected for model construction and a total of five models were then calculated using MODELLER v.9.19 [150]. The Homo-multimer multi template model type was used to build the query sequence on each of the selected templates. This allowed for the models to be built with each chain being constructed and refined in the presence of others.

To validate the model and to assess the accuracy of the model building process, z-Dope scores for each model were calculated and ranked according to the best score (lowest score). Positive values depict poor quality models whereas scores less than -1 are an indication of a good quality model, similar to the native protein. All the models generated had z-Dope scores lower than -1, however,

the model with the lowest score was then selected for further analysis.

3.2.2.2. Protein Preparation and Protonation

The 3D structures of *Mtb* KasA and *Hsmt* KasA were prepared using the Protein Preparation Wizard of Maestro version 12.9 of the Schrödinger Suites software [151]. This process is divided into three main stages: preprocessing the structure, optimizing the hydrogen-bond network and minimizing the structure. To kickstart the process, the target protein's 3D structure was imported into Maestro using the PDB ID. In the preprocessing step, unwanted groups from the protein's structure were removed. This included the co-crystallized ligand, ions and other molecules that were added to help crystallize the structure. In addition, hydrogens were added to all atoms that originally had none and waters were removed from the structure. This step also involved the assigning of bond orders as well as the creation of disulfide bonds and zero-order bonds to metals. All other parameters were kept at default settings. Overlapping atoms and atoms that had alternate positions in the structure were also removed.

Optimization of the hydrogen-bond network involved reorienting hydroxyl and thiol groups, water molecules, amide groups as well as the imidazole ring [152]. Furthermore, this step also predicts the protonation states of histidine, aspartic acid and glutamic acid. Tautomeric states of histidine are also generated. The final step was the minimization of the protein structure, which involved optimizing the positions of the hydrogen atoms in order to avoid steric clashes in the protein. The protein structures were then protonated at a neutral pH (pH 7.0) using PROPKA [153]. This allowed for the generation of pKa values for the protein residues at the specified pH.

3.2.3 Pocket Analysis and Allosteric Site Search

3.2.3.1 CavityPlus

The protonated protein structures of *Mtb* KasA and *Hsmt* KasA (excluding ligands, ions and waters) were uploaded onto CAVITY, a submodule of CavityPlus [127]. The program starts by importing the protein's co-ordinates from a PDB file and performing a validity check on the number of chains in the protein's structure. This was followed by the generation of a 3D grid on the surface of the protein. This computation outputted the cavities identified by the server, alongside properties that define the identified cavities such as predicted maximal pKd, drug score and druggability. Only cavities with a medium to strong druggability were considered and were used in subsequent calculations.

To identify the allosteric sites, the cavities discovered by the previous computation were used with another submodule of CavityPlus, CorrSite [154]. This was achieved by firstly selecting a known orthosteric site from the cavities. Since the protein structure has two monomers with an orthosteric site on each monomer, the second orthosteric side was excluded from the computation by specifying it according to the cavity it corresponded to from the CAVITY results. Cavities that had overlapping residues with the orthosteric sites were excluded and residues that were shared with the active site were removed. This computation resulted in the generation of Z-scores which were used to identify probable allosteric sites.

3.2.3.2 AutoLigand

The prepared protein structures without the ligands, ions and waters were imported into AutoDock Tools (ADT). This was followed by adding polar hydrogens, merging nonpolar hydrogens, adding gasteiger charges as well as assigning AutoDock type atoms to the protein structure by use of a 'prepare_receptor4.py' script in ADT. In order to determine affinity potentials, default AutoGrid parameters were used [155]. The grid box was centered on the receptor protein and set to cover the entire protein by using a 1Å grid spacing. The grid box size for the x, y and z dimensions were set to 70 x 60 x 84 Å respectively for the *Mtb* KasA structure while the dimensions for *Hsmt* KasA was set to 82 x 72 x 84 Å respectively (Figure 3.2). The output of this computation was saved in a grid parameter file format for use with AutoLigand. In order to generate the flood-fill points, an in-built python script of the AutoDock Tools named AutoLigand.py was used with default settings.

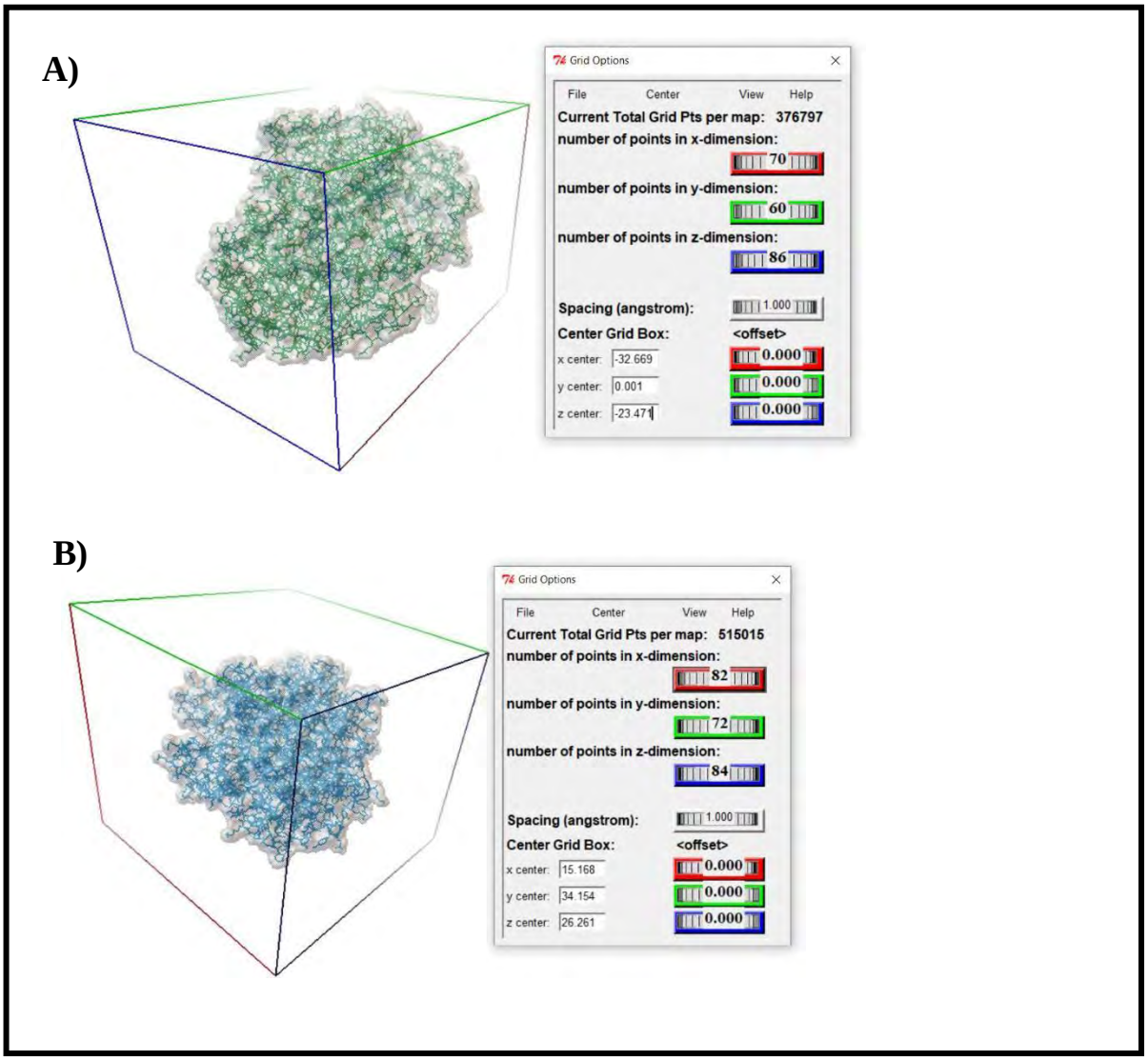


Figure 3.2: Grid box dimensions set to cover the entire protein for binding site search in AutoDock Tools**.** A) *Mtb* Kas A protein represented as a transparent surface and the atomic level shown in green. B) *Hsmt* KasA protein shown as a transparent surface with atomic detail of the protein in blue. The x-,y-,z- centers and dimensions parameters as well as the spacing used in grid generation are shown adjacent to the grid for both proteins.

3.2.3.3 Protein Plus DoGSiteScorer

Allosteric site search in DogSiteScorer was initiated by uploading the protonated structures of *Mtb* KasA and *Hsmt* KasA onto the webserver in PDB format. A validity check was then performed to evaluate the reliability of the 3D structures supplied. Both chains of the protein were selected, and the program was set to calculate pockets and sub-pockets in the protein structure. The output of

this computation included physicochemical properties such as volume, surface area, lipophilic surface area, depth as well as druggability for each pocket identified.

3.2.3.4 SiteMap

Binding site prediction in SiteMap was initiated using the protonated *Mtb* KasA and *Hsmt* KasA structures. SiteMap was set to identify the top ranked potential receptor binding sites. In addition, the program was set to identify at least 15 site points per binding site as well as detecting shallow binding sites. A default standard grid and a more restrictive hydrophobicity were used as search parameters. Sites that had a SiteScore > 0.8 were selected for further analysis.

3.3 RESULTS AND DISCUSSION

3.3.1 CavityPlus results on *Mtb* KasA and *Hsmt* KasA

Cavity, a submodule of CavityPlus identified a total of 22 possible cavities in both *Mtb* Kas A and *Hsmt* KasA. However, four of these cavities were classified as 'druggable' in *Mtb* KasA whereas only three cavities were druggable in *Hsmt* KasA (Table 3.1). The selection criteria for these cavities included the predicted maximal pKd, drug score and druggability. A predicted pKd value less than 6 indicates that the site identified is not suitable as a binding site. Cavity detection was performed with protein structures with no ligands bound to them. This is because ligand binding causes a conformational change in the protein structure that could bias the binding site detection process. Two key terms are utilized by the algorithm during protein cavity detection and these are 'ligandability' and 'druggability'. Ligandability refers to the possibility of designing small ligands with high binding affinity to a certain binding site whereas the latter refers to the possibility of a cavity being a good target for drug-like molecules [156]. Only sites that had medium to strong druggability were considered for further analysis.

Table 3.1: Output of the CAVITY module on *Mtb* KasA and *Hsmt* KasA.

Protein	Cavity	Predicted maximal pKd	Drug score	Druggability
***Mtb* KasA**	1	11.36	2450.00	Strong
	2	11.56	2418.00	Strong
	3	9.76	861.00	Strong
	4	9.42	864.00	Strong
***Hsmt* KasA**	1	10.46	293.00	Medium
	2	10.34	108.00	Medium
	3	10.64	-71.00	Medium

The Cavity drug score is influenced by factors such as cavity volume, hydrophobic volume, cavity surface area as well as hydrogen-bonding surface area. Cavities 1 and 2 on *Mtb* KasA were identified as potential allosteric sites whilst cavities 3 and 4 constitute residues of the active site. In *Hsmt* KasA, cavities 1 and 2 were identified as the active site cavities whilst cavity 3 is a potential allosteric site. Cavities 1 and 2 in *Mtb* KasA constitute hydrophobic residues that make up the acyl-binding pockets which are connected to the active site via an acyl-binding tunnel. However, it is interesting to note that the program failed to detect these cavities in *Hsmt* KasA, thus bringing the reliability of the program in accurately identifying cavities of structural and functional importance into question.

The cavity drug scores calculated for *Mtb* KasA are higher than those of *Hsmt* KasA. This provides a good foundation for structure-based drug design as inhibitors are designed based on selectivity. The druggability of the identified sites is strong in *Mtb* KasA and moderate in the human homolog protein. This suggests that *Mtb* KasA is more likely to be therapeutically modulated by a drug compound than *Hsmt* KasA. This is important because it enables the design of novel compounds with a high affinity for the pathogen, thus minimizing the risk of adverse effects in humans. Figure 3.3 shows the identified druggable cavities on the structures of *Mtb* KasA and *Hsmt* KasA.

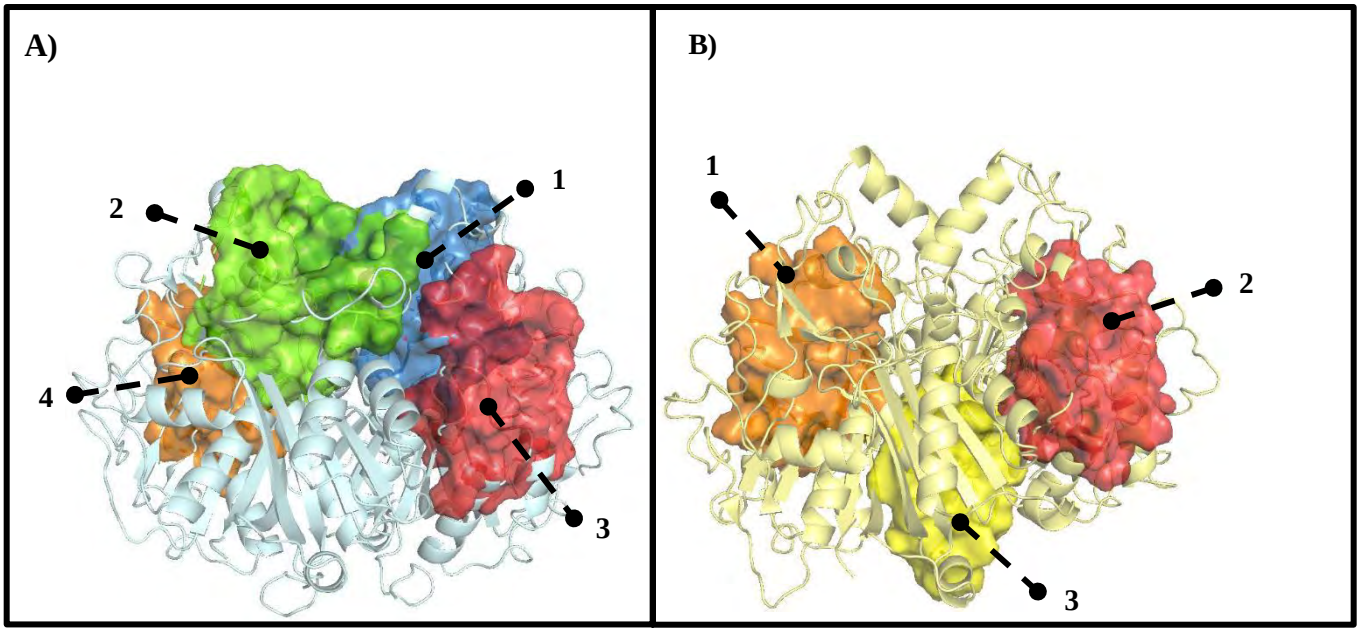


Figure 3.3: Detected cavities in *Mtb* KasA and *Hsmt* KasA by CavityPlus. A) *Mtb* KasA represented as a cartoon in pale cyan and the identified cavities shown as a closed surface numbered according to the Cavity output and colored in blue, green, red and orange respectively. B) *Hsmt* KasA represented as a cartoon in pale yellow and the identified cavities shown as a closed surface in orange, red and yellow respectively.

In order to validate the Cavity output, Z-scores were generated for each site using CorrSite. Values greater than 0.5 suggest that the cavities are potential allosteric sites. All the predicted cavities used for this calculation (orthosteric sites excluded) had Z-Scores greater than 0.5 as shown in Table 3.2. The Z-scores for the cavities identified in *Mtb* KasA are higher than those of *Hsmt* KasA. These sites are of interest as they do not compete with the endogenous substrate during catalysis.

Table 3.2: Predicted Z-scores of the potential allosteric sites identified by CorrSite.

Protein	Cavity	Z-Score
***Mtb* KasA**	1	1.97
	2	1.95
***Hsmt* KasA**	3	1.27

3.3.2 AutoLigand results on *Mtb* KasA and *Hsmt* KasA

AutoLigand identified 10 binding sites on both the *Mtb* KasA and *Hsmt* KasA protein. The FILL files with the co-ordinates of the residues were used to characterize the binding sites. What is interesting to note is that most of these sites were clustered together upon visualization. In *Mtb* KasA, fills 7 and 10 clustered together to form pocket 4, fills 2,4 and 5 cluster to form pocket 2, fills 1,3 and 6 form pocket 1 and fills 8 and 9 form pocket 3 (Figure 3.4). Pockets 3 and 4 in *Mtb* KasA are the active site pockets and pockets 1 and 2 are potential allosteric sites. However, in *Hsmt* KasA, an interesting observation is made when fill 5 forms a stand-alone pocket 4 while the rest cluster together to form separate pockets. This pocket consists of hydrophobic residues that make up the acyl-binding tunnel. It is also important to note that while the program identified the acyl-binding tunnel, it failed to accurately identify the acyl-binding pockets located at the surface of the protein in *Hsmt* KasA, thus raising concerns on the reliability of the algorithm used. The active site pockets, pocket 1 and 2 are formed by the clustering of fills 7 and 10 as well as fills 2, 3 and 4 respectively. Pocket 3 is formed by the clustering of fills 1, 9, 6 and 8 and this is the largest pocket identified by the algorithm.

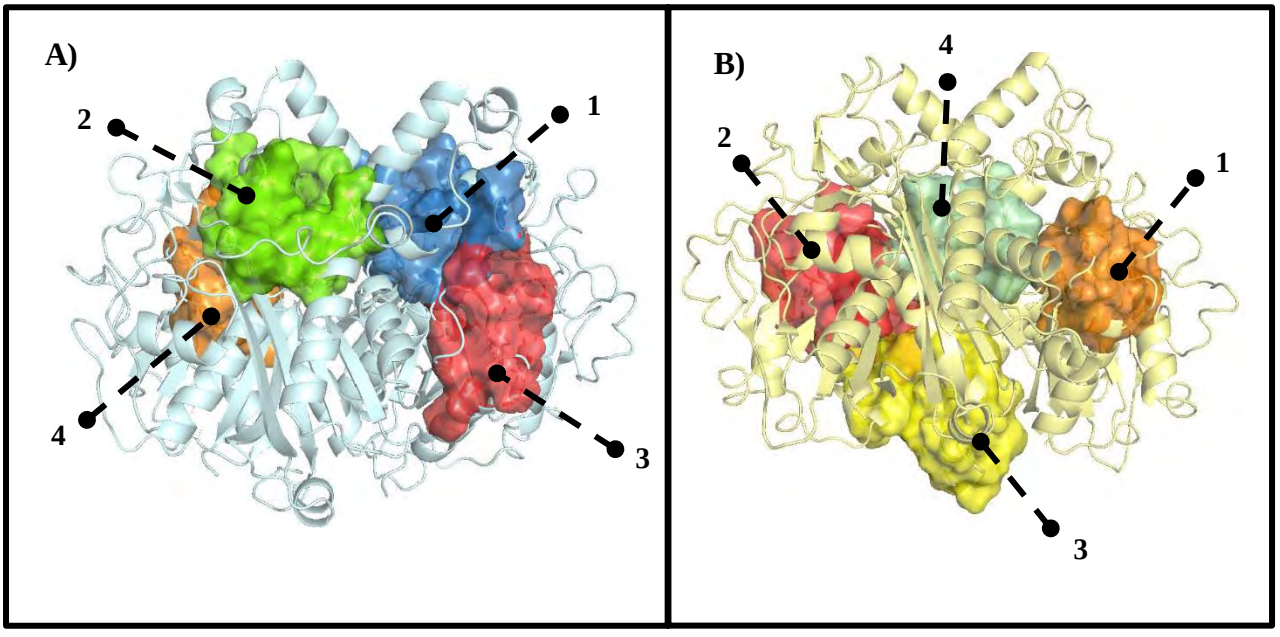


Figure 3.4: Detected cavities in *Mtb* KasA and *Hsmt* KasA by AutoLigand. A)*Mtb* KasA represented as a cartoon in pale cyan and the identified cavities shown as a closed surface in blue, green, orange and red. B) *Hsmt* KasA represented as a cartoon in pale yellow and the identified cavities shown as a closed surface in orange, red, yellow and pale-green respectively.

3.3.3 DoGSiteScorer results on *Mtb* KasA and *Hsmt* KasA

DogSiteScorer identified 18 probable binding pockets on both *Mtb* KasA and *Hsmt* KasA. The identified pockets were characterized by various geometric and physicochemical parameters such as pocket volume, surface area, lipophilic character as well as the pocket enclosure. The simple score and drug score were used to rank the pockets. Out of the 18 pockets identified by the algorithm, only seven pockets were found to be druggable in *Mtb* KasA whereas only four druggable pockets were found in *Hsmt* KasA. Pockets were only deemed druggable if the druggability score was greater than 0.5. Figure 3.5 shows the identified pockets on the 3D structures of *Mtb* KasA and *Hsmt* KasA.

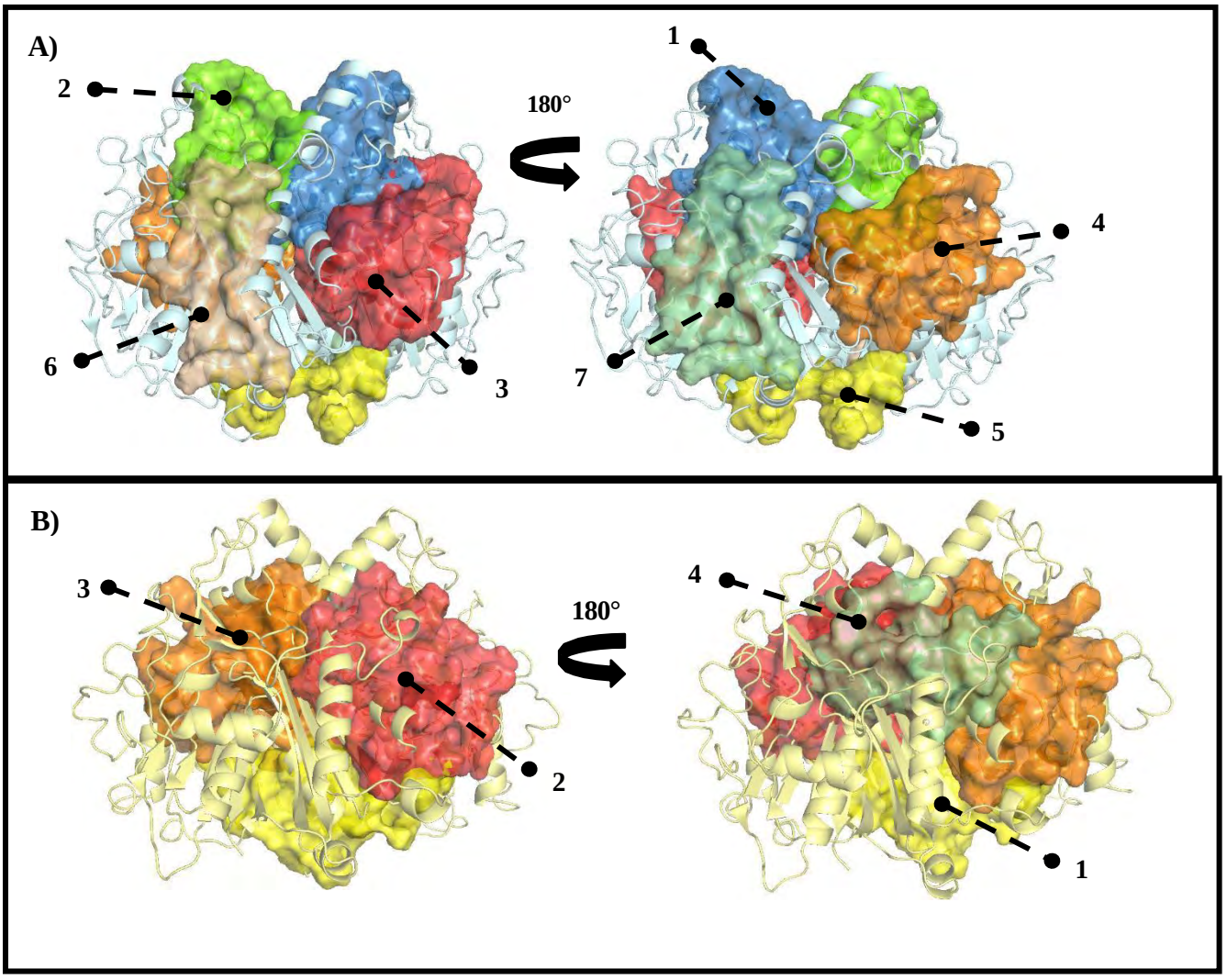


Figure 3.5: Detected cavities of *Mtb* KasA and *Hsmt* KasA by DogSiteScorer**.** A)*Mtb* KasA represented as a cartoon in pale cyan and the identified cavities shown as a closed surface and numbered accordingly. B) *Hsmt* KasA shown as a cartoon in pale yellow and the identified cavities shown as closed surfaces.

Pockets with a greater volume and surface area had a higher drug score and simple score compared to those with a lesser volume and surface area. What was interesting to note however is that some pockets had a relatively low simple score and a reasonably high drug score. This was the case with pockets 5, 6 and 7 in *Mtb* KasA and pocket 4 in *Hsmt* KasA (Table 3.3). The reason for this is that pocket volume is an important descriptor and indicator of druggability. Drugs are known to bind at large surface cavities and large pocket volumes increase the chances of finding the ligand in the probable pocket. In both cases, the algorithm correctly identified the active site pockets namely pockets 3 and 4 in *Mtb* KasA and pockets 2 and 3 in *Hsmt* KasA. An interesting observation made was that although the algorithm accurately identified the acyl-binding pockets in *Mtb* KasA (Pockets 1 and 2), it failed to identify these pockets in *Hsmt* KasA. However, the program correctly identified a connecting tunnel (Pocket 4) that serves as a shuttle between the active site and the acyl binding site. It is important to note that although most endogenous ligands are known to interact with the active sites, these pockets have a slightly lower drug score compared to other potential allosteric sites identified. This result allows us to explore allosteric drug design in an effort to develop therapeutic agents that have a different mechanism of action compared to the conventional orthosteric drug design.

Table 3.3: DogSiteScorer identified pockets and main pocket descriptors for the input structures *Mtb* KasA and *Hsmt* KasA.

Protein	Pocket	Volume(Å^3)	Surface Area(Å^2)	Drug score	Simple score
***Mtb* KasA**	1	1366.59	1224.69	0.82	0.63
	2	1126.00	1051.47	0.81	0.66
	3	880.96	1112.74	0.84	0.57
	4	778.71	909.89	0.85	0.48
	5	449.95	864.91	0.71	0.31
	6	353.88	618.17	0.87	0.15
	7	351.58	611.47	0.86	0.16
***Hsmt* KasA**	1	834.54	920.26	0.84	0.51
	2	894.54	766.57	0.82	0.52
	3	892.15	822.13	0.84	0.55
	4	478.62	512.50	0.83	0.2

3.3.4 SiteMap results on *Mtb* KasA and *Hsmt* KasA

SiteMap identified the top five sites on both *Mtb* KasA and *Hsmt* KasA via a combination of properties calculated at each site point. These properties include the site score (S-score), drug score (D-score), size and volume. In addition, physicochemical parameters such as hydrophobicity, hydrophilicity, enclosure as well as hydrogen bond donors and acceptors were also used to characterize the pockets. SiteMap employs a S-score threshold of 0.80 to discriminate sites that bind ligands from those that do not. The druggability of the site is referred to as the D-score and this is used to categorize sites as being "druggable", "undruggable" or "difficult to drug" [130, 131]. D-score incorporates terms that promote ligand binding such as hydrophilicity and degree of enclosure. Using the D-score, sites with a score value less than 0.83 are deemed undruggable, those with a D-score in the range 0.83<D-score<0.98 are difficult to drug and those with a D-score greater than 0.98 are druggable. Table 3.4 shows the top 5 sites identified by SiteMap in both *Mtb* KasA and *Hsmt* KasA.

Table 3.4: Binding sites identified by SiteMap on *Mtb* KasA and *Hsmt* KasA.

Protein	Sites	S-score	D-score	Size	Volume
***Mtb* KasA**	1	0.869	0.988	476	289.492
	2	0.892	0.994	358	302.869
	3	0.973	1.086	710	582.412
	4	0.991	1.098	702	536.412
	5	0.895	1.111	2186	1342.501
***Hsmt* KasA**	1	0.909	1.024	559	379.701
	2	0.911	1.006	526	443.156
	3	0.899	1.01	2027	1134.301
	4	0.898	1.018	1237	834.519
	5	0.918	0.998	658	568.694

From these results, it can be clearly observed that all the sites identified in *Mtb* KasA and *Hsmt* Kas A have an S-score greater than the threshold value of 0.8. This indicates that these sites are indeed potential ligand binding sites. In addition, the D-score also reveals that all the sites identified by the algorithm in both proteins are druggable. However, it is important to note that the D-score is influenced by the size and volume of the identified sites. Sites that had a greater size and volume in *Mtb* KasA (Sites 3,4 and 5) also had corresponding high druggability scores. What

was interesting to note however is that in *Hsmt* KasA, contrary to size and volume, Site 1 had the highest druggability score despite the relatively small size and volume when compared to other sites. This indicates that size is not overtly a factor in the Cheng approach used by SiteMap because it takes into consideration the ratio of non-polar residues rather than the nonpolar surface itself and thus scales all sites to a common size [130]. Figure 3.6 shows the sites identified mapped onto the respective protein structures.

SiteMap identified the greatest number of residues constituting a binding site compared to the other algorithms, as shown by the area of coverage of the binding sites on the respective protein structures. In *Mtb* KasA, site 5 has the greatest D-score and is the largest site that was identified. In *Hsmt* Kas A, site 3 was the largest site identified, however site 1 had the largest D-score. These results show that in *Hsmt* KasA, the active site (Site 1) has a greater druggability compared to other potential allosteric sites and this is also true for sites 3 and 4 in *Mtb* KasA. In addition, the druggability of the potential allosteric sites is higher in *Hsmt* KasA compared to *Mtb* KasA. This provides useful insights in competitive inhibitor drug design as drugs that bind selectively to the *Mtb* protein are preferred over the drugs that have a similar binding affinity to the human homolog. All in all, this algorithm correctly identified the key functional pockets (active site and acyl-binding site) in both *Mtb* KasA and *Hsmt* KasA.

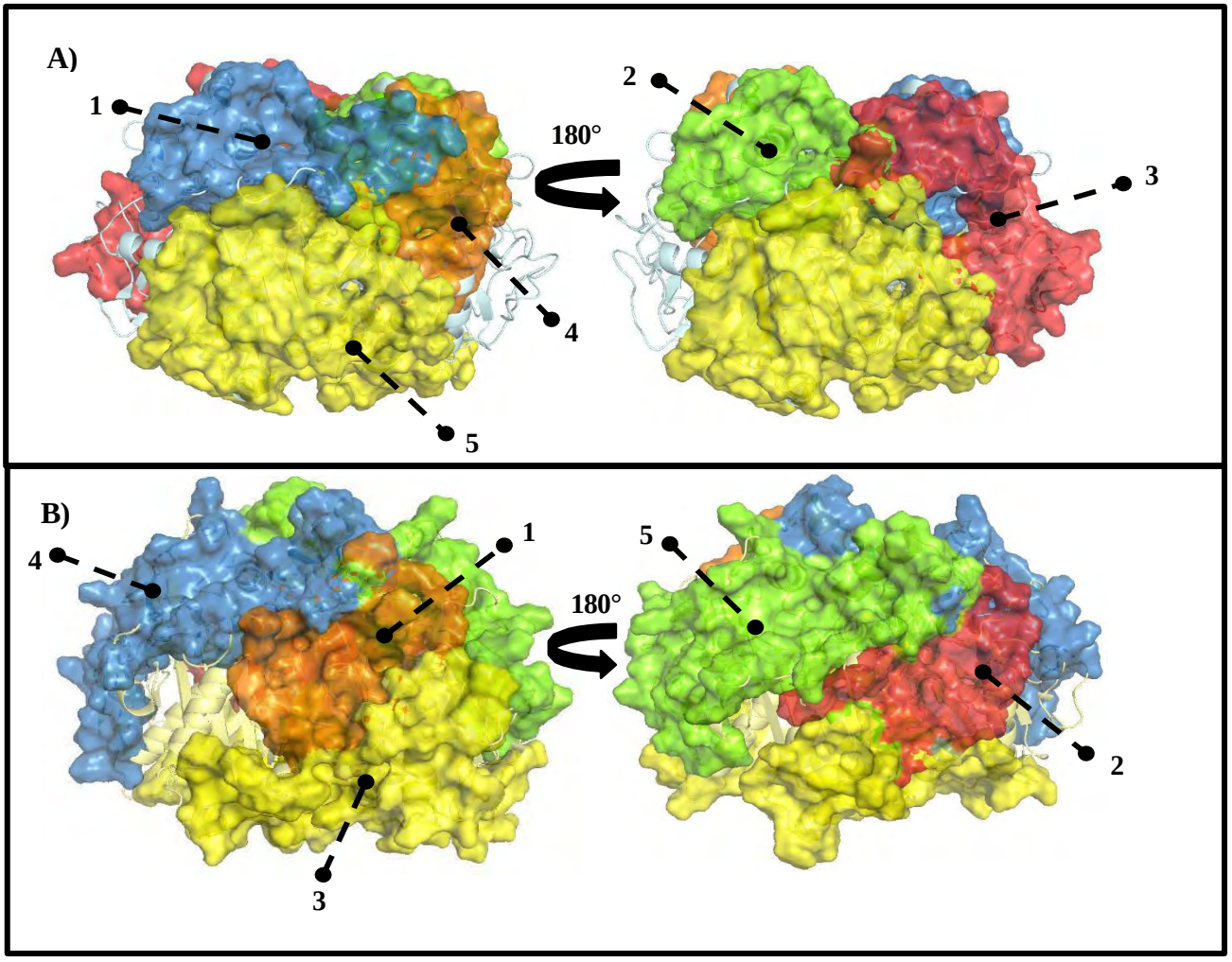


Figure 3.6: Binding sites identified by SiteMap on A) *Mtb* KasA and B) *Hsmt* KasA.

3.3.5 Validation of the pocket detection algorithms

In order to validate the protocol used by the various webservers in the detection of putative pockets in *Mtb* KasA and *Hsmt* KasA, a consensus prediction of the druggable sites was used. This involved identifying the pockets that were detected in at least 3 of the pocket detection tools used. This validation is important because the different pocket detection tools used in this study employ different algorithms, and hence discovering the residues making up the consensus sites would be a useful and unbiased method of characterizing the probable allosteric pockets in the respective protein structures. An *ad hoc* python script was used to assess the residues that constitute the consensus sites, and the result of this analysis yielded a total of 2 potential allosteric sites in *Mtb* KasA and only 1 potential allosteric site in *Hsmt* KasA as shown in Figure 3.7. The active site

pockets were excluded from this result.

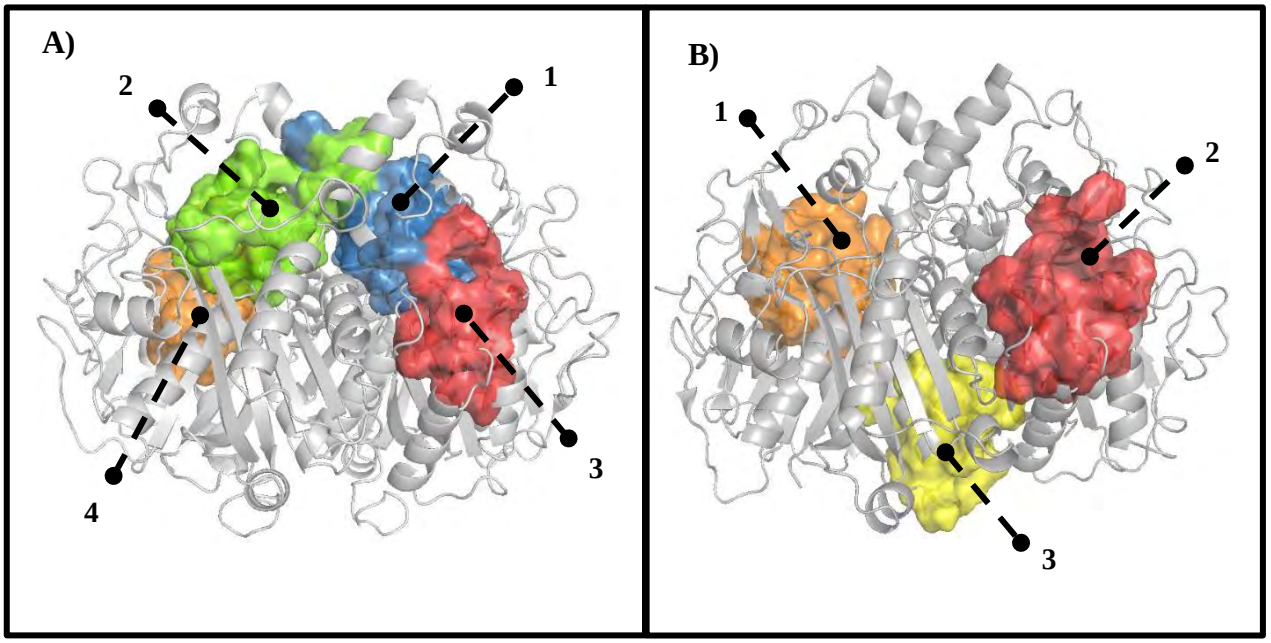


Figure 3.7: Mapping of the consensus binding pockets on the structures of *Mtb* Kas A and *Hsmt* KasA. A) *Mtb* KasA structure shown as a grey cartoon and the identified pockets colored in blue, green, red and orange and numbered accordingly. B) *Hsmt* KasA structure shown as a grey cartoon and the identified pockets colored in orange, red and yellow and numbered accordingly. The allosteric site is represented in yellow and the active sites are shown in orange and red.

The residues constituting the respective pockets are shown in Table S1. To further validate the allosteric site search results, the residues constituting the respective pockets were investigated for their contribution to the functionality of the protein. In *Mtb* KasA, pockets 1 and 2 are lined with hydrophobic residues that make up the acyl-binding tunnel, which stems from the surface of the protein and extends and terminates into the malonyl-binding pockets (pockets 3 and 4). Pockets 3 and 4 make up the active site and the catalytic residues Cys171 and His311 are also found in these pockets. In *Hsmt* KasA, the active site residues make up the active site pockets 1 and 2. Motifs that are associated with the identified pockets in both proteins include motifs 2, 3, 4, 7, 8 and 17 that are found in the active site as well as motifs 4, 10, 11 and 14 in the allosteric sites. These motifs and their contribution to function have been discussed in Chapter 2. Figure 3.8 shows the position of the acyl-binding tunnel relative to the active and allosteric sites in *Mtb* KasA.

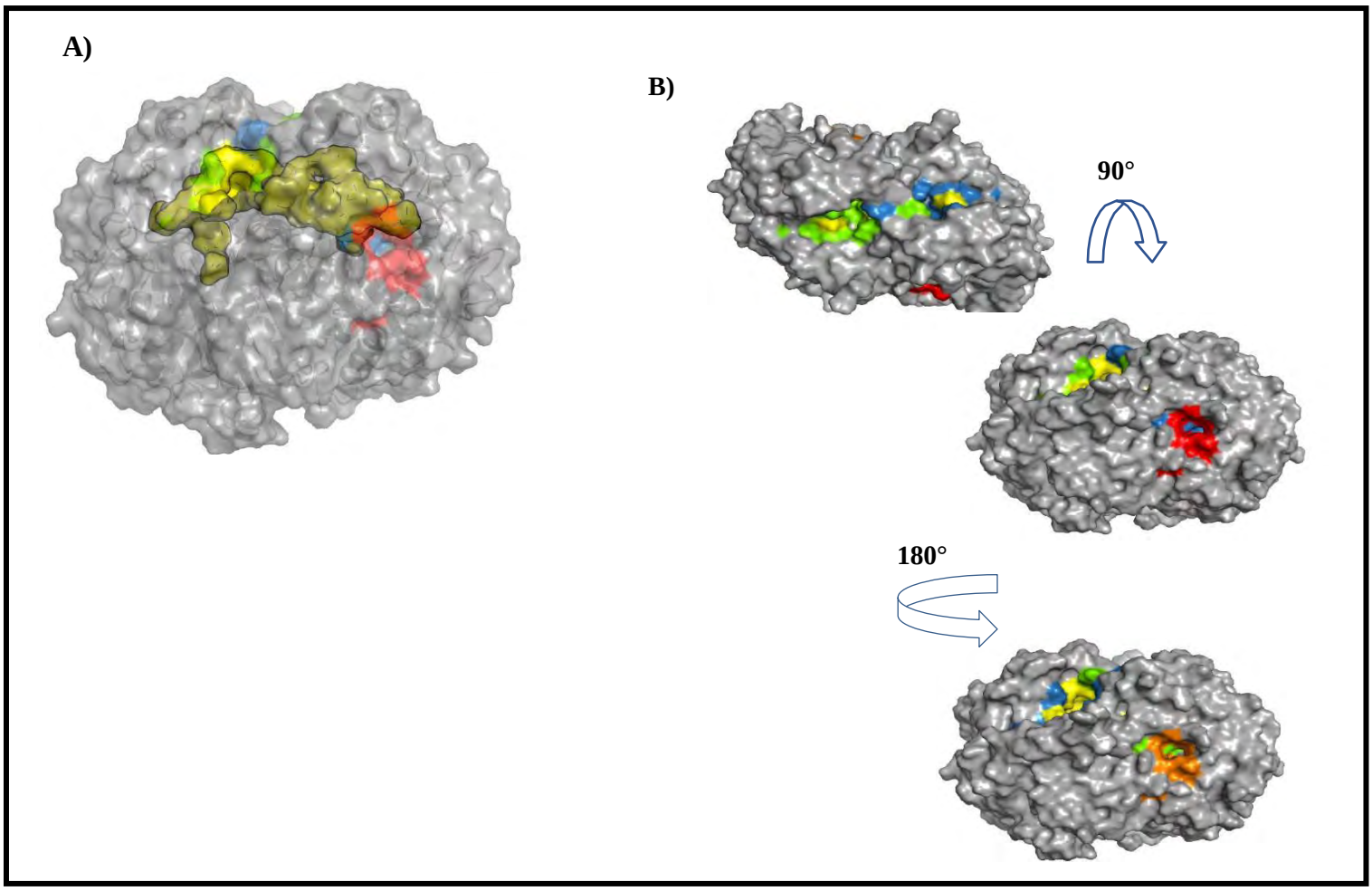


Figure 3.8: The position of the acyl-binding tunnel relative to the allosteric pockets and the active site pockets in *Mtb* KasA. A) *Mtb* KasA represented as a transparent surface in grey and the acyl binding tunnel shown as a closed surface in yellow. B) Rotations of the surface representation of *Mtb* KasA at 90° and 180°. The allosteric pockets are colored in green and blue whilst the active sites are represented in red and orange. The acyl-binding tunnel is shown in yellow

3.4 CHAPTER CONCLUSION

Allostery is a key molecular mechanism supporting the control and modulation in various cellular processes. Identifying druggable sites is a crucial step in structure-based drug design as it enables the design of small drug-like molecules to bind to these sites and induce the required therapeutic effects. The protein structures used in this study were *Mtb* KasA (PDB ID: 6P9K) and *Hsmt* KasA (PDB ID: 2IWY). Allosteric site search on these protein structures was conducted via a combination of computational tools. Various tools were used in order to increase the chances of accurately predicting the presence of putative allosteric sites of interest. These tools employ different algorithms in the prediction of allosteric sites, and this includes geometry-based, grid-based, energetic as well as machine learning models. A search for consensus pockets predicted by

at least 3 out of the 4 programs used was done and the result of this computation revealed that two allosteric sites were identified in *Mtb* KasA and only one site in *Hsmt* KasA. In order to further validate the results, the residues constituting the sites were analyzed for the role they play in the functioning of the protein. In addition, common sequence patterns (motifs) that are associated with these pockets also gave insights into the function of the identified pockets. All in all, it is evident that all of the tools used accurately predicted the probable allosteric sites in *Mtb* KasA. However, three out of the four tools used failed to predict the acyl-binding pockets of the human homolog protein, revealing that the algorithm employed in these tools cannot be fully relied on in predicting binding sites. What is also interesting to note is that both AutoLigand and DogSiteScorer identified a pocket with residues that constitute the acyl-binding tunnel in *Hsmt* KasA (Pocket 4), and this was not detected by the other tools. DogSiteScorer predicted the greatest number of pockets in *Mtb* KasA compared other tools. In addition, the program also identified two surface pockets in *Mtb* KasA that were not predicted by the other tools. A further investigation into these sites could provide important insights for use in drug discovery and design. Of all the programs used, SiteMap accurately identified the key pockets in *Mtb* KasA and *Hsmt* KasA and also had the greatest number of residues per binding site.

The next chapter will focus on the identification of ligands that interact with the identified allosteric sites in an attempt to identify potential hits that can be used in the successful inhibition of the *Mtb* KasA protein.

CHAPTER 4

STRUCTURE-BASED VIRTUAL SCREENING

Virtual screening (VS) is an approach that has gained popularity over the decades and is seen as a complementary approach to the experimental high throughput screening (HTS). However, the high cost and low hit rate associated with HTS has motivated the need to develop computational alternatives that utilize *in-silico* approaches in drug compound screening [157]. Structure-based drug design (SBDD) aims at understanding the molecular basis of a disease by utilizing the knowledge of the three-dimensional (3D) structure of the biological target. The structure-based screening process involves a variety of sequential computational stages such as target preparation, docking and post-docking analysis as well as hierarchizing compounds of interest for testing [158]. The implementation of computer-aided drug design has led to positive outcomes in drug discovery. New biologically active compounds have been predicted along with their receptor-bound structures at high hit rate success compared to the conventional HTS. In addition, the VS workflow can greatly shorten the cycle of hit discovery [159, 160].

This chapter aims at discovering novel allosteric modulators by virtually screening DrugBank compounds against the *Mtb* KasA and *Hsmt* KasA protein structures [161]. The entire surface of the protein structures was screened against 2089 drug compounds retrieved from the DrugBank provided by Sheik Amamuddy *et al*., [162] using AutoDock Vina [163]. Protein and ligand preparation was done using AutoDock 4.2 [164]. N-(2-cyano-3-methyl-1H-indol-5-yl)butane-1-sulfonamide (Ligand ID: O6G) was used as a control and the docking parameters were validated by redocking O6G on the wildtype *Mtb* KasA. The ligand poses and interactions were then evaluated, and the promising candidate compounds were selected based on their binding energies relative to *Mtb* KasA and their interaction with the allosteric sites identified in Chapter 3.

4.1 INTRODUCTION

The development of new drugs is characterized by long development cycles, high costs as well as low success rates [165]. Several medical conditions have treatment regimens that are inadequate or missing, and this has prompted the need to use artificial intelligence, virtual screening and machine learning approaches in the development of new therapeutics. This is because the traditional high-throughput technologies are unable to fully address the novel challenges that come with huge data being generated with respect to understanding the molecular mechanisms involved in numerous diseases.

In silico approaches have contributed vastly to drug discovery as they assist not only the wet-lab researchers but also computer experts who specialize in developing tools that are used to integrate the different types of biological data [166, 167]. Virtual screening was first coined in literature in 1997 and this technique is used to search small libraries of small molecules that are likely to bind to one or more drug targets. This approach usually generates a large data set of about 30-500 compounds that need to be validated experimentally for their suitability as inhibitors [168]. Despite several advances of *in-silico* screening, there are also pitfalls associated with this technique. Firstly, the high number of false-positives limits virtual screening to initial screening only. Structure-based virtual screening lacks reliable scoring systems that estimate the free binding energy as well as identifying inactive compounds that do not bind to the target to produce the desired therapeutic outcome [169]. Furthermore, this technique remains only a theoretical approach that requires validation by empirical means.

4.2 VIRTUAL SCREENING IN STRUCTURE-BASED DRUG DISCOVERY

The general scheme of a SBVS workflow begins with the processing of a 3D target structure that has been solved experimentally (X-ray, NMR, neutron scattering spectroscopy and cryo-electron microscopy) or computer modelled (homology modeling). Before considering a structure for SBVS, factors such as the druggability of the receptor, the choice of the binding site, selection of the appropriate protein structure, the identification of ligand binding sites as well as assigning the appropriate protonation states have to be considered [160]. This is followed by the careful selection of the compound library of small molecules that are to be screened according to the target in question. Each compound in the library is virtually docked into the target binding site(s) via a

docking program.

The aim of docking is to predict the protein-ligand complex structures by analyzing the conformational space of the ligand within the binding site of the protein. A scoring function is then used to evaluate the free binding energy between the docked compound and the target in each docking pose. This produces ranked compounds which are then subjected to post-processing by examining the calculated binding scores, validity of the generated poses and desired physicochemical properties. Post-processing produces a small group of top-ranked compounds that will be selected as candidates for experimental assays.

4.3 MOLECULAR DOCKING

Molecular docking is an important tool used in drug discovery and molecular modelling applications, that attempts to find the best binding orientation of a ligand into a protein molecule [170]. Docking can be achieved via two interconnected steps namely sampling the conformations of the ligand in the protein and ranking them via a scoring function. The molecular docking is thus split into two: the searching algorithm and the scoring algorithm. The searching algorithm looks into all the conformations of the ligand within the space available. However, this is a time-consuming process and it is practically impossible to sample all possible conformations for a compound, therefore, each compound is investigated within a given threshold of conformations. The conformational search seeks for structural parameters of the ligand such as torsional, translation and rotational degrees of freedom. The second feature of the docking algorithm is the scoring function. The purpose of the scoring function is to characterize the correct poses from the incorrect poses or binding of inactive compounds. The scoring algorithms estimate the binding affinity between the protein and the ligand as well as predicting the energy profiles [171]. Low energy profiles are generally preferred as they provide more stable interactions. Scoring functions are classed according to physics-based, empirical, knowledge-based and machine learning based methods. The first three methods are known to use linear regression whilst the fourth one uses non-linear regression methods [170].

In order to perform molecular docking, two types of approaches are generally used. These are the simulation approach in which the energy profiling is estimated for the protein-ligand complex and the complementarity approach that calculates surface complementarity between the ligand and the protein. In the simulation approach, the ligand and the protein are separated, and the ligand is then

allowed to bind into the grooves or binding sites of the protein in its conformational space. As the ligand moves within its conformational space, it generates energy that is known as the "total energy of the system". The shape complementarity approach uses the ligand and target as a set of surface structural features in order to enable molecular docking [172]. The molecular surface of the target is evaluated for its solvent accessible surface area and the ligand is described in terms of how best it matches the surface.

4.3.1 Types of Molecular Docking

Following protein and ligand preparation, the type of molecular docking to be formed must be considered. The two types of molecular docking experiments are flexible and rigid docking. Rigid docking treats both the receptor and the small molecule as rigid whilst flexible docking can involve either a flexible ligand and a rigid receptor or both the ligand and protein being flexible [173]. Rigid body docking produces a large number of docked conformations with suitable surface complementarity. This type of docking uses the fast Fourier transform (FFT) correlation approach to analyze the space of docked conformations by use of electrostatic interactions and solvent terms. Rigid docking resembles the 'lock and key' model and is mainly used for protein-protein docking. Docking using the rigid body docking approach is much faster than flexible ligand docking because the size of search space is smaller. However, if the incorrect conformation of a ligand is used, the chances of finding the best complementary fit are lowered.

Flexible docking on the other hand takes into consideration the flexibility of the ligand side chains. The flexibility of the side chains is known to play an important role in protein-ligand complexes. These changes allow for the alteration of the binding site of the receptor according to the orientation of the ligand [174]. Four strategies are currently used to dock flexible ligands namely: (1) Monte Carlo methods, (2) in site combinatorial search, (3) ligand build-up and (4) site-mapping and fragment assembly. This technique is computationally expensive; however, it provides better accuracy at reasonable speed.

4.3.2 AutoDock 4.2 and AutoDock Vina

Docking experiments are performed using various programs such as AutoDock vs 4.2.6 [155], AutoDock Vina vs 1.1.2 [163], Glide vs 7.7 [175] and Dock vs 6.8 [176], only to mention a few, and these have gained popularity over the years due to the accuracy of the docking algorithm. AutoDock Tools (ADT) is part of MGL Tools, from the Molecular Graphics Laboratory at the Scripps Research Institute, built on the Python Molecular Viewer (PMV) [177]. AutoDock utilizes

the Lamarckian genetic algorithm (LGA) for conformational searching. A number of conformations are created but only those with the lowest binding energy are selected. AutoDock4 uses a semi-empirical free energy force field in order to predict free binding energies of ligand molecules to receptor proteins [178]. This force field is based on a detailed thermodynamic model that incorporates intramolecular energies into the predicted free binding energy. In addition, terms of desolvation are also included in the model that use a set of atom types and charges. AutoDock4 allows for flexible protein and ligand docking, and during simulation, these are treated categorically to allow for the rotation of the ligand around the torsion degrees of freedom. AutoDock facilitates the input of molecule files and subjects them to a set of methods that the user specifies for protonation, calculating gasteiger charges, specifying rotatable bonds and torsions in the ligand molecule and launch docking calculations.

AutoDock Vina was designed to be used with the file format type of AutoDock4, pdbqt file format which is an extension of the PDB file. This program uses an iterated local search global optimizer that is based on stochastic global and local optimization methods. Since protein-ligand docking experiments require the availability of a 3D structure of a target protein and a library of small compounds (databases) in order to identify potential hits, the DrugBank database was explored for FDA approved compounds that could potentially have inhibitory effects against *Mtb* KasA for the purposes of drug repurposing.

4.4 DRUGBANK DATABASE

DrugBank is a freely available web database that contains extensive molecular information about the drug, drug target, mechanism of action of drug and drug interactions of FDA approved drugs as well as drugs that await the FDA approval process [161]. This database is the most commonly used reference drug resource worldwide as it contains high quality information of primary origin. DrugBank 1.0 was first released in 2006 and primarily provided the physicochemical data on a few chosen FDA approved drugs and their drug targets. In 2008, a new version was released, and it added pharmacological, pharmacogenomic as well as molecular biological data [179]. DrugBank3.0 was developed in 2010 and included drug-drug and drug-food interactions. Furthermore, the pharmacokinetic information of the data was released. This is followed by the subsequent addition of drug metabolism data, qualitative structure activity relationships (QSAR) and ADMET (absorption, distribution, metabolism, excretion and toxicity) data in 2014. The current DrugBank

database, DrugBank 5.0 contains 2358 FDA approved drugs and 4501 drugs still undergoing clinical trials. This new version has additional information such as pharmacometabolomics, gene expression levels and protein expression levels. The information contained in this database is routinely used by educators, biologists, pharmaceutical researchers, bioinformaticians as well as the general public.

4.5 METHODOLOGY

Figure 4.1 outlines the procedure followed in the molecular docking of Drug Bank compounds onto the KasA protein structures.

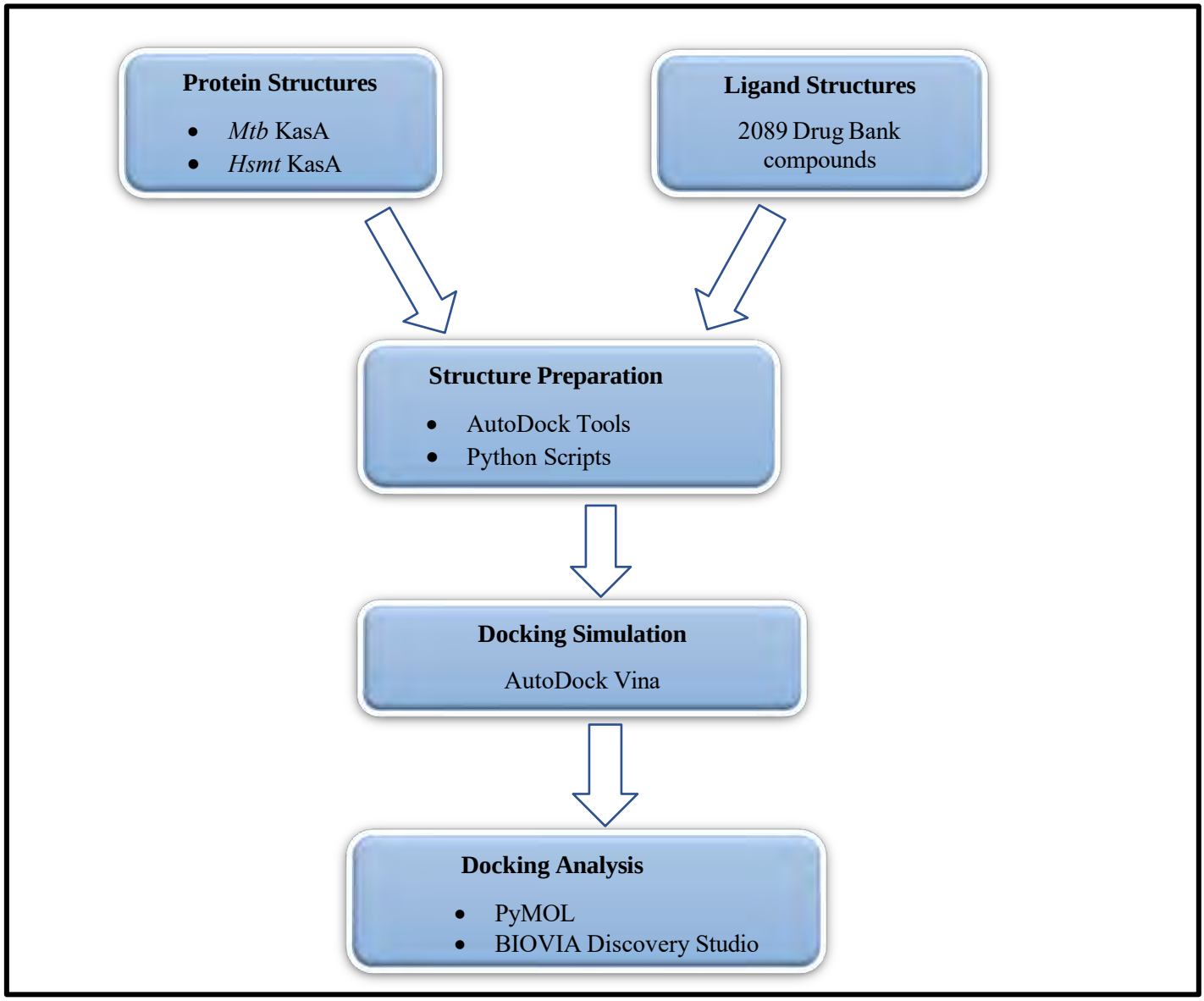


Figure 4.1: Summarized workflow of the molecular docking procedure showing all the steps and tools used.

4.5.1 Data Retrieval

The crystallographic structures of *Mtb* KasA (6P9K) and *Hsmt* KasA (2IWY) were retrieved from the Protein Data Bank. The inhibitor ligands used were a data set of minimized 2089 DrugBank compounds provided by Sheik Amamuddy et al. [162]. These structures are all FDA-approved compounds that were provided in pdbqt file format. Prior to molecular docking, all crystallographic water molecules and ligands were removed from the target protein structures using the Protein Preparation Wizard of Maestro vs 12.9 of the Schrödinger Suites software [151].

4.5.2 Structure Preparation

The protonated protein structures of *Mtb* KasA and *Hsmt* KasA were prepared for docking using AutoDock Tools by firstly adding polar hydrogens, merging non-polar hydrogens and adding gasteiger charges using the python script, "*prepare_receptor4.py*". The distinct parameters of this script also allow for assigning of AutoDock type atoms to the structures as well as to remove waters if any [180]. The resulting structure was saved as a pdbqt file format for both proteins.

As the *Mtb* KasA protein had a co-crystallized ligand in its structure, **N**-(2-cyano-3-methyl-1H-indol-5-yl) butane-1-sulfonamide (O6G) was used as a positive control for the docking procedure. This molecule was prepared for redocking using the python script of the AutoDock Tools software, "*prepare_ligand4.py*". This script adds polar hydrogens, merges non-polar hydrogens, adds gasteiger charges as well as assigns AutoDock type atoms to the ligand structure. In addition, the torsion degrees of freedom, aromatic carbons and rotatable bonds are also calculated [181]. The 2089 DrugBank compounds were prepared following the same procedure as the control ligand.

4.5.3 Initial Docking Validation

Validation of a docking protocol is a necessary step in the analysis of biological data as it determines the reproducibility of the results as well as to validate the docking parameters used [182]. Since the 3D structures of *Mtb* KasA and *Hsmt* KasA are similar, it was imperative to validate the initial docking procedure on both structures. This was done by redocking the co-crystallized ligand of *Mtb* KasA, O6G onto the 3D structures of *Mtb* KasA and the human homolog using AutoDock Vina. Both proteins and the ligand were prepared for docking using the receptor and ligand preparation tools of the AutoDock Suite. A docking box size of 70 x

60 x 84 Å was centered at co- ordinates (-32.669, 0.001, -23.471) on the *Mtb* KasA protein whilst a box size of 82 x 72 x 84 Å was centered at co-ordinates (15.168, 34.154, 26.261) on *Hsmt* KasA. A grid spacing of 1.0 Å and an exhaustiveness of 320 was used. The redocked ligand poses were analyzed and the one with the best pose was selected based on its nature to imitate the pose of co-crystallized ligand. The interactions of the redocked ligand with the proteins were visualized in PyMOL and BIOVIA Discovery Studio. The parameters used in the redocking experiment were then used for docking of the 2089 DrugBank compounds.

4.5.4 Docking Parameters and Grid Evaluation

Blind docking was performed using AutoDock Vina. Two docking experiments were set up: blind docking on *Mtb* KasA and another on *Hsmt* KasA. The AutoGrid package of AutoDock Tools was used to generate a grid box of interaction energies based on the receptor co- ordinates [122]. The grid box was set to cover the entire surface of the protein structures. The grid parameters shown in Table 4.1 were generated for the x, y, z dimensions and a grid spacing of 1.0Å was used. An *ad hoc* python script was used to write the grid parameters to the vina configuration file for each ligand.

Table 4.1: Molecular Docking parameters used for *Mtb* KasA and *Hsmt* KasA.

Protein	Box size	Centre co-ordinates
***Mtb* KasA**	X = 70 Y = 60 Z = 84	x = -32.669 y = 0.001 z = -23.471
***Hsmt* KasA**	X = 82 Y = 72 Z = 84	x = 15.168 y = 34.154 z = 26.261

4.5.5 Docking Simulation

For each DrugBank compound, a vina configuration file containing the grid parameters was created in preparation for docking. A python script was used to create commands to run Vina on each of the vina format files by creating a "*gnu_parallel.jobs*" file, which uses gnu parallel to execute jobs in parallel depending on the number of central processing unit (CPU) cores assigned [183]. In this analysis, 8 cores and an exhaustiveness of 320 was used. The high exhaustiveness allows for the prediction of crystallographic poses of the ligand in order to select the best conformational pose for binding. Docking was initiated on the Center for High

Performance Computing cluster (CHPC) by running the job file using the command "*vina –config ligand_name. vina*" for the protein-ligand pairs. The job was then submitted onto the cluster using a "*docking.pbs*" file, with a walltime of 48:00:00 and a normal queue. Following this computation, the output files were then subjected to the command "*vina_split –input ligand_name_vina.pdbqt*", which generated nine poses for each compound docking on the receptor. Log files were also generated for each ligand showing the poses ranked according to the binding affinities.

4.5.6 Docking Analysis

Docking analysis was done using python scripts in order to extract ligand poses with the lowest binding energies. Further screening was done by utilizing the Jupyter Lab software [184] to extract ligands with binding energies less than or equal to the binding energy of the control. A python script running PyMOL commands was used to visualize the poses of the docked ligands onto the protein structures. Due to the protein structures being homodimers, the script enabled the superimposition of both chains and pockets identified in Chapter 3 so as to compare the ligands accurately. Ligands binding to the allosteric site of the *Mtb* KasA protein were extracted using PyMOL commands by selecting ligands within 10Å of the pocket. The protein-ligand interactions of promising candidate compounds were visualized in BIOVIA Discovery Studio.

4.6 RESULTS AND DISCUSSION

4.6.1 Docking Validation

The quality of reproducibility of a known ligand binding pose is used to validate the docking protocol. In addition, validation is an essential step used to authenticate the docking parameters used. This step was performed by redocking O6G onto the crystal structures of *Mtb* KasA and *Hsmt* KasA and comparing the resulting RSMD value on all the atoms. The redocked ligand was selected based on its ability to imitate the pose and interactions exhibited in the wildtype structure. The RMSD value between the two binding poses was computed in PyMOL and this gave a value of 0.218Å for *Mtb* KasA and 2.623Å for *Hsmt* KasA. A low RSMD value (less than 1Å) indicates less divergence from the expected pose, thus the more similar the two structures. Figure 4.2 shows the superimposed 3D structures of the redocked ligand and original co-crystallized ligand in both protein structures. The redocked ligand assumed a

similar orientation as the as the co-crystallized ligand in *Mtb* KasA (Figure 4.2A), however a slightly different orientation was observed for *Hsmt* KasA (Figure 4.2B). This result is also further validated by the high RSMD value obtained. However, what is interesting to note is that the ligand binds to both protein structures with a binding energy of -7.7kcal/mol, which poses a great challenge as there is need to identify compounds that bind preferentially to *Mtb* KasA and having little to no effects in *Hsmt* KasA.

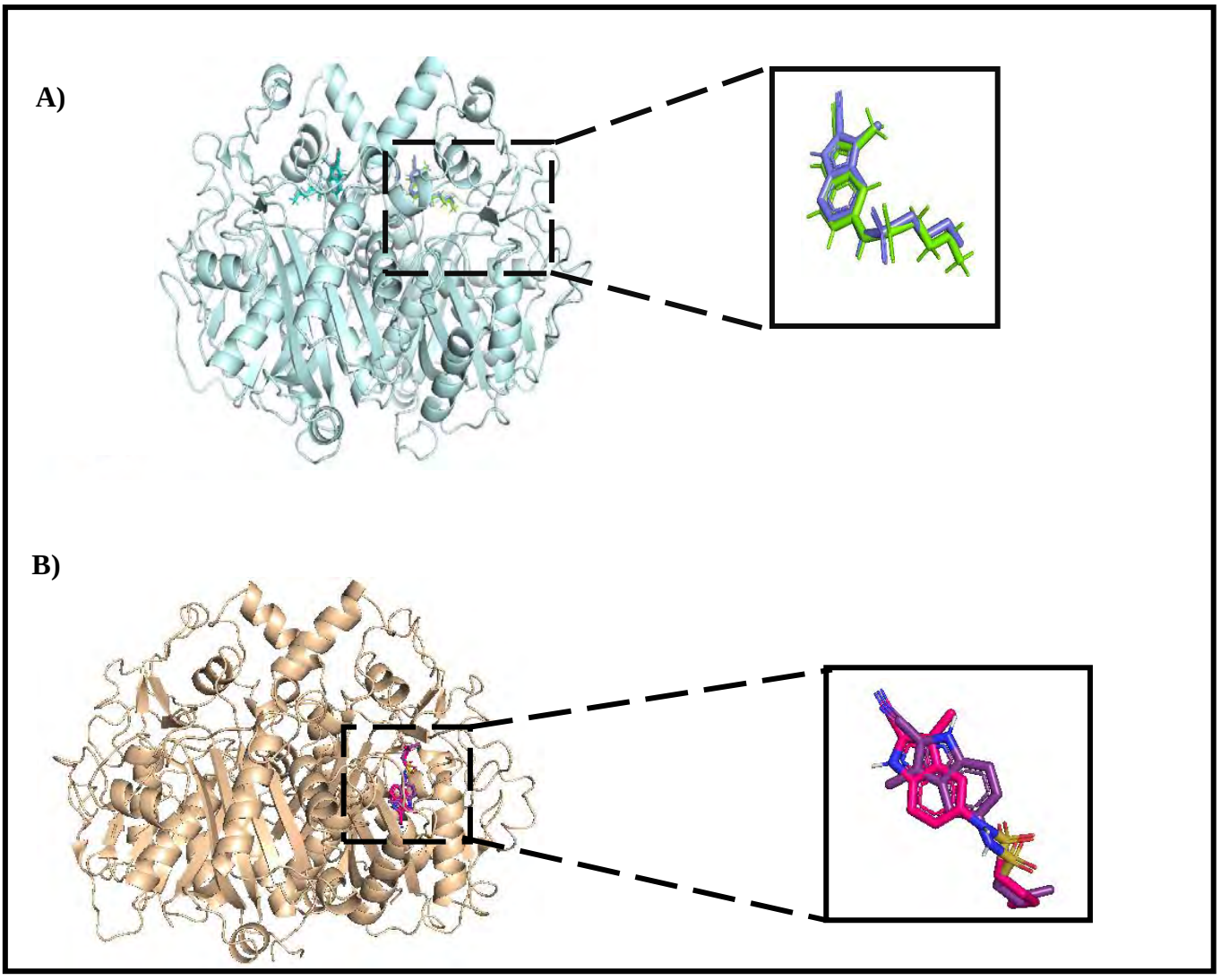


Figure 4.2: Docking validation of O6G onto the crystal structures of *Mtb* KasA and *Hsmt* KasA as visualized in PyMOL. A)*Mtb* KasA dimer represented as a cartoon in pale cyan, with the superimposed redocked O6G ligand (blue) against the reference ligand (green). B) *Hsmt* KasA dimer represented as a cartoon in wheat, with the superimposed redocked ligand (pink) against the reference ligand (purple).

The protein ligand interactions of the original co-crystallized ligand and the redocked ligand in *Mtb* KasA and *Hsmt* KasA were visualized in BIOVIA Discovery Studio and are shown in Figure S6. In *Mtb* KasA, both Pro201 and Gly200 interacted with additional amide-pi bonds with the redocked ligand pose compared to only pi-pi interactions as seen with the reference ligand . Key interactions included two hydrogen bonds formed by Glu199 and Gly200 (shown in green) as well as the formation of a pi-sigma bond with the aromatic ring by Gly200 compared to the pi-alkyl bond it forms with the aromatic ring in the reference ligand. All other proximal residues interacted with Van der Waals forces of attraction. In *Hsmt* KasA, the reference ligand formed more hydrogen bonds with the protein compared to *Mtb* KasA, as seen with Gly323, Ala324, Asn451, Asp310 and Thr352. However, the hydrogen bonds were reduced to only three in the redocked ligand pose. The orientation of the redocked ligand's side chain allowed for the formation of a hydrogen bond with Thr350, which was absent with the reference ligand. Phe445 forms a pi-alkyl bond with the reference ligand and an amide-pi bond with the redocked ligand., whilst Pro317 forms a pi-sigma bond with the reference ligand and a pi-alkyl bond with the redocked ligand. The pi-alkyl bonds formed between Ala358, Gly355 and Phe274 with the reference ligand are lost in the redocked dose. An unfavorable donor-donor bond is formed between His348 and both the reference and redocked ligand pose. The formation of unfavorable bonds in protein-ligand complexes reduces the stability of the complex as these bonds indicate a force of repulsion occurring at atomic level. However, factors such as protein flexibility and the type of docking program used need to be taken into account before ruling out the compound as a potential inhibitor.

4.6.2 Blind docking analysis

Blind docking refers to the docking of a ligand molecule to the whole surface of the protein without any prior knowledge of the target binding site. The 3D structures of *Mtb* KasA and *Hsmt* KasA were subjected to blind docking using 2089 prepared DrugBank compounds on the CHPC cluster using AutoDock Vina. Of these compounds, 9 failed to dock to the protein structures due to size constraints. These are DB0006, DB00638, DB01278, DB01284, DB05528, DB08869, DB09067, DB11322 and DB13928. The data set was further reduced to 1087 compounds after screening for compounds with binding energies less than or equal to -7.7 kcal/mol, the binding energy of the control ligand, O6G. The overall binding affinities of the initial 2080 docked compounds were represented on an heatmap produced using an *ad hoc* python script as shown

in Figure S7. The heatmap was centered on the binding energy value of -7.7 kcal/mol, which was the cut-off used in screening of the compounds. The binding energies of these compounds indicate the stability of the ligands when bound to the respective protein structures. A lower binding energy indicates a higher binding affinity for the receptor, which in turn leads to stable protein- ligand complexes. On the other hand, a higher binding energy indicates a lower binding affinity for the receptor, and this results in unstable protein-ligand complexes as the ligand can be easily displaced by other compounds of a relatively higher affinity. Many of the compounds shown on the heatmap had binding energies greater than that of O6G, whilst few compounds had binding energies lower than the control for both proteins. The similarities in the binding energies of the DrugBank compounds to the respective protein structures could be attributed to the highly conserved and similar 3D structures of the proteins.

4.6.2.1 Blind docking analysis on *Mtb* KasA

Figure 4.3 shows the blind docking results of the screened 1087 DrugBank compounds. Ligand binding was distributed across the identified allosteric and active site, with a few compounds binding to the peripheral regions of the protein. The docking analysis revealed that only four compounds were bound to the active site of *Mtb* KasA. These compounds included DB00922, DB01014, DB12954, and DB14086. What is interesting to note is that all these compounds were preferentially bound to the active site of one monomer of the protein structure, which questions the nature of similar active sites binding the same ligand. However, this result could be attributed to several factors. Could it be that one monomer is active at a given time in *Mtb* KasA? Recent studies on some homodimeric structures have revealed that while dimerization is crucial for enzyme activity, one protomer is active and the other is inactive [185]. This has not been investigated in *Mtb* KasA, and some insights into this phenomenon could help in understanding the inhibition mechanisms of the protein. To expound further, this finding could be associated with the ligands having adopted a conformation that is able to interact with one monomer of the dimer. It is also important to note that not all ligands are able to interact simultaneously with two protomers, as is the case with bivalent ligands. Nonetheless, these compounds displayed a higher binding affinity for *Hsmt* KasA compared to the *Mtb* protein, hence proving them unsuitable as potential inhibitors.

Since the main focus of this study is exploring the compounds that bind to the potential allosteric sites of *Mtb* KasA, ligands binding within 10 Å of the identified allosteric sites were selected in

order to investigate ligands occupying the acyl-binding tunnel, which plays a crucial role in enzyme function.

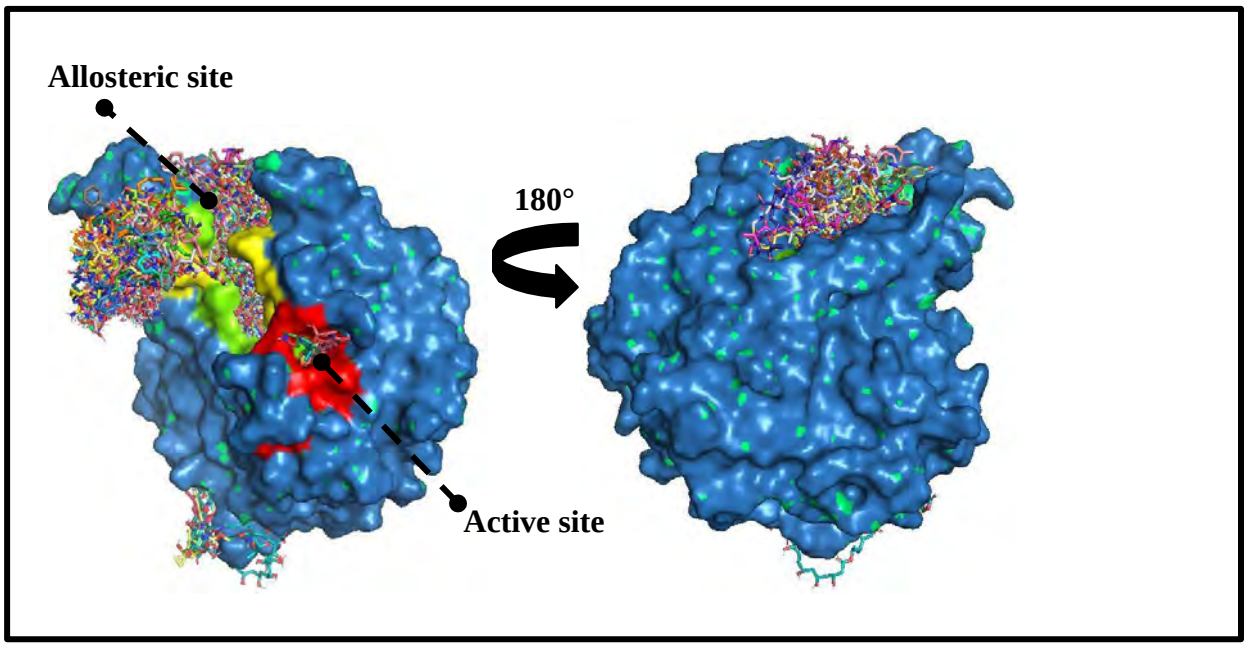


Figure 4.3: Superimposed structures of *Mtb* KasA in surface representation. One monomer is shown in blue and the other in lime green. The allosteric and active site pockets are clearly labelled and colored in green and red respectively. The acyl-binding tunnel is colored in yellow.

4.6.2.2 Blind docking analysis on *Hsmt* KasA

Ligand binding in *Hsmt* KasA was not distributed evenly as seen in *Mtb* KasA as most of the ligands were scattered on the surface of the protein, with a number of the ligands occupying the active site and allosteric site regions (Figure 4.4). It is important to state that the potential allosteric site identified in *Hsmt* KasA is not associated with the acyl-binding tunnel region. The docking analysis revealed a number of ligand compounds that bind to the active site of *Hsmt* KasA. Upon analysis, most of these compounds had a higher binding affinity for *Hsmt* KasA compared to *Mtb* KasA, where they were bound either to the identified allosteric sites or peripheral regions of the protein not characterized as potential allosteric sites. However, since the aim of this study is to investigate compounds that bind to the allosteric sites of *Mtb* KasA with more emphasis on those occupying the acyl-binding tunnel, the active site compounds were excluded.

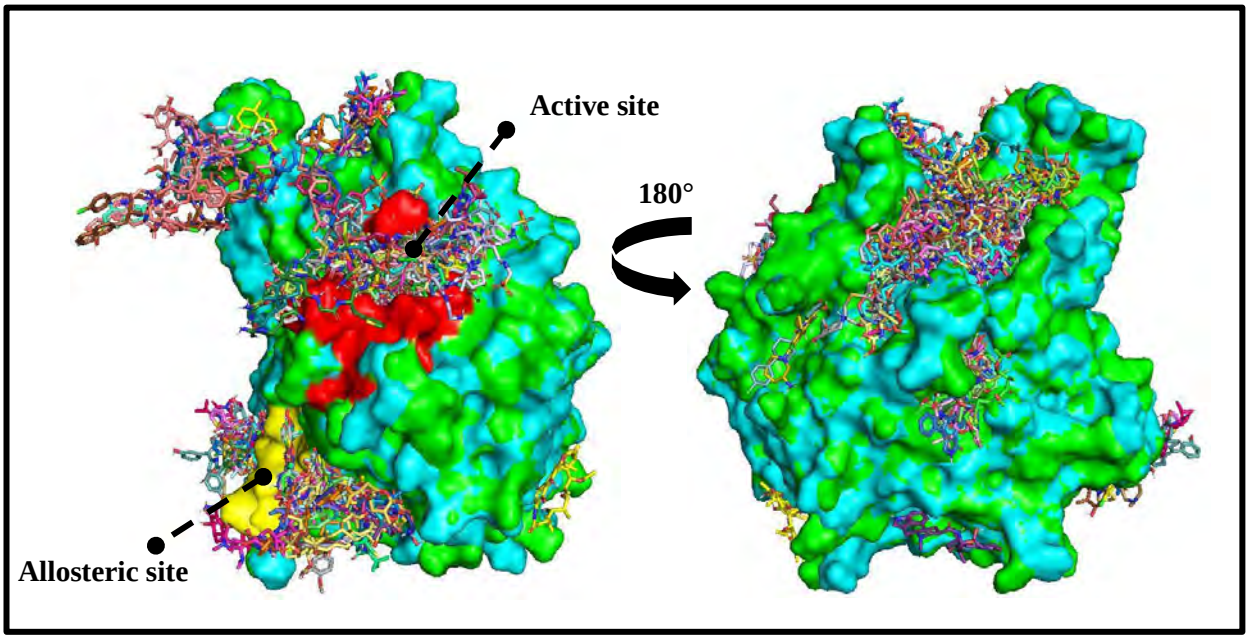


Figure 4.4: Superimposed structure of *Hsmt* KasA in surface representation. One monomer is shown in cyan and the other monomer in green. The active site and potential allosteric site are clearly labelled and colored in red and yellow respectively.

4.6.3 Binding energies of promising hit compounds

Binding energies are used to determine the binding affinity between a ligand and a protein after docking. A cut off of -7.7kcal/mol was used to screen for potential hit compounds as this was the binding energy of the control ligand (O6G). The result of this computation yielded 27 hit compounds, based on their lower binding energies and slightly higher affinity for *Mtb* KasA compared to *Hsmt* KasA, as shown in Table 4.2. It is important to note that these compounds bind to the human homolog protein with appreciable affinity, thus posing a challenge in inhibitor drug design. Further screening was done by identifying compounds that had relatively low binding affinity for *Hsmt* KasA (high binding energy) and reasonably high affinity for *Mtb* KasA, with the assumption that at relatively high concentrations of the compound *in-vitro*, these compounds would preferentially bind to *Mtb* KasA and saturate the binding sites. This led to the identification of ten promising drug compounds highlighted in yellow in Table 4.2. DB08889 had the highest binding affinity for the *Mtb* protein in comparison with the human protein, with a binding energy of -9.9 kcal/mol. This was followed by compounds DB06755 and DB09270, with binding energies -9.8 kcal/mol and -9.1 kcal/mol respectively.

Table 4.2: Binding energies of promising candidate compounds against *Mtb* KasA and *Hsmt* Kas A in kcal/mol. Compounds with relatively low binding energies (high affinity) for *Mtb* KasA were highlighted in yellow.

Drug Compound	***Mtb* KasA binding energy**	***Hsmt* KasA binding energy**
DB08889	-9.9	-6.6
DB06755	-9.8	-6
DB09270	-9.1	-6.6
DB12278	-8.9	-6.6
DB13720	-8.9	-6.7
DB06720	-8.8	-6.5
DB08936	-8.7	-6.5
DB11226	-8.5	-6.1
DB00392	-8.5	-6.3
DB01625	-8.5	-6.6
DB00343	-8.5	-6.7
DB01105	-8.2	-6.2
DB01246	-8.2	-6.4
DB13225	-8.1	-6.4
DB00420	-8.1	-6.7
DB00290	-8	-6.3
DB13854	-8	-6.5
DB04711	-8	-6.6
DB11583	-7.9	-6.3
DB00887	-7.9	-6.3
DB13179	-7.9	-6.4
DB00449	-7.9	-6.7
DB00539	-7.8	-6.4
DB00781	-7.7	-5.6
DB00777	-7.7	-6.4
DB08887	-7.7	-6.6
DB14104	-7.7	-6.7

An interesting observation made when analyzing these results was that all the 27 candidate compounds were found to be occupying the acyl-binding tunnel in *Mtb* KasA (Figure 4.5). The same compounds were investigated for their interaction with *Hsmt* KasA and it was found that none of these compounds occupied the acyl-binding tunnel of the protein. This means that the identified compounds exclusively bind to the acyl-binding tunnel of *Mtb* KasA, thus making them interesting compounds for further investigation. However, it is important to note that binding energies alone cannot be sufficiently used to determine ligand suitability in this case.

This is because binding energies are calculated from the scoring function used in docking programs, and these scoring functions are not always accurate for divergent ligands as those in the DrugBank database [186]. In addition, cross-docking of ligands to non-native protein structures has been associated with the prediction of the wrong binding mode, which in turn affects the binding energy generated [187]. As such, other factors such as the presence of conventional hydrogen bonds (strongest molecular bond) in the protein-ligand complexes and the molecular mass of the compound need be evaluated before a ligand is deemed suitable as a potential inhibitor.

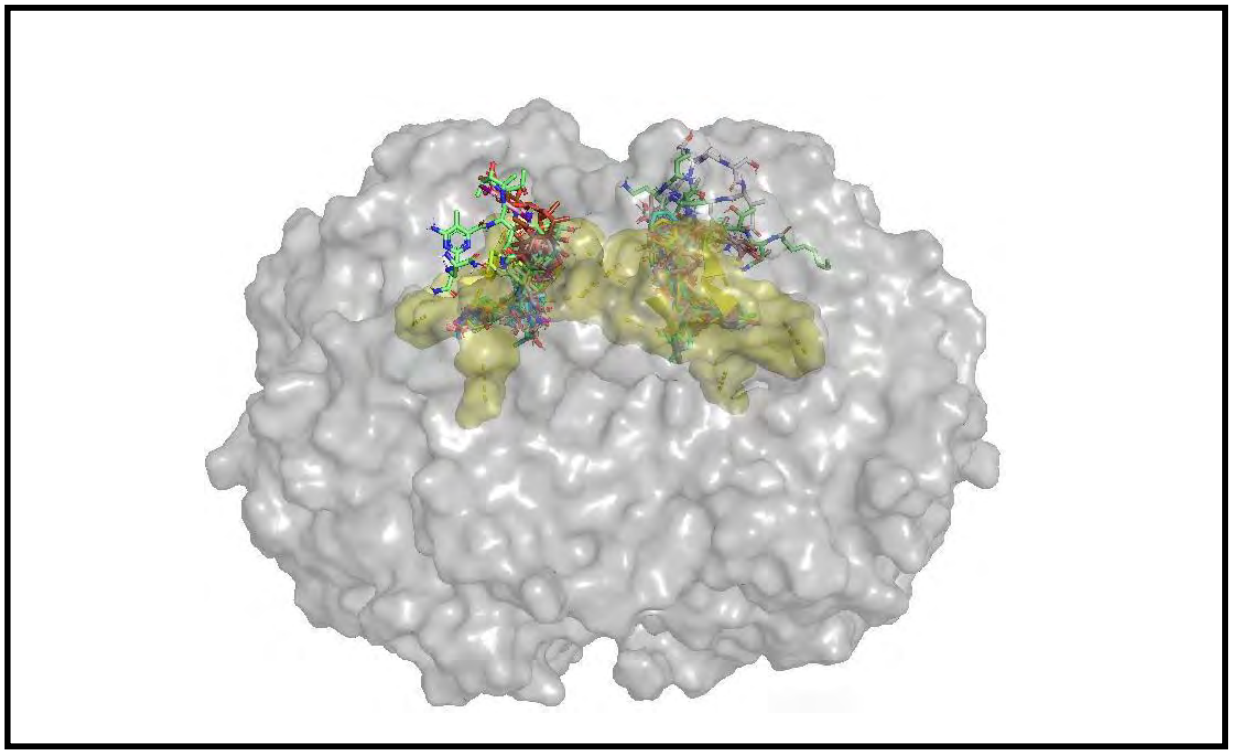


Figure 4.5: Binding of the promising candidate compounds in the acyl-binding tunnel of *Mtb* KasA. The protein is represented as a transparent grey surface and the acyl-binding tunnel is shown as a closed surface in yellow. The ligand molecules shown as licorice sticks are seen to interact with the acyl-binding tunnel.

4.7 CHAPTER CONCLUSION

Molecular docking is a widely used approach in computer aided drug design that enables us to predict the binding affinity between a ligand and a protein as well as the structure of the protein-ligand complex. In this chapter, 2089 FDA approved DrugBank compounds were screened against the structures of *Mtb* KasA and *Hsmt* KasA, with the aim of identifying potential allosteric modulators. The main focus was on the blind docking outcome associated with the consensus allosteric sites identified by the allosteric site search algorithms in Chapter 3. Attention was given to the compounds binding exclusively to the acyl-binding tunnel of *Mtb* KasA, which is important for shuttling substrates into the active site for catalysis as well as accommodating the growing fatty acid chain products. A heatmap was constructed to show the overall binding energies of docked DrugBank compounds to *Mtb* KasA and *Hsmt* KasA. A binding energy threshold of less than or equal to -7.7kcal/mol was used to screen ligands as determined by the control ligand in the validation of the docking protocol. Twenty-seven hit compounds were identified based on their relatively high binding affinity for *Mtb* KasA compared to *Hsmt* KasA. However, due to the appreciable affinity that these compounds displayed for the human homolog protein, binding energies alone could not be used to determine the ligands' suitability as potential inhibitors, thus further screening experiments ought to be performed. The identified compounds however serve as a starting point in screening for potential hit compounds with inhibitory activity against *Mtb* KasA.

CHAPTER FIVE

CONCLUSION AND FUTURE REMARKS

Drug resistance in TB still remains a challenge and this prompts for the discovery of new drug targets and compounds with novel mechanisms of action. The main focus of this thesis was to provide insights into the *Mtb* KasA protein which has been identified as a potential drug target by adding to the existing and expanding knowledge of TB drug discovery. *In-silico* based approaches were used to assess the attractiveness of *Mtb* KasA as drug target and to identify novel compounds that inhibit the protein's allosteric sites. The approaches used to achieve this aim included sequence and structural analysis, pocket identification and allosteric site search as well as molecular docking. This chapter provides a summary of the key findings of this study and presents future work plans to validate the present results and discover new compounds that can be used to successfully inhibit *Mtb* KasA.

5.1 CONCLUDING REMARKS

This study details the analysis of *Mtb* KasA as a potential drug target in the treatment of Tuberculosis. Although this protein has been identified as an attractive target for the development of antituberculosis agents in literature, little is known about the protein at sequence analysis level. A total of 15 KasA homolog sequences were retrieved from the UniProt database, with 8 of these sequences derived from bacteria, 3 from fungi and 4 from mammals. Multiple sequence alignment, motif analysis and phylogenetic analysis revealed similarities of the homolog proteins at sequence level, showing the conservation of residues and sequence patterns that have structural and functional roles.

Comparative multiple sequence alignment was carried out in order to observe the conservation of residues across the homolog sequences. Although there were variations at residue level, owing to substitutions and insertions in the alignment, the catalytic residues were conserved across all the homolog sequences. Motif analysis revealed common sequence patterns that play both structural and functional roles in the KasA protein family. Particularly outstanding was motif 6 that contains the helix-turn-helix or hypothetical gate segment of the KasA proteins. This region is essential as

it facilitates the opening of the acyl-binding tunnel during catalysis in order to accommodate the growing fatty acid chains. Phylogenetic analysis of the homolog proteins revealed a distinct clustering of the prokaryotic sequences from the eukaryotic sequences. These results were consistent with the sequence identities shared among the sequences and this was also further supported by the all-versus-all sequence identity heatmap.

Chapter 3 focused on allosteric site search on the *Mtb* protein and the human homolog by firstly employing cavity calculations in order to identify binding pockets on the respective protein structures. A combination of allosteric site search tools were used for this computation namely CavityPlus, AutoLigand, Protein Plus DogSiteScorer and SiteMap. These tools were used to determine consensus sites identified in at least 3 of the 4 programs used, in order to increase the accuracy of the results. Although all the programs correctly identified the functional sites of *Mtb* KasA, only one program correctly identified all functional sites in *Hsmt* KasA. All in all, two allosteric sites were identified in *Mtb* KasA and only one allosteric site was identified in *Hsmt* KasA. Allosteric site search was conducted in order to identify novel drug compounds that would be suitable as allosteric modulators, binding with a high binding affinity to the *Mtb* protein compared to the human homolog protein.

Molecular docking of the 2089 FDA approved DrugBank compounds was performed in Chapter 4 using the AutoDock Tools software. Firstly, validation of the initial docking protocol was performed by redocking the co-crystallized ligand, O6G, onto *Mtb* KasA and *Hsmt* KasA as a positive control. The protein structures and ligands were prepared for docking in AutoDock Vina and the docking simulations were performed on the CHPC cluster. The ligands were screened according to the binding energies, with -7.7kcal/mol used as the threshold. Main focus was given to ligands that bound exclusively to the *Mtb* KasA acyl-binding site, reducing the dataset to 27 hit compounds. However, due to the similarities in the binding energies of the compounds to both proteins, the use of binding energies alone as an indicator of ligand suitability was ruled out. Thus, the identified compounds serve as a starting point for further screening in order to identify potential inhibitors against *Mtb* KasA.

Overall, the study presented in this thesis has provided useful insights in the structure and function of *Mtb* KasA, with particular reference to allosteric inhibition. Although previous studies have attempted to discover drugs that target the allosteric site, compounds inhibiting the acyl-binding

tunnel have not been studied before, which make this research novel and if further studies are performed, this could positively impact the field of research in overcoming the TB resistance problem.

5.2 FUTURE REMARKS

Further research and improvement of this study may include analyzing the hydrogen bonds formed between the ligand and the protein's residues. Hydrogen bonds are the strongest molecular interactions and are a major contributor to the stability of the protein-ligand complexes. Thus, complexes with more hydrogen bonds would be more desirable compared to those with little to none. The molecular weight of a compound is also an important indicator of ligand suitability, and compounds with a molecular mass less than 500g/mol are preferred (Lipinski Rule of 5).

In order to determine the stability of the protein-ligand complexes formed, molecular dynamic simulations must be employed and the protein-ligand trajectories can be analyzed using the Root Mean Square Deviation (RMSD), the Root Mean Square Fluctuations (RMSF), Radius of gyration and hydrogen bond profiling. In addition, the effects of mutations on the *Mtb* KasA-ligand systems can also be explored by mutating *Mtb* KasA and analyzing how the ligands behave in comparison to the wild-type system.

REFERENCES

[1] K. Zaman, “Tuberculosis: A Global Health Problem”, Accessed: Aug. 03, 2022. [Online]. Available: http://www.umdnj.

[2] S. H. Lee, “Tuberculosis infection and latent tuberculosis,” *Tuberculosis and Respiratory Diseases*, vol. 79, no. 4. The Korean Academy of Tuberculosis and Respiratory Diseases, pp. 201–206, Oct. 05, 2016. doi: 10.4046/trd.2016.79.4.201.

[3] W. Cruz-Knight and L. Blake-Gumbs, “Tuberculosis: An Overview,” *Prim. Care Clin. Off. Pract.*, vol. 40, no. 3, pp. 743–756, Sep. 2013, doi: 10.1016/J.POP.2013.06.003.

[4] R. Sharan and D. Kaushal, “Vaccine strategies for the Mtb/HIV copandemic,” *npj Vaccines 2020 51*, vol. 5, no. 1, pp. 1–10, Oct. 2020, doi: 10.1038/s41541-020-00245-9.

[5] R. M. G. J. Houben and P. J. Dodd, “The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling,” *PLOS Med.*, vol. 13, no. 10, p. e1002152, Oct. 2016, doi: 10.1371/JOURNAL.PMED.1002152.

[6] “Global tuberculosis report 2020.” https://www.who.int/publications/i/item/9789240013131 (accessed Aug. 03, 2022).

[7] “2019 Report on TB Research Funding Trends – Treatment Action Group.” https://www.treatmentactiongroup.org/resources/tbrd-report/tbrd-report-2019/ (accessed Aug. 03, 2022).

[8] H. Fu, J. A. Lewnard, I. Frost, R. Laxminarayan, and N. Arinaminpathy, “Modelling the global burden of drug-resistant tuberculosis avertable by a post-exposure vaccine,” *Nat. Commun.*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/S41467-020-20731-X.

[9] J. Chakaya *et al.*, “Global Tuberculosis Report 2020 - Reflections on the Global TB burden, treatment and prevention efforts,” *Int. J. Infect. Dis.*, vol. 113 Suppl 1, no. Suppl 1, pp. S7–S12, Dec. 2021, doi: 10.1016/J.IJID.2021.02.107.

[10] J. R. Ncayiyana *et al.*, “Prevalence of latent tuberculosis infection and predictive factors in an urban informal settlement in Johannesburg, South Africa: a cross-sectional study,” *BMC Infect. Dis.*, vol. 16, no. 1, Nov. 2016, doi: 10.1186/S12879-016-1989-X.

[11] “GLOBAL STRATEGY FOR TUBERCULOSIS RESEARCH AND INNOVATION 2 0 3 0.” Accessed: Aug. 03, 2022. [Online]. Available: http://www.who.int/tb/Moscow_Declaration_MinisterialConference_TB/en/

[12] “Global Tuberculosis Report. WHO.s.l: WHO 2019”.

[13] “The end TB strategy.” https://www.who.int/publications/i/item/WHO-HTM-TB-2015.19 (accessed Aug. 04, 2022).

[14] K. Weyer, J. Levin, J. Lancaster, J. Brand, and M. Van der Walt, “Determinants of multidrug-resistant tuberculosis (MDR-TB) in South Africa,” *South African Med. J.*, vol. 97, no. 11, p. 1120, Nov. 2007, doi: 10.7196/SAMJ.298.

[15] S. Satyanarayana *et al.*, “An Opportunity to END TB: Using the Sustainable Development Goals for Action on Socio-Economic Determinants of TB in High Burden Countries in WHO South-East Asia and the Western Pacific Regions,” *Trop. Med. Infect. Dis.*, vol. 5, no. 2, Jun. 2020, doi: 10.3390/TROPICALMED5020101.

[16] “END TB BY 2030”, Accessed: Aug. 03, 2022. [Online]. Available: http://apps.who.int/bookorders.

[17] O. Oxlade and M. Murray, “Tuberculosis and Poverty: Why Are the Poor at Greater Risk in India?,” *PLoS One*, vol. 7, no. 11, p. e47533, Nov. 2012, doi: 10.1371/JOURNAL.PONE.0047533.

[18] N. Foster, A. Vassall, S. Cleary, L. Cunnama, G. Churchyard, and E. Sinanovic, "The economic burden of TB diagnosis and treatment in South Africa," *Soc. Sci. Med.*, vol. 130, pp. 42–50, Apr. 2015, doi: 10.1016/J.SOCSCIMED.2015.01.046.
[19] T. Tanimura, E. Jaramillo, D. Weil, M. Raviglione, and K. Lönnroth, "Financial burden for tuberculosis patients in low- and middle-income countries: a systematic review," *Eur. Respir. J.*, vol. 43, no. 6, pp. 1763–1775, 2014, doi: 10.1183/09031936.00193413.
[20] S. Keshavjee and P. E. Farmer, "Tuberculosis, Drug Resistance, and the History of Modern Medicine," *N. Engl. J. Med.*, vol. 367, no. 10, pp. 931–936, Sep. 2012, doi: 10.1056/NEJMRA1205429/SUPPL_FILE/NEJMRA1205429_DISCLOSURES.PDF.
[21] D. Heemskerk, M. Caws, B. Marais, and J. Farrar, *Tuberculosis in adults and children*. 2015. Accessed: Aug. 05, 2022. [Online]. Available: https://library.oapen.org/handle/20.500.12657/32827
[22] R. W.-S. S. M. Journal and undefined 2013, "Tuberculosis 2: Pathophysiology and microbiology of pulmonary tuberculosis," *ajol.info*, vol. 6, no. 1, 2013, Accessed: Aug. 05, 2022. [Online]. Available: https://www.ajol.info/index.php/ssmj/article/view/132586/122185
[23] M. Thillai, K. Pollock, … M. P.-E. R. of, and undefined 2014, "Interferon-gamma release assays for tuberculosis: current and future applications," *Taylor Fr.*, Accessed: Aug. 05, 2022. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1586/17476348.2014.852471
[24] M. Uplekar *et al.*, "WHO's new end TB strategy," *thelancet.com*, vol. 385, pp. 1799–1801, 2015, doi: 10.1016/S0140-6736(15)60570-0.
[25] J. E.-N. R. Immunology and undefined 2012, "The immunological life cycle of tuberculosis," *nature.com*, Accessed: Aug. 05, 2022. [Online]. Available: https://www.nature.com/articles/nri3259
[26] M. Pai *et al.*, "Tuberculosis," *Nat. Rev. Dis. Prim. 2016 21*, vol. 2, no. 1, pp. 1–23, Oct. 2016, doi: 10.1038/nrdp.2016.76.
[27] F. G. J. Cobelens, "False-positive tuberculin reactions due to non-tuberculous mycobacterial infections [5]," *Int. J. Tuberc. Lung Dis.*, vol. 11, no. 8, pp. 934–935, Aug. 2007.
[28] S. Kurz, J. Furin, C. B.-I. D. Clinics, and undefined 2016, "Drug-resistant tuberculosis: challenges and progress," *id.theclinics.com*, Accessed: Aug. 06, 2022. [Online]. Available: https://www.id.theclinics.com/article/S0891-5520(16)30016-2/abstract
[29] E. L. Nuermberger, M. K. Spigelman, and W. W. Yew, "Current Development and Future Prospects in Chemotherapy of Tuberculosis," *Respirology*, vol. 15, no. 5, p. 764, Jul. 2010, doi: 10.1111/J.1440-1843.2010.01775.X.
[30] M. Hijjar, G. Gerhardt, … G. T.-R. de S., and undefined 2007, "Retrospect of tuberculosis control in Brazil," *SciELO Bras.*, vol. 41, 2007, Accessed: Aug. 06, 2022. [Online]. Available: https://www.scielo.br/j/rsp/a/hQdTLVHssMBb86tdQMPhhWR/abstract/?lang=en
[31] J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, and T. A. Funkhouser, "Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure," *PLOS Comput. Biol.*, vol. 5, no. 12, p. e1000585, Dec. 2009, doi: 10.1371/JOURNAL.PCBI.1000585.
[32] J. Caminero, G. Sotgiu, A. Zumla, G. M.-T. L. infectious, and undefined 2010, "Best drug treatment for multidrug-resistant and extensively drug-resistant tuberculosis," *Elsevier*, Accessed: Aug. 06, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1473309910701390?casa_token=UrV6dUvoMQkAAAAA:wJZjNeyuDK--KhmTCa52d7oC0bxF7gPKseYi-J-vcxUUTvywPVdTk6TTXNN-nlYRBDsGhTLiwXU
[33] "Treatment of multidrug-resistant tuberculosis: evidence and contr...: Ingenta Connect."

https://www.ingentaconnect.com/content/iuatld/ijtld/2006/00000010/00000008/art00003 (accessed Aug. 06, 2022).
[34] C. Vilchèze, "Mycobacterial Cell Wall: A Source of Successful Targets for Old and New Drugs," *Appl. Sci. 2020, Vol. 10, Page 2278*, vol. 10, no. 7, p. 2278, Mar. 2020, doi: 10.3390/APP10072278.
[35] S. M. Gygli, S. Borrell, A. Trauner, and S. Gagneux, "Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives," *FEMS Microbiol. Rev.*, vol. 41, no. 3, pp. 354–373, May 2017, doi: 10.1093/FEMSRE/FUX011.
[36] M. Laws, P. Jin, K. R.-T. in microbiology, and undefined 2022, "Efflux pumps in Mycobacterium tuberculosis and their inhibition to tackle antimicrobial resistance," *Elsevier*, Accessed: Aug. 06, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0966842X21001232?casa_token=i5hUlDStf18AAAAA:tb5g8-x3IlLnYkpZIJuUGxD55_jCXhrJlPnDNff9KZYTNrA8e7VPEgBiYhHSym8biwiXdcWohKXT
[37] P. Li, Y. Gu, J. Li, L. Xie, X. Li, and J. Xie, "Mycobacterium tuberculosis Major Facilitator Superfamily Transporters," *J. Membr. Biol.*, vol. 250, no. 6, pp. 573–585, Dec. 2017, doi: 10.1007/S00232-017-9982-X.
[38] E. A. Campbell *et al.*, "Structural mechanism for rifampicin inhibition of bacterial rna polymerase," *Cell*, vol. 104, no. 6, pp. 901–912, Mar. 2001, doi: 10.1016/S0092-8674(01)00286-0.
[39] K. J. Seung, S. Keshavjee, and M. L. Rich, "Multidrug-Resistant Tuberculosis and Extensively Drug-Resistant Tuberculosis," *Cold Spring Harb. Perspect. Med.*, vol. 5, no. 9, Sep. 2015, doi: 10.1101/CSHPERSPECT.A017863.
[40] B. Dagne *et al.*, "The Epidemiology of first and second-line drug-resistance Mycobacterium tuberculosis complex common species: Evidence from selected TB treatment initiating centers in Ethiopia," *PLoS One*, vol. 16, no. 1, p. e0245687, Jan. 2021, doi: 10.1371/JOURNAL.PONE.0245687.
[41] Y. Zhang and W. W. Yew, "Mechanisms of drug resistance in Mycobacterium tuberculosis: Update 2015," *Int. J. Tuberc. Lung Dis.*, vol. 19, no. 11, pp. 1276–1289, Nov. 2015, doi: 10.5588/IJTLD.15.0389.
[42] J. G. Jang and J. H. Chung, "Diagnosis and treatment of multidrug-resistant tuberculosis," *Yeungnam Univ. J. Med.*, vol. 37, no. 4, p. 277, Oct. 2020, doi: 10.12701/YUJM.2020.00626.
[43] S. D. Hamusse, D. Teshome, M. S. Hussen, M. Demissie, and B. Lindtjørn, "Primary and secondary anti-tuberculosis drug resistance in Hitossa District of Arsi Zone, Oromia Regional State, Central Ethiopia," *BMC Public Health*, vol. 16, no. 1, pp. 1–10, Jul. 2016, doi: 10.1186/S12889-016-3210-Y/TABLES/4.
[44] N. Dookie, S. Rambaran, N. Padayatchi, S. Mahomed, and K. Naidoo, "Evolution of drug resistance in Mycobacterium tuberculosis: a review on the molecular determinants of resistance and implications for personalized care," *J. Antimicrob. Chemother.*, vol. 73, no. 5, pp. 1138–1151, May 2018, doi: 10.1093/JAC/DKX506.
[45] P. Brennan, D. C.-C. topics in medicinal chemistry, and undefined 2007, "The cell-wall core of Mycobacterium tuberculosis in the context of drug discovery.," *ingentaconnect.com*, Accessed: Aug. 07, 2022. [Online]. Available: https://www.ingentaconnect.com/content/ben/ctmc/2007/00000007/00000005/art00004
[46] X. Duan, X. Xiang, and J. Xie, "Crucial components of mycobacterium type II fatty acid biosynthesis (Fas-II) and their inhibitors," *FEMS Microbiol. Lett.*, vol. 360, no. 2, pp. 87–99,

Nov. 2014, doi: 10.1111/1574-6968.12597.
[47] J. PaweŁczyk and L. Kremer, "The Molecular Genetics of Mycolic Acid Biosynthesis," *Microbiol. Spectr.*, vol. 2, no. 4, Aug. 2014, doi: 10.1128/MICROBIOLSPEC.MGM2-0003-2013.
[48] L. Kremer *et al.*, "Biochemical characterization of acyl carrier protein (AcpM) and malonyl-CoA:AcpM transacylase (mtFabD), two major components of Mycobacterium tuberculosis fatty acid synthase II," *J. Biol. Chem.*, vol. 276, no. 30, pp. 27967–27974, Jul. 2001, doi: 10.1074/JBC.M103687200.
[49] M. Mathieu *et al.*, "The 1.8 A crystal structure of the dimeric peroxisomal 3-ketoacyl-CoA thiolase of Saccharomyces cerevisiae: implications for substrate binding and reaction mechanism," *J. Mol. Biol.*, vol. 273, no. 3, pp. 714–728, Oct. 1997, doi: 10.1006/JMBI.1997.1331.
[50] S. R. Luckner, C. A. Machutta, P. J. Tonge, and C. Kisker, "Crystal structures of Mycobacterium tuberculosis KasA show mode of action within cell wall biosynthesis and its inhibition by thiolactomycin," *Structure*, vol. 17, no. 7, pp. 1004–1013, Jul. 2009, doi: 10.1016/J.STR.2009.04.012.
[51] H. Zhang, C. A. Machutta, and P. J. Tonge, "8.07 Fatty Acid Biosynthesis and Oxidation," *booksite.elsevier.com*, Accessed: Aug. 08, 2022. [Online]. Available: https://booksite.elsevier.com/brochures/conap2/PDFs/Vol8FattyAcidBiosynthesis-and-Oxidation.pdf
[52] J. Schiebel *et al.*, "Structural Basis for the Recognition of Mycolic Acid Precursors by KasA, a Condensing Enzyme and Drug Target from Mycobacterium Tuberculosis," *J. Biol. Chem.*, vol. 288, no. 47, p. 34190, Nov. 2013, doi: 10.1074/JBC.M113.511436.
[53] R. A. Slayden and C. E. Barry, "The role of KasA and KasB in the biosynthesis of meromycolic acids and isoniazid resistance in Mycobacterium tuberculosis," *Tuberculosis (Edinb).*, vol. 82, no. 4–5, pp. 149–160, 2002, doi: 10.1054/TUBE.2002.0333.
[54] R. Heath, C. R.-N. product reports, and undefined 2002, "The Claisen condensation in biology," *pubs.rsc.org*, Accessed: Aug. 08, 2022. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2002/np/b110221b?casa_token=dO8QpWu-e0wAAAAA:g_vtMf1pWIFFIJ-7j0RPHrpliRr8R9NiA6b46Ria6eSrhZeqnmMCAcDJPbpcCCs9uwsztwrDYeYHVWe_
[55] R. J. Heath and C. O. Rock, "Fatty acid biosynthesis as a target for novel antibacterials," *Curr. Opin. Investig. Drugs*, vol. 5, no. 2, p. 146, Feb. 2004, Accessed: Aug. 08, 2022. [Online]. Available: /pmc/articles/PMC1618763/
[56] A. Bhatt, V. Molle, G. S. Besra, W. R. Jacobs, and L. Kremer, "The Mycobacterium tuberculosis FAS-II condensing enzymes: their role in mycolic acid biosynthesis, acid-fastness, pathogenesis and in future drug development," *Mol. Microbiol.*, vol. 64, no. 6, pp. 1442–1454, Jun. 2007, doi: 10.1111/J.1365-2958.2007.05761.X.
[57] L. Kremer *et al.*, "Thiolactomycin and related analogues as novel anti-mycobacterial agents targeting KasA and KasB condensing enzymes in Mycobacterium tuberculosis," *J. Biol. Chem.*, vol. 275, no. 22, pp. 16857–16864, Jun. 2000, doi: 10.1074/JBC.M000569200.
[58] A. K. Brown, R. C. Taylor, A. Bhatt, K. Fütterer, and G. S. Besra, "Platensimycin Activity against Mycobacterial β-Ketoacyl-ACP Synthases," *PLoS One*, vol. 4, no. 7, p. e6306, Jul. 2009, doi: 10.1371/JOURNAL.PONE.0006306.
[59] N. M. Parrish, F. P. Kuhajda, H. S. Heine, W. R. Bishai, and J. D. Dick, "Antimycobacterial activity of cerulenin and its effects on lipid biosynthesis," *J. Antimicrob. Chemother.*, vol. 43, no. 2, pp. 219–226, Feb. 1999, doi: 10.1093/JAC/43.2.219.

[60] M. L. Schaeffer, G. Agnihotri, C. Volker, H. Kallender, P. J. Brennan, and J. T. Lonsdale, "Purification and biochemical characterization of the Mycobacterium tuberculosis beta-ketoacyl-acyl carrier protein synthases KasA and KasB," *J. Biol. Chem.*, vol. 276, no. 50, pp. 47029–47037, Dec. 2001, doi: 10.1074/JBC.M108903200.
[61] D. R. Durairaj and P. Shanmughavel, "In Silico Drug Design of Thiolactomycin Derivatives Against Mtb-KasA Enzyme to Inhibit Multidrug Resistance of Mycobacterium tuberculosis," *Interdiscip. Sci.*, vol. 11, no. 2, pp. 215–225, Jun. 2019, doi: 10.1007/S12539-017-0257-0.
[62] L. Kremer *et al.*, "Mycolic acid biosynthesis and enzymic characterization of the beta-ketoacyl-ACP synthase A-condensing enzyme from Mycobacterium tuberculosis.," *Biochem. J.*, vol. 364, no. Pt 2, p. 423, Jun. 2002, doi: 10.1042/BJ20011628.
[63] S. Pushpakom *et al.*, "Drug repurposing: progress, challenges and recommendations," *nature.com*, vol. 18, 2018, doi: 10.1038/nrd.2018.168.
[64] M. Rudrapal, … S. K.-, M. A. and, and undefined 2020, "Drug repurposing (DR): an emerging approach in drug discovery," *books.google.com*, Accessed: Aug. 08, 2022. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=_JQtEAAAQBAJ&oi=fnd&pg=PA3&dq=66.%09Rudrapal,+M.,+J.+Khairnar,+and+G.+Jadhav,+Drug+repurposing+(DR):+an+emerging+approach+in+drug+discovery.+Drug+Repurposing+Hypothesis+Mol.+Asp.+Ther.+Appl,+2020.&ots=_UmpvVY35o&sig=vhowYIvjBnN0l_DD2e22YTuEhVU
[65] Q. An, C. Li, Y. Chen, Y. Deng, T. Yang, and Y. Luo, "Repurposed drug candidates for antituberculosis therapy," *Eur. J. Med. Chem.*, vol. 192, p. 112175, Apr. 2020, doi: 10.1016/J.EJMECH.2020.112175.
[66] M. Mori, S. Villa, S. Ciceri, D. Colombo, P. Ferraboschi, and F. Meneghetti, "An Outline of the Latest Crystallographic Studies on Inhibitor-Enzyme Complexes for the Design and Development of New Therapeutics against Tuberculosis," *Molecules*, vol. 26, no. 23, Dec. 2021, doi: 10.3390/MOLECULES26237082.
[67] H. Marrakchi, M. Lanéelle, M. D.-C. & biology, and undefined 2014, "Mycolic acids: structures, biosynthesis, and beyond," *Elsevier*, Accessed: Aug. 09, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1074552113004201
[68] J. Asselineau, E. L.- Nature, and undefined 1950, "Structure of the mycolic acids of mycobacteria," *nature.com*, Accessed: Aug. 09, 2022. [Online]. Available: https://www.nature.com/articles/166782a0
[69] Y. Wang, H. Wu, and Y. Cai, "A benchmark study of sequence alignment methods for protein clustering," *BMC Bioinformatics*, vol. 19, Dec. 2018, doi: 10.1186/S12859-018-2524-4.
[70] P. Bork, E. K.-C. opinion in structural biology, and undefined 1996, "Protein sequence motifs," *Elsevier*, Accessed: Aug. 09, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959440X96800571
[71] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, Mar. 1981, doi: 10.1016/0022-2836(81)90087-5.
[72] V. L.-7th M. B. Course, B. Molecular, and undefined 2008, "The Needleman-Wunsch algorithm for sequence alignment," *cs.sjsu.edu*, Accessed: Aug. 09, 2022. [Online]. Available: https://www.cs.sjsu.edu/~aid/cs152/NeedlemanWunsch.pdf
[73] R. M. Schwartz and M. O. Dayhoff, "Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts," *Science (80-.).*, vol. 199, no. 4327, pp. 395–403, 1978, doi: 10.1126/SCIENCE.202030.
[74] S. Henikoff, J. H.-A. in protein chemistry, and undefined 2000, "Amino acid substitution matrices," *books.google.com*, Accessed: Aug. 09, 2022. [Online]. Available:

https://books.google.com/books?hl=en&lr=&id=9p3E2sS1aJUC&oi=fnd&pg=PA73&dq=75.%09Henikoff,+S.+and+J.G.+Henikoff,+Performance+evaluation+of+amino+acid+substitution+matrices.+Proteins:+Structure,+Function,+and+Bioinformatics,+1993.+17(1):+p.+49-61.&ots=eMZkwjziUc&sig=1UsGxdWasiPtnQXEi4rM4JHriCI

[75] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 89, no. 22, pp. 10915–10919, 1992, doi: 10.1073/PNAS.89.22.10915.

[76] T. Müller, R. Spang, M. V.-M. biology and evolution, and undefined 2002, "Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method," *academic.oup.com*, Accessed: Aug. 09, 2022. [Online]. Available: https://academic.oup.com/mbe/article-abstract/19/1/8/1066678

[77] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, "NCBI BLAST: a better web interface," *Nucleic Acids Res.*, vol. 36, no. Web Server issue, 2008, doi: 10.1093/NAR/GKN201.

[78] S. F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997, doi: 10.1093/NAR/25.17.3389.

[79] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 85, no. 8, pp. 2444–2448, 1988, doi: 10.1073/PNAS.85.8.2444.

[80] J. Söding, A. Biegert, and A. N. Lupas, "The HHpred interactive server for protein homology detection and structure prediction," *Nucleic Acids Res.*, vol. 33, no. Web Server issue, Jul. 2005, doi: 10.1093/NAR/GKI408.

[81] F. Plewniak, "Database similarity searches," *Methods Mol. Biol.*, vol. 484, pp. 361–378, 2008, doi: 10.1007/978-1-59745-398-1_24.

[82] J. Daugelaite, A. O' Driscoll, and R. D. Sleator, "An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics," *ISRN Biomath.*, vol. 2013, pp. 1–14, Aug. 2013, doi: 10.1155/2013/615630.

[83] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Mol. Biol. Evol.*, vol. 4, no. 4, pp. 406–425, 1987, doi: 10.1093/OXFORDJOURNALS.MOLBEV.A040454.

[84] I. Gronau, S. M.-I. P. Letters, and undefined 2007, "Optimal implementations of UPGMA and other common clustering algorithms," *Elsevier*, 2007, Accessed: Aug. 09, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020019007001767

[85] F. Sievers, A. Wilm, D. Dineen, … T. G.-M. systems, and undefined 2011, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *embopress.org*, Accessed: Aug. 09, 2022. [Online]. Available: https://www.embopress.org/doi/abs/10.1038/Msb.2011.75

[86] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, Apr. 2013, doi: 10.1093/MOLBEV/MST010.

[87] K. Katoh, K. Misawa, K. I. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, doi: 10.1093/NAR/GKF436.

[88] J. Pei, B. H. Kim, and N. V. Grishin, "PROMALS3D: a tool for multiple protein sequence and structure alignments," *Nucleic Acids Res.*, vol. 36, no. 7, p. 2295, Apr. 2008, doi: 10.1093/NAR/GKN072.

[89] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high

throughput," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004, doi: 10.1093/NAR/GKH340.
[90] J. Thompson, P. Koehl, R. R.-… : Structure, undefined Function, undefined and, and undefined 2005, "BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark," *Wiley Online Libr.*, vol. 61, no. 1, pp. 127–136, Oct. 2005, doi: 10.1002/prot.20527.
[91] I. Van Walle, I. Lasters, L. W.- Bioinformatics, and undefined 2005, "SABmark—a benchmark for sequence alignment that covers the entire known fold space," *academic.oup.com*, Accessed: Aug. 26, 2022. [Online]. Available: https://academic.oup.com/bioinformatics/article-abstract/21/7/1267/268759
[92] J. M. Chang, P. Di Tommaso, V. Lefort, O. Gascuel, and C. Notredame, "TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W3–W6, Jul. 2015, doi: 10.1093/NAR/GKV310.
[93] J. Xiong, *Essential bioinformatics*. 2006. Accessed: Aug. 11, 2022. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=AFsu7_goA8kC&oi=fnd&pg=PA14&dq=105.%09Xiong,+J.,+Essential+bioinformatics.+2006:+Cambridge+University+Press.&ots=hIznwQ8wkh&sig=fwDvdJfosHYQ4GeAoqLWBzj12EA
[94] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert, "Approaches to the automatic discovery of patterns in biosequences," *J. Comput. Biol.*, vol. 5, no. 2, pp. 279–305, 1998, doi: 10.1089/CMB.1998.5.279.
[95] S. Henikoff, J. H.-J. of molecular biology, and undefined 1994, "Position-based sequence weights," *Elsevier*, Accessed: Aug. 11, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/0022283694900329
[96] A. X. L. Id *et al.*, "Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning," *PLOS Comput. Biol.*, vol. 18, no. 6, p. e1010238, Jun. 2022, doi: 10.1371/JOURNAL.PCBI.1010238.
[97] A. Bairoch, "PROSITE: a dictionary of sites and patterns in proteins," *Nucleic Acids Res.*, vol. 19 Suppl, no. Suppl, p. 2241, Apr. 1991, doi: 10.1093/NAR/19.SUPPL.2241.
[98] E. L. Sonnhammer, S. R. Eddy, and R. Durbin, "Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments", doi: 10.1002/(SICI)1097-0134(199707)28:3.
[99] T. Bailey, J. Johnson, C. G.-… acids research, and undefined 2015, "The MEME suite," *academic.oup.com*, Accessed: Aug. 11, 2022. [Online]. Available: https://academic.oup.com/nar/article-abstract/43/W1/W39/2467905
[100] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in bipolymers," 1994, Accessed: Aug. 11, 2022. [Online]. Available: http://www.cs.toronto.edu/~brudno/csc2417_15/10.1.1.121.7056.pdf
[101] T. Sanyanga, B. Nizami, Ö. T. B.- Molecules, and undefined 2019, "Mechanism of Action of Non-Synonymous Single Nucleotide Variations Associated with α-Carbonic Anhydrase II Deficiency," *mdpi.com*, doi: 10.3390/molecules24213987.
[102] M. Charleston, "Phylogeny," *Brenner's Encycl. Genet. Second Ed.*, pp. 324–325, Jan. 2013, doi: 10.1016/B978-0-12-374984-0.01160-8.
[103] B. G. Hall, "Building phylogenetic trees from molecular data with MEGA," *Mol. Biol. Evol.*, vol. 30, no. 5, pp. 1229–1235, May 2013, doi: 10.1093/MOLBEV/MST012.
[104] D. H.-C. Biology and undefined 1997, "Phylogenetic analysis," *cell.com*, Accessed: Aug. 11, 2022. [Online]. Available: https://www.cell.com/current-biology/pdf/S0960-9822(97)70070-8.pdf
[105] P. A.-N. Biology and undefined 2016, "Molecular phylogenetics: Concepts for a newcomer,"

Springer, vol. 160, pp. 185–196, 2017, doi: 10.1007/10_2016_49.
[106] K. Tamura, G. Stecher, and S. Kumar, "MEGA11: Molecular Evolutionary Genetics Analysis Version 11," *Mol. Biol. Evol.*, vol. 38, no. 7, pp. 3022–3027, Jun. 2021, doi: 10.1093/MOLBEV/MSAB120.
[107] F. Ronquist *et al.*, "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space," *Syst. Biol.*, vol. 61, no. 3, pp. 539–542, May 2012, doi: 10.1093/SYSBIO/SYS029.
[108] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, May 2014, doi: 10.1093/BIOINFORMATICS/BTU033.
[109] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut, "Bayesian Phylogenetics with BEAUti and the BEAST 1.7," *Mol. Biol. Evol.*, vol. 29, no. 8, pp. 1969–1973, Aug. 2012, doi: 10.1093/MOLBEV/MSS075.
[110] A. Bateman, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/NAR/GKY1049.
[111] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.
[112] H. Berman, J. Westbrook, Z. F.-… acids research, and undefined 2000, "The protein data bank," *academic.oup.com*, Accessed: Aug. 12, 2022. [Online]. Available: https://academic.oup.com/nar/article-abstract/28/1/235/2384399
[113] W. D.-C. N. P. Crystallogr and undefined 2002, "Pymol: An open-source molecular graphics tool," *Citeseer*, Accessed: Aug. 12, 2022. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.231.5879&rep=rep1&type=pdf#page=44
[114] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton, "Jalview Version 2--a multiple sequence alignment editor and analysis workbench," *Bioinformatics*, vol. 25, no. 9, pp. 1189–1191, 2009, doi: 10.1093/BIOINFORMATICS/BTP033.
[115] T. Allan Sanyanga, B. Nizami, and Ö. T. Bishop, "Mechanism of Action of Non-Synonymous Single Nucleotide Variations Associated with α-Carbonic Anhydrase II Deficiency," *Molecules*, vol. 24, no. 21, Nov. 2019, doi: 10.3390/MOLECULES24213987.
[116] T. L. Bailey and M. Gribskov, "Combining evidence using p-values: application to sequence homology searches," *Bioinformatics*, vol. 14, no. 1, pp. 48–54, 1998, doi: 10.1093/BIOINFORMATICS/14.1.48.
[117] S. Q. Le and O. Gascuel, "An Improved General Amino Acid Replacement Matrix," *Mol. Biol. Evol.*, vol. 25, no. 7, pp. 1307–1320, Jul. 2008, doi: 10.1093/MOLBEV/MSN067.
[118] S. Whelan and N. Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach," *Mol. Biol. Evol.*, vol. 18, no. 5, pp. 691–699, 2001, doi: 10.1093/OXFORDJOURNALS.MOLBEV.A003851.
[119] D. W. Nyamai and Ö. Tastan Bishop, "Aminoacyl tRNA synthetases as malarial drug targets: a comparative bioinformatics study," *Malar. J. 2019 181*, vol. 18, no. 1, pp. 1–27, Feb. 2019, doi: 10.1186/S12936-019-2665-6.
[120] O. Emanuelsson, H. Nielsen, S. Brunak, and G. Von Heijne, "Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence," *J. Mol. Biol.*, vol. 300, no. 4, pp. 1005–1016, Jul. 2000, doi: 10.1006/JMBI.2000.3903.
[121] C. E. Christensen, B. B. Kragelund, P. von Wettstein-Knowles, and A. Henriksen, "Structure of the human β-ketoacyl [ACP] synthase from the mitochondrial type II fatty acid synthase,"

Protein Sci., vol. 16, no. 2, p. 261, Dec. 2007, doi: 10.1110/PS.062473707.
[122] T. M. Musyoka, A. M. Kanzi, K. A. Lobb, and Ö. Tastan Bishop, "Analysis of non-peptidic compounds as potential malarial inhibitors against Plasmodial cysteine proteases via integrated virtual screening workflow," *J. Biomol. Struct. Dyn.*, vol. 34, no. 10, pp. 2084–2101, Oct. 2016, doi: 10.1080/07391102.2015.1108231.
[123] A. F. Abdel-Magid, "Allosteric Modulators: An Emerging Concept in Drug Discovery," *ACS Med. Chem. Lett.*, vol. 6, no. 2, p. 104, Feb. 2015, doi: 10.1021/ML5005365.
[124] T. Kenakin, *A pharmacology primer: theory, application and methods*. 2009. Accessed: Aug. 26, 2022. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=8yM2vAwRmj8C&oi=fnd&pg=PP1&dq=132.%09Kenakin,+T.P.,+Chapter+7+-+Allosteric+Drug+Antagonism,+in+A+Pharmacology+Primer+(Third+Edition),+T.P.+Kenakin,+Editor.+2009,+Academic+Press:+New+York.+p.+129-147.&ots=vF3J8D_qeU&sig=VuENrUQna4ZXaTDBahj_93AfcrA
[125] O. S. Amamuddy *et al.*, "Integrated computational approaches and tools for allosteric drug discovery," *mdpi.com*, doi: 10.3390/ijms21030847.
[126] A. G.-M. P. and Practice and undefined 2013, "Use of allosteric targets in the discovery of safer drugs," *karger.com*, Accessed: Aug. 26, 2022. [Online]. Available: https://www.karger.com/Article/Abstract/350417
[127] Y. Xu *et al.*, "CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction," *academic.oup.com*, Accessed: Aug. 28, 2022. [Online]. Available: https://academic.oup.com/nar/article-abstract/46/W1/W374/4994680
[128] R. Harris, A. J. Olson, and D. S. Goodsell, "Automated prediction of ligand-binding sites in proteins," *Proteins*, vol. 70, no. 4, pp. 1506–1517, Mar. 2008, doi: 10.1002/PROT.21645.
[129] A. Volkamer, D. Kuhn, F. Rippmann, and M. Rarey, "DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment," *Bioinformatics*, vol. 28, no. 15, pp. 2074–2075, Aug. 2012, doi: 10.1093/BIOINFORMATICS/BTS310.
[130] T. A. Halgren, "Identifying and characterizing binding sites and assessing druggability," *J. Chem. Inf. Model.*, vol. 49, no. 2, pp. 377–389, Feb. 2009, doi: 10.1021/CI800324M.
[131] T. Halgren, "New method for fast and accurate binding-site identification and analysis," *Chem. Biol. Drug Des.*, vol. 69, no. 2, pp. 146–148, Feb. 2007, doi: 10.1111/J.1747-0285.2007.00483.X.
[132] R. Nussinov and C. J. Tsai, "Allostery in disease and in drug discovery," *Cell*, vol. 153, no. 2, pp. 293–305, Apr. 2013, doi: 10.1016/J.CELL.2013.03.034.
[133] J. P. Changeux, "The concept of allosteric modulation: an overview," *Drug Discov. Today. Technol.*, vol. 10, no. 2, 2013, doi: 10.1016/J.DDTEC.2012.07.007.
[134] G. Collier and V. Ortiz, "Emerging computational approaches for the study of protein allostery," *Arch. Biochem. Biophys.*, vol. 538, no. 1, pp. 6–15, 2013, doi: 10.1016/J.ABB.2013.07.025.
[135] E. Özkan, H. Yu, and J. Deisenhofer, "Mechanistic insight into the allosteric activation of a ubiquitin-conjugating enzyme by RING-type ubiquitin ligases," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 52, pp. 18890–18895, Dec. 2005, doi: 10.1073/PNAS.0509418102/SUPPL_FILE/09418FIG9.JPG.
[136] N. Sinha and R. Nussinov, "Point mutations and sequence variability in proteins: redistributions of preexisting populations," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 6, pp. 3139–3144, Mar. 2001, doi: 10.1073/PNAS.051399098.

[137] E. A. Wold and J. Zhou, "GPCR Allosteric Modulators: Mechanistic Advantages and Therapeutic Applications," *Curr. Top. Med. Chem.*, vol. 18, no. 23, p. 2002, Jan. 2018, doi: 10.2174/1568026619999190101151837.
[138] W. Jiao and E. J. Parker, "Using a combination of computational and experimental techniques to understand the molecular basis for protein allostery," *Adv. Protein Chem. Struct. Biol.*, vol. 87, pp. 391–413, 2012, doi: 10.1016/B978-0-12-398312-1.00013-5.
[139] D. E. Koshland, J. G. Nemethy, and D. Filmer, "Comparison of experimental binding data and theoretical models in proteins containing subunits," *Biochemistry*, vol. 5, no. 1, pp. 365–385, Jan. 1966, doi: 10.1021/BI00865A047.
[140] S. Lu, S. Li, and J. Zhang, "Harnessing allostery: a novel approach to drug discovery," *Med. Res. Rev.*, vol. 34, no. 6, pp. 1242–1285, Nov. 2014, doi: 10.1002/MED.21317.
[141] X. Li *et al.*, "Toward an understanding of the sequence and structural basis of allosteric proteins," *J. Mol. Graph. Model.*, vol. 40, pp. 30–39, Mar. 2013, doi: 10.1016/J.JMGM.2012.12.011.
[142] J. Zheng, B. S. Avvaru, C. Tu, R. McKenna, and D. N. Silverman, "Role of Hydrophilic Residues in Proton Transfer During Catalysis by Human Carbonic Anhydrase II," *Biochemistry*, vol. 47, no. 46, p. 12028, Nov. 2008, doi: 10.1021/BI801473W.
[143] J. S. Yang, S. W. Seo, S. Jang, G. Y. Jung, and S. Kim, "Rational Engineering of Enzyme Allosteric Regulation through Sequence Evolution Analysis," *PLOS Comput. Biol.*, vol. 8, no. 7, p. e1002612, Jul. 2012, doi: 10.1371/JOURNAL.PCBI.1002612.
[144] R. Nussinov and C.-J. Tsai, "The different ways through which specificity works in orthosteric and allosteric drugs," *Curr. Pharm. Des.*, vol. 18, no. 9, pp. 1311–1316, Feb. 2012, doi: 10.2174/138161212799436377.
[145] R. Nussinov and C. J. Tsai, "The design of covalent allosteric drugs," *Annu. Rev. Pharmacol. Toxicol.*, vol. 55, pp. 249–267, Jan. 2015, doi: 10.1146/ANNUREV-PHARMTOX-010814-124401.
[146] S. Lu, W. Huang, and J. Zhang, "Recent computational advances in the identification of allosteric sites in proteins," *Drug Discov. Today*, vol. 19, no. 10, pp. 1595–1600, 2014, doi: 10.1016/J.DRUDIS.2014.07.012.
[147] P. J. Goodford, "A computational procedure for determining energetically favorable binding sites on biologically important macromolecules," *J. Med. Chem.*, vol. 28, no. 7, pp. 849–857, 1985, doi: 10.1021/JM00145A002.
[148] M. P. Jacobson *et al.*, "A hierarchical approach to all-atom protein loop prediction," *Proteins*, vol. 55, no. 2, pp. 351–367, May 2004, doi: 10.1002/PROT.10613.
[149] M. P. Jacobson, R. A. Friesner, Z. Xiang, and B. Honig, "On the role of the crystal environment in determining protein side-chain conformations," *J. Mol. Biol.*, vol. 320, no. 3, pp. 597–608, 2002, doi: 10.1016/S0022-2836(02)00470-9.
[150] A. Šali, L. Potterton, F. Yuan, H. van Vlijmen, and M. Karplus, "Evaluation of comparative protein modeling by MODELLER," *Proteins*, vol. 23, no. 3, pp. 318–326, 1995, doi: 10.1002/PROT.340230306.
[151] G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, and W. Sherman, "Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments," *J. Comput. Aided. Mol. Des.*, vol. 27, no. 3, pp. 221–234, Mar. 2013, doi: 10.1007/S10822-013-9644-8.
[152] H. Li, A. D. Robertson, and J. H. Jensen, "Very fast empirical prediction and rationalization of protein pKa values," *Proteins*, vol. 61, no. 4, pp. 704–721, Dec. 2005, doi: 10.1002/PROT.20660.

[153] M. Rostkowski, M. H. Olsson, C. R. Søndergaard, and J. H. Jensen, "Graphical analysis of pH-dependent properties of proteins predicted using PROPKA," *BMC Struct. Biol.*, vol. 11, no. 1, pp. 1–6, Jan. 2011, doi: 10.1186/1472-6807-11-6/FIGURES/3.
[154] X. Ma, H. Meng, and L. Lai, "Motions of Allosteric and Orthosteric Ligand-Binding Sites in Proteins are Highly Correlated," *J. Chem. Inf. Model.*, vol. 56, no. 9, pp. 1725–1733, Sep. 2016, doi: 10.1021/ACS.JCIM.6B00039.
[155] G. M. Morris *et al.*, "Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function," *J. Comput. Chem.*, vol. 19, no. 14, 1639.
[156] Y. Yuan, J. Pei, and L. Lai, "Binding site detection and druggability prediction of protein targets for structure-based drug design," *Curr. Pharm. Des.*, vol. 19, no. 12, pp. 2326–2333, Mar. 2013, doi: 10.2174/1381612811319120019.
[157] T. Cheng, Q. Li, Z. Zhou, Y. Wang, and S. H. Bryant, "Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review," *AAPS J.*, vol. 14, no. 1, p. 133, Mar. 2012, doi: 10.1208/S12248-012-9322-0.
[158] P. D. Lyne, "Structure-based virtual screening: An overview," *Drug Discov. Today*, vol. 7, no. 20, pp. 1047–1055, Oct. 2002, doi: 10.1016/S1359-6446(02)02483-2.
[159] Z. Wang *et al.*, "Combined strategies in structure-based virtual screening," *Phys. Chem. Chem. Phys.*, vol. 22, no. 6, pp. 3149–3159, Feb. 2020, doi: 10.1039/C9CP06303J.
[160] E. Lionta, G. Spyrou, D. Vassilatis, and Z. Cournia, "Structure-based virtual screening for drug discovery: principles, applications and recent advances," *Curr. Top. Med. Chem.*, vol. 14, no. 16, pp. 1923–1938, Oct. 2014, doi: 10.2174/1568026614666140929124445.
[161] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018, doi: 10.1093/NAR/GKX1037.
[162] O. Sheik Amamuddy, T. M. Musyoka, R. A. Boateng, S. Zabo, and Ö. Tastan Bishop, "Determining the unbinding events and conserved motions associated with the pyrazinamide release due to resistance mutations of Mycobacterium tuberculosis pyrazinamidase," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1103–1120, Jan. 2020, doi: 10.1016/J.CSBJ.2020.05.009.
[163] O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, p. 455, Jan. 2010, doi: 10.1002/JCC.21334.
[164] G. M. Morris *et al.*, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, Dec. 2009, doi: 10.1002/JCC.21256.
[165] N. Singh, L. Chaput, and B. O. Villoutreix, "Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace," *Brief. Bioinform.*, vol. 22, no. 2, pp. 1790–1818, Mar. 2021, doi: 10.1093/BIB/BBAA034.
[166] D. Rognan, "The impact of in silico screening in the discovery of novel and safer drug candidates," *Pharmacol. Ther.*, vol. 175, pp. 47–66, Jul. 2017, doi: 10.1016/J.PHARMTHERA.2017.02.034.
[167] O. Taboureau, J. B. Baell, J. Fernández-Recio, and B. O. Villoutreix, "Established and Emerging Trends in Computational Drug Discovery in the Structural Genomics Era," *Chem. Biol.*, vol. 19, no. 1, pp. 29–41, Jan. 2012, doi: 10.1016/J.CHEMBIOL.2011.12.007.
[168] Y. Tanrikulu, B. Krüger, and E. Proschak, "The holistic integration of virtual screening in drug discovery," *Drug Discov. Today*, vol. 18, no. 7–8, pp. 358–364, Apr. 2013, doi: 10.1016/J.DRUDIS.2013.01.007.
[169] D. Stumpfe and J. Bajorath, "Current Trends, Overlooked Issues, and Unmet Challenges in

Virtual Screening," *J. Chem. Inf. Model.*, vol. 60, no. 9, pp. 4112–4115, Sep. 2020, doi: 10.1021/ACS.JCIM.9B01101.
[170] J. Li, A. Fu, and L. Zhang, "An Overview of Scoring Functions Used for Protein-Ligand Interactions in Molecular Docking," *Interdiscip. Sci.*, vol. 11, no. 2, pp. 320–328, Jun. 2019, doi: 10.1007/S12539-019-00327-W.
[171] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, "Molecular Docking: A powerful approach for structure-based drug discovery," *Curr. Comput. Aided. Drug Des.*, vol. 7, no. 2, p. 146, Jun. 2011, doi: 10.2174/157340911795677602.
[172] S. Agarwal and R. Mehrotra, "An overview of Molecular Docking," 2016.
[173] V. Mohan, A. Gibbs, M. Cummings, E. Jaeger, and R. DesJarlais, "Docking: successes and challenges," *Curr. Pharm. Des.*, vol. 11, no. 3, pp. 323–333, Mar. 2005, doi: 10.2174/1381612053382106.
[174] N. S. Pagadala, K. Syed, and J. Tuszynski, "Software for molecular docking: a review," *Biophys. Rev.*, vol. 9, no. 2, pp. 91–102, Apr. 2017, doi: 10.1007/S12551-016-0247-1.
[175] R. A. Friesner *et al.*, "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy," *J. Med. Chem.*, vol. 47, no. 7, pp. 1739–1749, Mar. 2004, doi: 10.1021/JM0306430.
[176] W. J. Allen *et al.*, "DOCK 6: Impact of new features and current docking performance," *J. Comput. Chem.*, vol. 36, no. 15, pp. 1132–1156, Jun. 2015, doi: 10.1002/JCC.23905.
[177] G. M. Morris, R. Huey, and A. J. Olson, "Using AutoDock for ligand-receptor docking," *Curr. Protoc. Bioinforma.*, vol. Chapter 8, no. SUPPL. 24, 2008, doi: 10.1002/0471250953.BI0814S24.
[178] R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell, "A semiempirical free energy force field with charge-based desolvation," *J. Comput. Chem.*, vol. 28, no. 6, pp. 1145–1152, Apr. 2007, doi: 10.1002/JCC.20634.
[179] D. S. Wishart *et al.*, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Res.*, vol. 36, no. Database issue, Jan. 2008, doi: 10.1093/NAR/GKM958.
[180] B. N. Diallo, "In silico study of Plasmodium 1-deoxy-dxylulose 5-phosphate reductoisomerase (DXR) for identification of novel inhibitors from SANCDB," 2018.
[181] R. Huey and G. M. Morris, "Using AutoDock with AutoDockTools: A Tutorial," 2003.
[182] P. Gowo, "In silico identification of selective novel hits against the active site of wild type mycobacterium tuberculosis pyrazinamidase and its mutants," 2021.
[183] A. Samdani and U. Vetrivel, "[논문]POAP: A GNU parallel based multithreaded pipeline of open babel and AutoDock suite for boosted high throughput virtual screening," *Comput. Biol. Chem.*, vol. 74, pp. 39–48, Jun. 2018, doi: 10.1016/J.COMPBIOLCHEM.2018.02.012.
[184] J. M. Perkel, "Why Jupyter is data scientists' computational notebook of choice," *Nature*, vol. 563, no. 7729, pp. 145–146, Nov. 2018, doi: 10.1038/D41586-018-07196-1.
[185] H. M. Mengist, T. Dilnessa, and T. Jin, "Structural Basis of Potential Inhibitors Targeting SARS-CoV-2 Main Protease," *Front. Chem.*, vol. 9, p. 7, Mar. 2021, doi: 10.3389/FCHEM.2021.622898/XML/NLM.
[186] T. Pantsar and A. Poso, "Binding Affinity via Docking: Fact and Fiction," *Mol. A J. Synth. Chem. Nat. Prod. Chem.*, vol. 23, no. 8, p. 1DUMMY, 2018, doi: 10.3390/MOLECULES23081899.
[187] H. Patel, X. Lucas, I. Bendik, S. Günther, and I. Merfort, "Target Fishing by Cross-Docking to Explain Polypharmacological Effects," *ChemMedChem*, vol. 10, no. 7, pp. 1209–1217, Jul. 2015, doi: 10.1002/CMDC.201500123.

SUPPLEMENTARY MATERIAL

Figure S1: Multiple sequence alignment as predicted by PROMALS3D. The structural information was obtained from PDB ID: 6P9K.

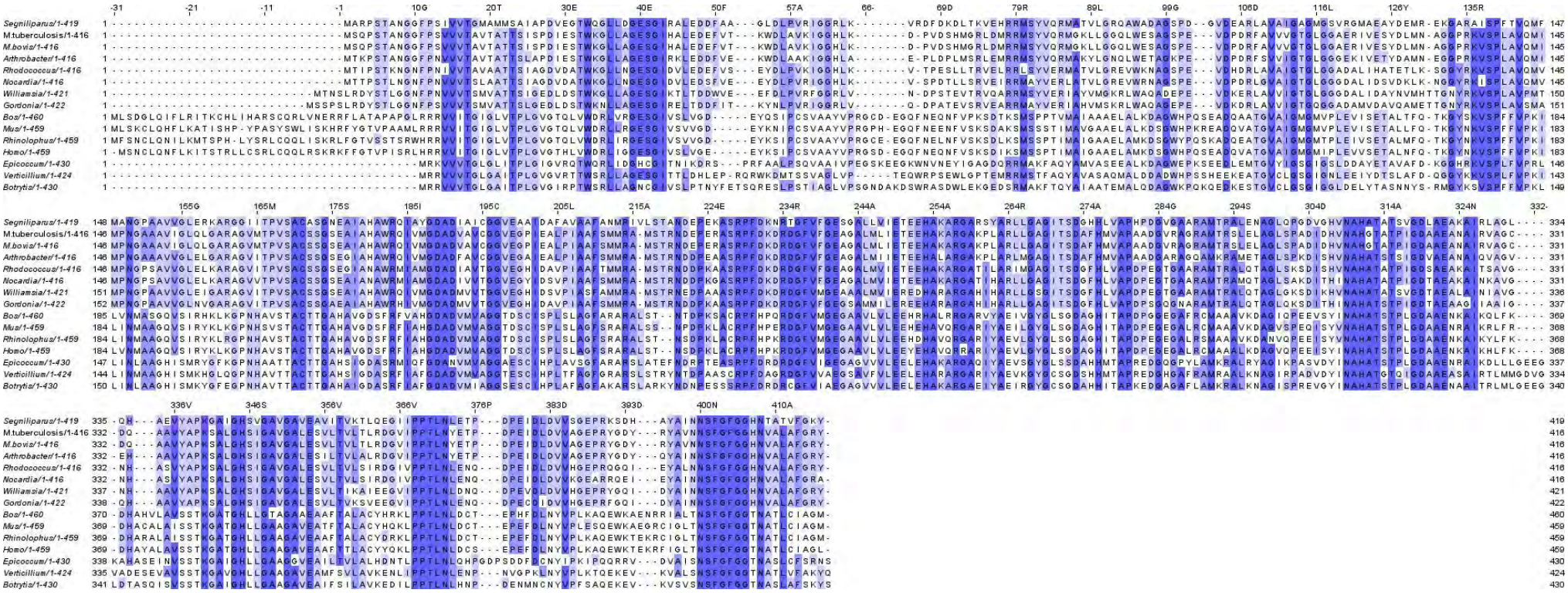


Figure S2: Clustal Omega alignment of KasA homologs. Residue numbering is given for *M. tuberculosis*. The alignment is colored by percentage identity.

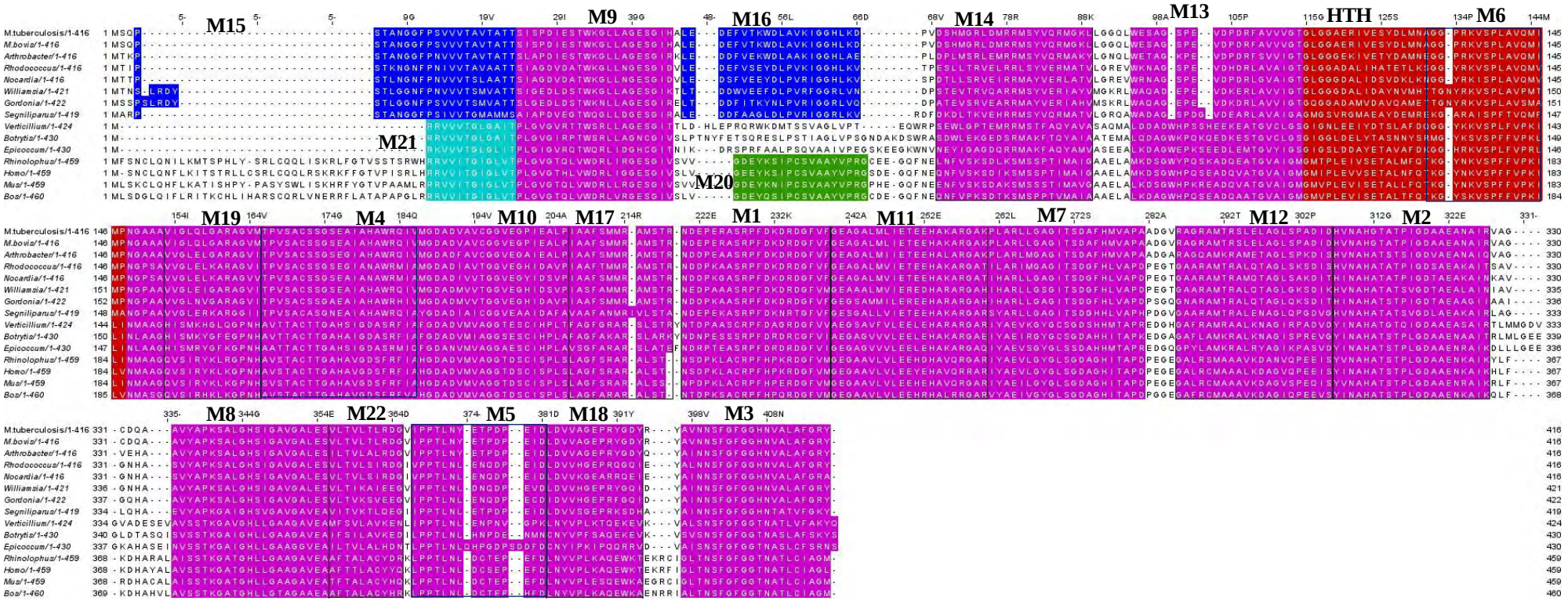


Figure S3: Mapping of discovered motifs in the KAS protein family onto the multiple sequence alignment. Motifs conserved in all sequences are colored in magenta, those unique to bacterial sequences are shown in blue, those unique to both fungi and mammals are shown in cyan and motifs only present in mammalian sequences are shown in green. Motif numbering is based on MEME results. The HTH region present in all KAS enzymes is shown in red.

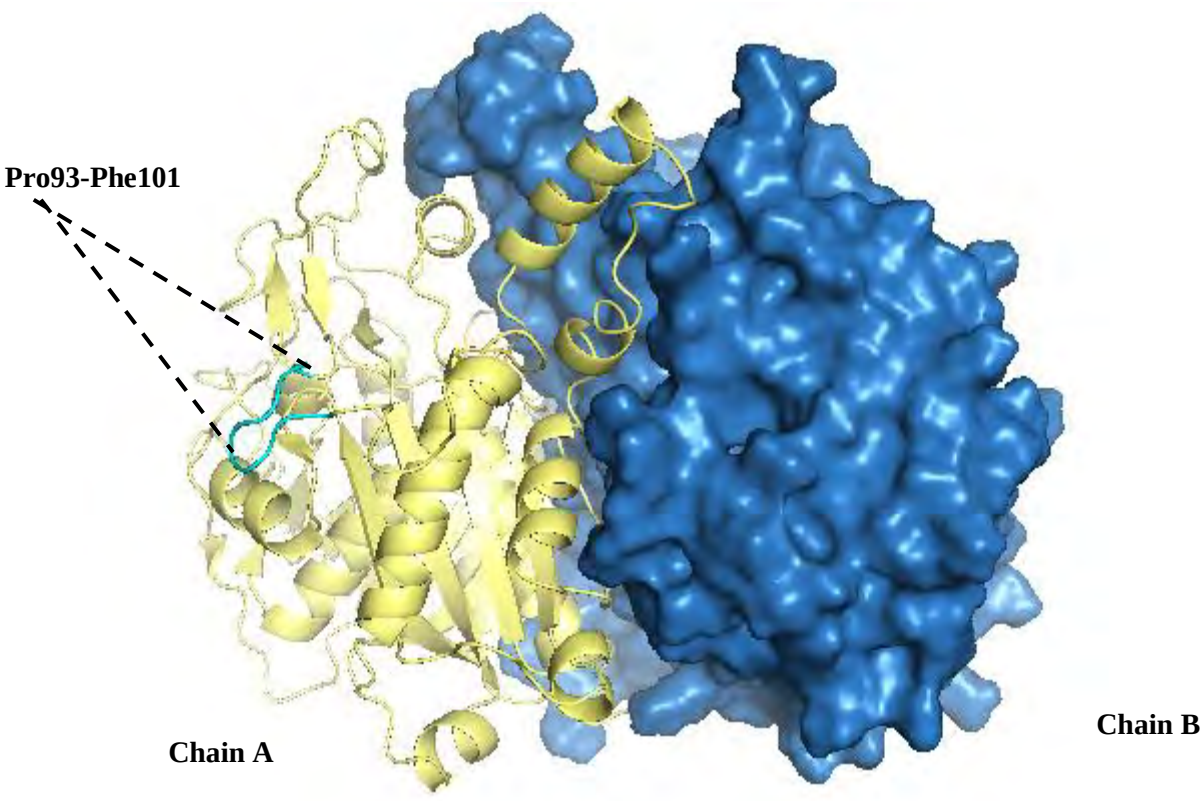


Figure S4: Crystal structure of *Hsmt* KasA, with one monomer represented as a closed surface in blue and the other monomer as a cartoon in pale yellow. The loop region made up of residues Pro93-Phe101 that make up insert II of the multiple sequence alignment is shown in cyan.

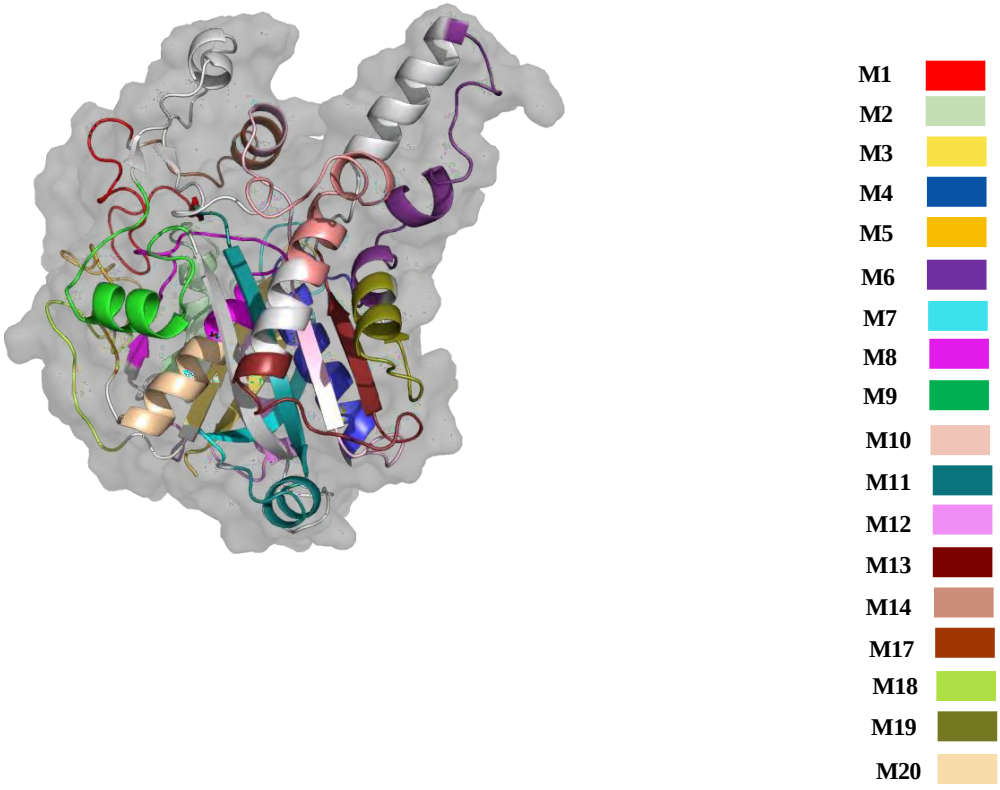


Figure S5: Mapping of highly conserved motifs on the structure of *Mtb* KasA.

Table S1: Residues constituting the allosteric and active site pockets identified by the pocket detection algorithms. Residues in red for *Mtb* KasA represent the acyl-binding tunnel residues whilst those in red for *Hsmt* KasA represent the active site residues.

Protein	Pocket	Residues
Mtb KasA	1	**Chain A** : 64,82, 115,116,117,119,120,122,123,126,170,199,200,201,202,203,205,206,209,210,239,240,241,345,346,347,404 **Chain B**: 141,142,143,145,146
	2	**Chain A**: 141,142,143,145,146 **Chain B**: 64,82, 115,116,117,119,120,122,123,126,170,199,200,201,202,203,205,206,209,210,239,240,241,345,346,347,404
	3	**Chain A**: 171,213,215,273,278,279,280,281,287,311,315,317,318,321,322,325,402,403,406
	4	**Chain B**: 171,213,215,273,278,279,280,281,287,311,315,317,318,321,322,325,402,403,406
Hsmt KasA	1	**Chain A** : 39,40,41,218,221,222,224,225,226,227, 305 **Chain B**: 39,40,41,218,221,222,224,225,226,227, 305
	2	**Chain A**: 209,251,252,253,274,308,310,313,314,315,316,317,318,321,323,324,349,348,350,352,353,354,355,356,3598,359,385,445,446,447,448,449, 451
	3	**Chain B**: 209,251,252,253,274,308,310,313,314,315,316,317,318,321,323,324,349,348,350,352,353,354,355,356,358,359,385,445,446,447,448,449, 451

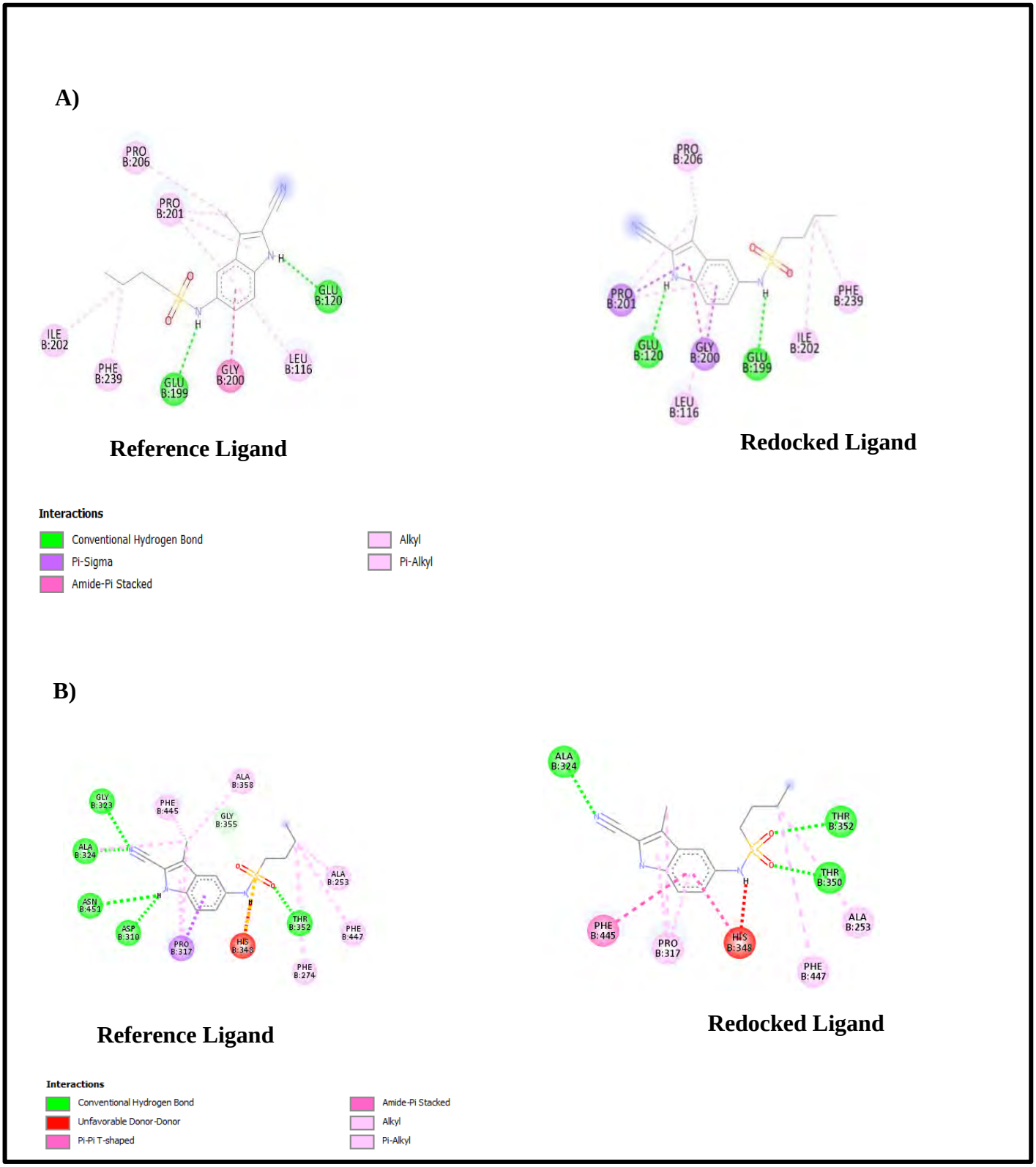


Figure S6: Protein-ligand interactions of the original co-crystallized ligand (O6G) and the redocked ligand as visualized in BIOVIA Discovery Studio. A) Interactions formed between *Mtb* KasA with the reference and redocked ligand. B) Interactions formed between *Hsmt* KasA with the reference and redocked ligand.

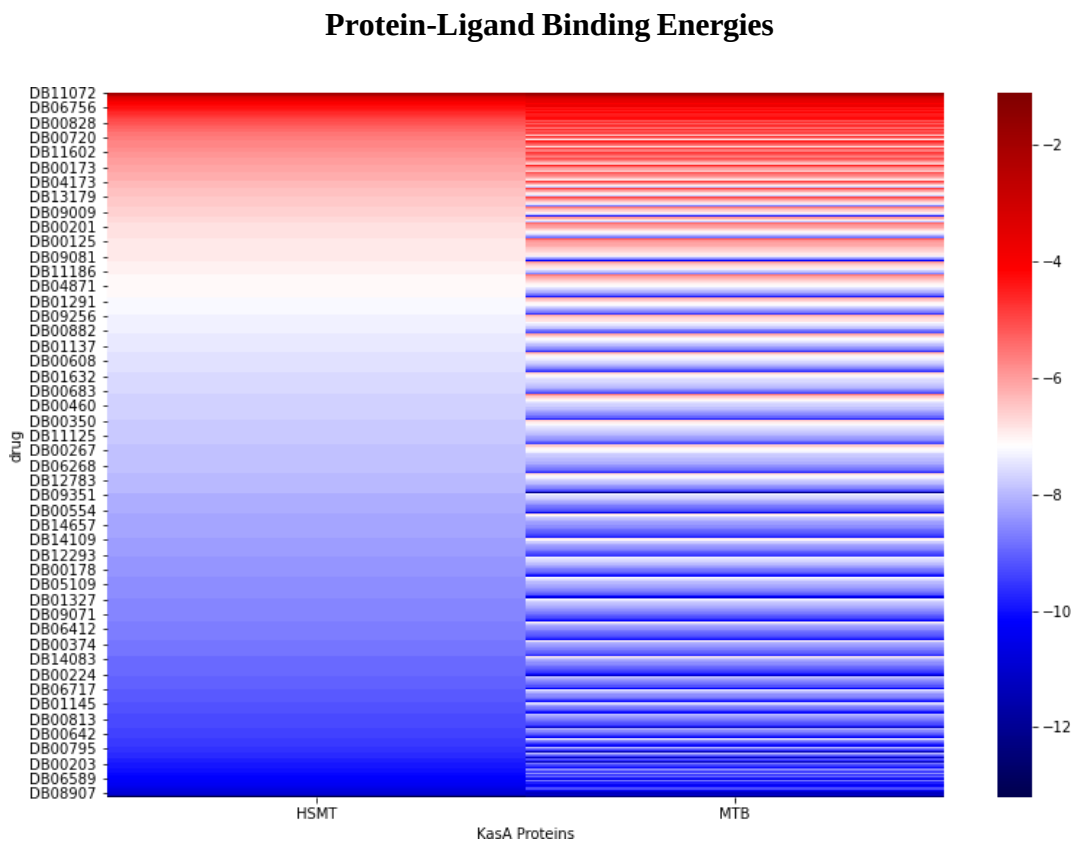


Figure S7: A heatmap showing the binding energies of the DrugBank compounds to Mtb KasA and Hsmt KasA. The heatmap was centered on the binding energy of the control ligand O6G (-7.7kcal/mol) and this value was used as a cut-off value for screening purposes.