TARGETING ALLOSTERIC SITES OF *Escherichia coli* HEAT SHOCK PROTEIN 70 FOR ANTIBIOTIC DEVELOPMENT

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

of

Master of Science in Bioinformatics and Computational Molecular Biology

(Coursework and Thesis)

of

RHODES UNIVERSITY, SOUTH AFRICA

Research Unit in Bioinformatics (RUBi)

DEPARTMENT OF BIOCHEMISTRY AND MICROBIOLOGY

Faculty of Science

By

CHIAMAKA JESSICA OKEKE

18O9027

RUBi
Research Unit in Bioinformatics

JANUARY 2019

ABSTRACT

Hsp70s are members of the heat shock proteins family with a molecular weight of 70-kDa and are the most abundant group in bacterial and eukaryotic systems, hence the most extensively studied ones. These proteins are molecular chaperones that play a significant role in protein homeostasis by facilitating appropriate folding of proteins, preventing proteins from aggregating and misfolding. They are also involved in translocation of proteins into subcellular compartments and protection of cells against stress. Stress caused by environmental or biological factors affects the functionality of the cell. In response to these stressful conditions, up-regulation of Hsp70s ensures that the cells are protected by balancing out unfolded proteins giving them ample time to repair denatured proteins. Hsp70s is connected to numerous illnesses such as autoimmune and neurodegenerative diseases, bacterial infection, cancer, malaria, and obesity. The multi-functional nature of Hsp70s predisposes them as promising therapeutic targets. Hsp70s play vital roles in various cell developments, and survival pathways, therefore targeting this protein will provide a new avenue towards the discovery of active therapeutic agents for the treatment of a wide range of diseases. Allosteric sites of these proteins in its multi-conformational states have not been explored for inhibitory properties hence the aim of this study. This study aims at identifying allosteric sites that inhibit the ATPase and substrate binding activities using computational approaches. Using *E. coli* as a model organism, molecular docking for high throughput virtual screening was carried out using 623 compounds from the South African Natural Compounds Database (SANCDB; https://sancdb.rubi.ru.ac.za/) against identified allosteric sites. Ligands with the highest binding affinity (good binders) interacting with critical allosteric residues that are druggable were identified. Molecular dynamics (MD) simulation was also performed on the identified hits to assess for protein-inhibitor complex stability. Finally, principal component analysis (PCA) was performed to understand the structural dynamics of the ligand-free and ligand-bound structures during MD simulation.

DECLARATION

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

………………………………   …………………………..

        Signature   Date

DEDICATION

This project is dedicated to Almighty God for his grace, strength, and favour granted to me during the course of this project. It is also dedicated to my late father Mr. I.G. Okeke, may his soul rest in peace, my wonderful mother Mrs. D.A. Okeke and my siblings for their love and support.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

x

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviations

Å	Ångström
ADP	Adenosine Di-Phosphate
ATP	Adenosine Tri-Phosphate
BE	Binding energy
Cα	Alpha Carbon
E. coli	*Escherichia coli*
FFT	Fast Fourier Transform
GIT	Gastrointestinal tract
GROMACS	GROningen MAchine for Chemical Simulations
HPC	High Performance Computing
Hsp	Heat shock protein
Hsps	Heat shock proteins
Hsp70	70kDa Heat shock protein
KDa	Kilo-Daltons
MD	Molecular dynamics
MM-PBSA	Molecular Mechanics energy combined with the Poisson-Boltzmann and Surface Area continuum solvation
NBD	Nucleotide binding domain
NMA	Normal mode analysis
PBC	Periodic Boundary Conditions
PDB	Protein Data Bank

PRS	Perturbation-Response Scanning
Rg	Radius of gyration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SANCDB	South African Natural Compound Database
SASA	Solvent-accessible surface area
SBD	Substrate binding domain
SBDα	Alpha-helical lid
SBDβ	Beta substrate binding domain
SVM	Support vector machine
UspA	Universal stress protein
VMD	Visual Molecular Dynamics
3D	Three-dimensional

AMINO ACID TABLE

Full Name	3-letter code	1-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

TABLE OF SOFTWARE USED

Software	Link
AlloSite	http://mdl.shsmu.edu.cn/AST/
FTMAP	http://ftmap.bu.edu/login.php
PDB	https://www.rcsb.org/
R studio	https://www.rstudio.com/
SiteMap	https://www.schrodinger.com/sitemap

CHAPTER ONE: INTRODUCTION

1.1 Background of the study

Microorganisms are part of the normal flora of the gastrointestinal tract (GIT) of human and animals (Canny and McCormick, 2008), and over time evolved to withstand its acidic nature by secreting enzymes that neutralize the pH of the GIT (Hao and Lee, 2004). Among the microorganisms found in the GIT, the family Enterobacteriaceae are the most predominant (Leão-Vasconcelos *et al.,* 2015). Enterobacteriaceae are a group of gram-negative bacteria that are facultative anaerobes and do not form spores (Health Protection Agency, 2013). This group of organisms can be symbionts, commensals or pathogens (Baylis *et al.,* 2011). Of notable interest in this group is *Escherichia coli (E. coli)*. Most strains of this organism are non-pathogenic (Nataro and Kaper, 1998) while the pathogenic strains have been implicated as the most frequent causes of many common bacterial infections, including cholecystitis (J. Liu *et al.,* 2015), bacteraemia (Davies *et al.,* 2017), cholangitis (Ahmed, 2018), urinary tract infection (UTI) (Alanazi *et al.,* 2018), and traveller's` diarrhoea (Lääveri *et al.,* 2018), and other clinical infections such as neonatal meningitis and pneumonia (Chalova *et al.,* 2009). In addition to surviving in the guts of human and some animals, *E. coli* can also survive outside the human host when passed out through faeces (Lamprecht *et al.,* 2014). *E. coli* has also being used as a model organism for drug discovery due to its ability to thrive in and out of its host (Samreen and Ling, 2014). Its survival can be affected by different environmental and biological factors such as the indigenous microbial communities, temperature, oxygen content, pH (Ishii and Sadowsky, 2008) and drug pressure. In response to these factors, *E. coli* is capable of producing stress response mechanisms that facilitate its survival (Chauret, 2011).

1.2 *E. coli* as a model organism

E. coli is arguably the most understood and broadly studied free-living organism (Blount, 2015). The extensive knowledge of *E. coli* makes it a useful model organism (Lee and Lee, 2003). This is because the organism is a single-celled organism that can reproduce rapidly in simple culture media, multiplying averagely once in every 30 minutes (Taj *et al.,* 2014). *E. coli* requires simple ingredients and nutrients to grow, and most naturally occurring strains of *E. coli* are non-pathogenic (Donnenberg and Whittam, 2001). The genetics of *E. coli* are well understood and can be easily manipulated or engineered (Samreen and Ling, 2014). The

ability of *E. coli* surviving in its primary host which is the GIT of human and its secondary host which is soil and water is still unknown (Elsas *et al.,* 2010). To withstand drastic environmental changes, various proteins including UspA (Nachin *et al.,* 2005), sigma factors (Chung *et al.,* 2006) and heat shock proteins (Hsps) (Benjamin and Mcmillan, 1998) are synthesized. Among these proteins, Hsps play a critical role in the organism's survival (Nylandsted *et al.,* 2004). In this study, we investigate this family of survival proteins.

Currently, the full-length crystallized 3D structures of *E. coli* Hsp70 in both functionally essential conformations open and closed, are available on PDB (https://www.rcsb.org/). However allosteric intermediate of *E. coli* Hsp70 was retrieved from geometric clustering of Hsp70 trajectory reported by Penkler *et al.,* 2017.

1.3 Biological significance of heat shock proteins (Hsps)

Hsps are a vast group of a protein family that are generally grouped based on their capacity and molecular weight in kiloDaltons (kDa) such as Hsp40, Hsp60, Hsp70, and Hsp90 (Li and Srivastava, 2004; Penkler *et al.,* 2017). The presence of these different types of Hsps has been accounted for in every single living life form (Cakmak, 2009). For this research, we investigate the 70kDa family of Hsps. Each group is structurally different, and they work collectively in maintaining the total cellular protein homeostasis. Hsp90 and Hsp70 have been associated with signal transduction (Nollen and Morimoto, 2002). The Hsp40 group of proteins are essential activators of the Hsp70 chaperones (Young, 2010). Some members of the Hsps family are fundamentally expressed while others are induced as a response to various environmental stimuli or growth conditions (Genevaux *et al.,* 2007). Hsps activity involves cycles of polypeptide binding and release (Bukau and Horwich, 1998). The biological significance of Hsps is highlighted by their enormous abundance in the cell (Park and Seo, 2015). In eukaryotic cells, they constitute about 1-2% of total proteins in unstressed cells (Buchner, 1996) and expanding to 4-6% of stressed cell proteins (Kim and Kim, 2011). Hsps are involved in numerous regulatory pathways, and they control the activity of regulatory proteins including heat shock transcription factors and plasmid replication initiator protein (Mayer *et al.,* 2001). Cellular stress caused by either biotic or abiotic factors is an unseemly condition for a cell and as such affect its functional role. Under these stressful conditions such as heat shock, a shift in pH, oxygen deprivation, and osmotic changes expanded articulation of Hsps ensures the cell is protected by balancing out unfolded

proteins, giving them ample time to repair or re-incorporate denatured proteins (Pirkkala *et al.,* 2001)

1.4 Heat shock protein70 (Hsp70)

Hsp70 are molecular chaperones that are essential for cell survival and play an important role in protein homeostasis (Fontaine *et al.,* 2015). They facilitate appropriate proteins folding, prevent aggregation and misfolding of proteins, transport of proteins into proper subcellular compartments, control of the cell cycle and signaling (Asea *et al.,* 2002), and protection of cells from stress (Li and Srivastava, 2004). Hsps are ubiquitously expressed in most living organisms (Schlesinger, 1990). Hsp70s are also involved in distinct cellular functions which include assembly, translocation, and degradation of proteins (Chow, 2005). Hsp70 prevents the aggregation of other proteins by binding firmly to the exposed peptide arrangements which subsequently prevents these proteins from becoming non-functional (Rudiger *et al.,* 1997). Regulation of conformational changes affecting biological processes, such as signaling is also a function of Hsp70s (Genevaux *et al.,* 2007).

1.5 Hsp70s architecture and nucleotide-dependent functional cycle

Members of the Hsps family with a molecular weight of 70kDa (Hsp70) are by far the most extensively studied and abundant group in bacterial and eukaryotic systems (Benjamin and Mcmillan, 1998) and in some members of the Archaea domain (Gribaldo *et al.,* 1999). In human, there are ten functionally and structurally distinct isoforms of Hsp70 proteins (Fontaine *et al.,* 2015) that are located in cellular compartments such as cytoplasm, endoplasmic reticulum (ER), mitochondria, and nucleus (Shonhai *et al.,* 2007). In *Plasmodium*, there are six functionally and structurally homologs of Hsp70 proteins (Hatherley *et al.,* 2014) in bacteria, the major homolog of Hsp70, DnaK, is located in the cytosol (Calloni *et al.,* 2012). Hsp70 consists of two integral domains; a nucleotide binding domain (NBD) located at the N-terminus and a substrate binding domain (SBD) located at the C-terminus (Figure 1). These domains are connected via a highly conserved hydrophobic linker region which is believed to facilitate inter-domain communication (Penkler *et al.,* 2017; Hatherley *et al.,* 2014). The NBD contains ATP/ADP binding site and is made up of two lobes – I and II which form a clamp-like structure with a large cleft in the center (Powers *et al.,* 2010). Each lobe consists of two subdomains – IA, IIA, IB, and IIB. The SBD contains

the substrate binding site which consists of a hydrophobic Beta-peptide binding pocket (SBDβ) and an alpha-helical lid (SBDα) (Penkler *et al.,* 2017; Bertelson *et al.,* 2009).

Hsp70 action relies on the intra-molecular interactions between its NBD and SBD, and also between the inter-molecular interactions with its cochaperone DnaJ (Li *et al.,* 2016). The specificity of Hsps70s is greatly dependent on a co-chaperone DnaJ (Jiang *et al.,* 2007). DnaJ is usually found in large numbers and in combination with Hsps70, function as a powerful molecular chaperone machine (Mayer and Bukau, 2005). The multi-functional nature of Hsp70 requires it to associate indiscriminately with misfolded proteins and also specifically with folded substrates (Laufen *et al.,* 1999). Appropriate selection of substrates is hypothesized to be a role of DnaJ co-chaperone, which controls the ATPase cycle of Hsp70 (Davis *et al.,* 1999). Hsp70 cycle operates between two conformations induced by ATP binding: open and closed conformations (Figure 1.1). Experimental studies using *E. coli* Hsp70 as a model has shown that in its open conformation when bound to ATP (Figure 1.1A), the hydrophobic linker is a highly-structured participant in the interdomain docking of the SBD onto the NBD such that both domains are dependent on each other (English *et al.,* 2017). In contrast to its closed conformation when bound to ADP (Figure 1.1B), the hydrophobic linker is flexible such that both NBD and SBD are undocked, making them largely independent of each other (Penkler *et al.,* 2017). Figure 1.3 provides s diagrammatic representation of the series of events taking place during the Hsp70 nucleotide-dependent functional cycle. Exchange of ADP for ATP governs the refolding role of Hsp70 and their affinity for substrates. In its open conformation, when bound to ATP, Hsp70 has a low affinity and high on-off transfer rates for peptide binding (Zhuravleva *et al.,* 2012). Upon ATP binding, DnaJ facilitates the binding of a peptide substrate at the SBD balances out the NBD leading to an allosterically active intermediate state (Figure 1.2) in which the SBD partially separates from the NBD (Penkler *et al.,* 2017). Hydrolysis of ATP destabilizes the SBD resulting in the eventual transition to the closed state (Fontaine *et al.,* 2015). This brings about an unwinding of the SBD conformation, enabling substrates to withdraw and the binding of new substrates. The exchange of ATP for ADP by the nucleotide exchange factor, GrpE, initiates an allosteric signal at the NBD that is communicated to the SBD, which results in the partial docking of the SBD to the NBD and an allosterically active intermediate. Structural rearrangements at the SBD leads to the dissociation of the peptide substrate, bringing the chaperone back to the start of the cycle (Penkler *et al.,* 2017). The hydrolysis of ATP is triggered by DnaJ proteins (Kampinga and Craig, 2010). In its closed conformation

where ATP is being hydrolyzed to ADP, Hsp70 has a high affinity for peptide binding, while there is a low affinity for peptide binding in its open conformation (Shonhai *et al.,* 2007).
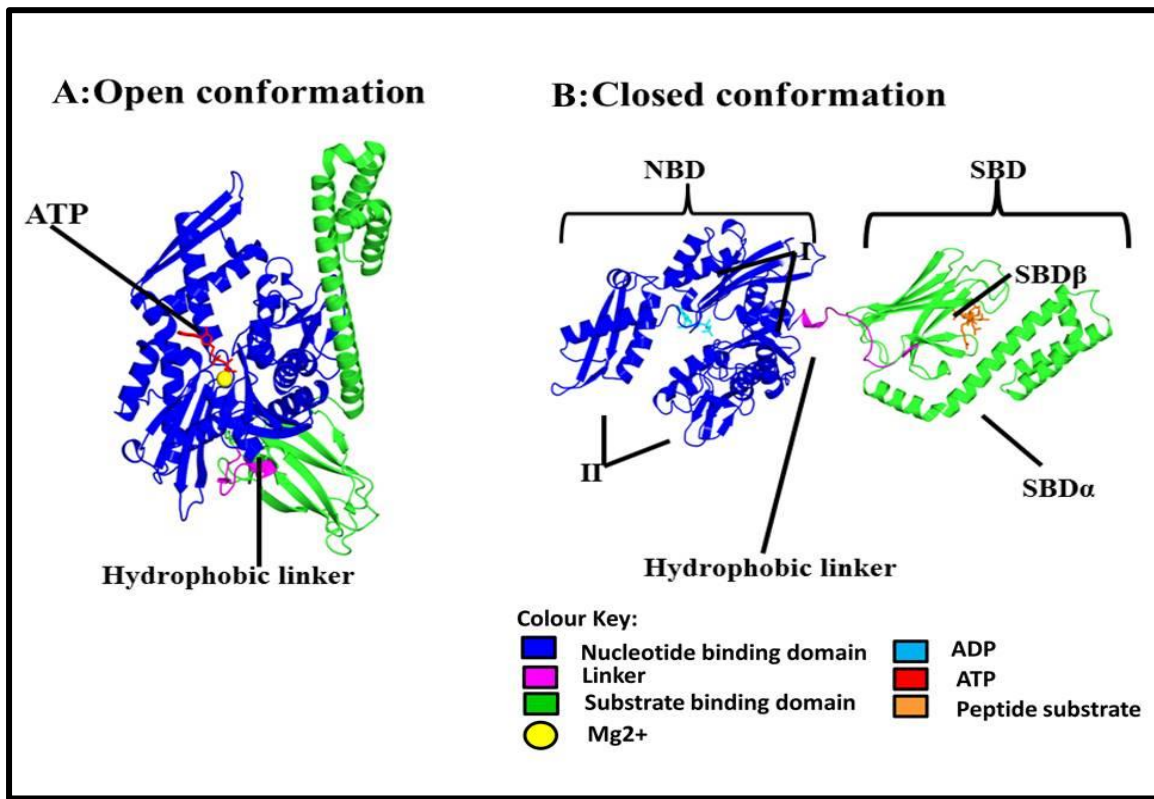
Figure 1.1: A): Experimental crystal structure of Hsp70 in its open conformation (PDB ID 4B9Q) (Kityk et al., 2012). NBD (blue) res 1-370, SBD (green) res 390-600, hydrophobic linker (pink), Mg^{2+} (yellow), ADP (Sky blue), ATP (red), and peptide substrate (Orange). **B**: Experimental crystal structure of E. coli Hsp70 in its closed conformation (PDB ID 2KHO) (Bertelsen et al., 2009).
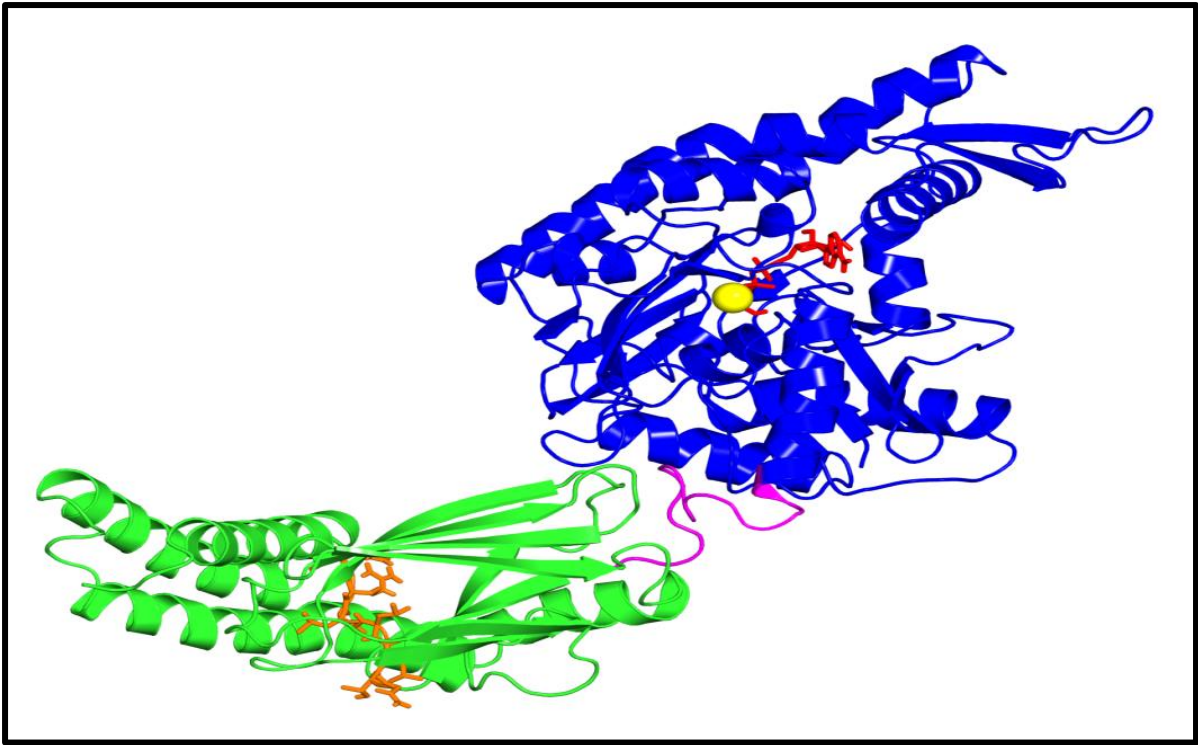
Figure 1.2: Allosteric intermediate conformation of E. coli Hsp70 (Penkler et al., 2017).NBD (blue) res 1-370, SBD (green) res 390-600, hydrophobic linker (pink), Mg^{2+} (yellow), ADP (Sky blue), ATP (red), and peptide substrate (Orange).
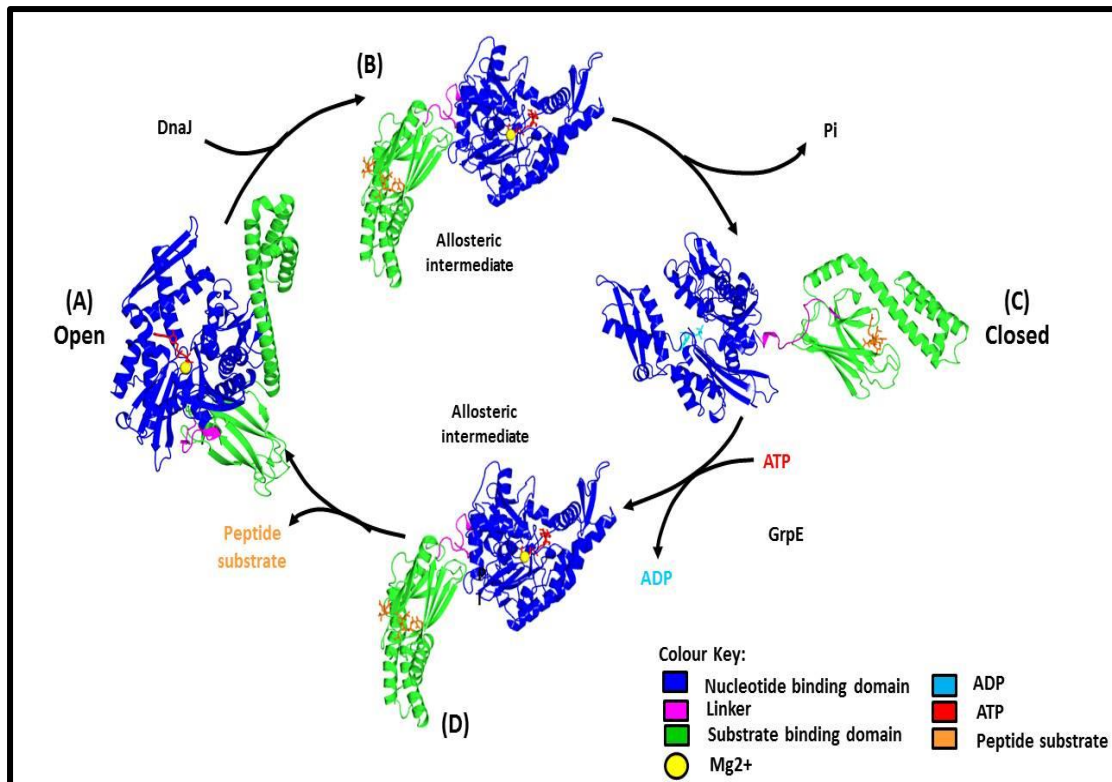
Figure 1.3: Schematic demonstration of Hsp70 nucleotide-dependent cycle. In its open conformation, where the NBD and SBD are attached to each other, DnaJ facilitates binding of a peptide substrate in the SBD triggering the hydrolysis of ATP to ADP (A-B). Exchange of ADP for ATP with the help of GrpE stimulates a conformational change into a high-affinity stable substrate complex such that the NBD detaches from the SBD making them largely independent of each other.

1.6 Hsp70s as therapeutic targets

Alterations in proteins have been implicated as the reason for a vast number of human diseases also called protein conformation disorders (Morimoto and Cuervo, 2009). Hsp70s play critical roles in proteostasis, and the presence of these proteins in the organism can enhance its pathogenicity. These proteins are notably conserved across species and are important for cell survival making them a key factor for potential drug target (Przyborski *et al.,* 2015). The SBD in Hsp70s enables its binding with a wide range of molecules which provides a cytoprotective role against different cell stress (Kumar *et al.,* 2016). Hsp70s is connected to numerous illnesses, to a great extent by hereditary examinations or articulation investigation. Recent reviews have confirmed the connection of Hsp70 to cancer (Schlecht *et al.,* 2013), neurodegenerative disease, autoimmunity, and other disorders (Li *et al.,* 2016). In malaria, *Plasmodium falciparum* genome encodes six homolog Hsp70s (PfHsp70) which play distinctive roles in the parasite's development, survival and pathogenesis (Przyborski *et al.,* 2015). The resistance of *P. falciparum* to available drugs is a significant concern, and the

quest for novel approaches in combating the disease caused by the parasite and eliminating the parasite is a continuous effort (Sinha *et al.,* 2014). Targeting PfHsp70 whose presence suggests a possible role in the pathogenesis of the parasite and its life cycle represent novel therapeutic targets (Bell *et al.,* 2011). Hsp70 has been linked to being a major cause of tumor cell resistance against different cells. Increased expression of this protein in cancer cells is associated with tumorigenesis and for tumor movement by conferring protection from chemotherapy. Targeting Hsp70 from being expressed diminishes the size of tumors and can cause their total relapse (Kumar *et al.,* 2016). Numerous studies have implicated the presence of Hsp70 in *Helicobacter pylori* to the occurrence of gastric cancer. Hsp70 aids attachment of *H. pylori* to the gastric epithelia enhancing its pathogenicity in the gastric system of humans (Hoffman and Garduno, 1999).

1.6.1 E. coli Hsp70 as a therapeutic agent

E. coli Hsp70 also known as DnaK interact with a host of different substrates or clients by binding to unprotected hydrophobic regions of partially folded or unfolded proteins. ATP binding and hydrolysis induces structural rearrangements in the NBD which perpetuates interdomain docking between the NBD and SBD. Allosteric communication between the domains is vital for protein function and the basic characteristics of this allostery have long been investigated using *E. coli* DnaK as a model (Leu *et al.,* 2014). DnaK is important in maintaining proteostasis and has been characterized in certain pathogenic bacteria as an important factor in the pathogenicity and stress resistance in multi-drug-resistance bacteria (Chiappori *et al.,* 2015). DnaK has proven to be extremely attractive targets for developing new treatments for several diseases such as microbial infections, neurodegenerative diseases, and other protein folding disorders (Julia I Ju Leu *et al.,* 2014). Despite the incredible interest in the preclinical potential of small molecules targeting this protein, it has proved difficult to identify and characterize effective modulators for therapeutic use due to the scarcity of drug-like inhibitors as a therapeutic target (Miyata *et al.,* 2010). As a result, few selective inhibitors have been identified which include 2-phenylethynesulfonamide (PES) (Leu *et al.,* 2009), chlorinated derivative (PES-Cl) (Balaburski *et al.,* 2013), triphenyl(phenylethynyl)phosphonium bromide (PET-16) (Leu *et al.,* 2014) and the ATP analog VER-155008 (Schlecht *et al.,* 2013).

Targeting Hsp70s has previously involved inhibiting the ATPase activity of the protein with natural compounds that acts as ATP-competitive inhibitors and thus prevents allosteric communication between NBD and SBD (Assimon *et al.,* 2013; Schlecht *et al.,* 2013). This process also includes targeting its interactions with important co-chaperone, DnaJ (Patury *et al.,* 2009). Another approach includes preventing the SBD from binding to its substrates. Pyrrhocoricin, a member of the Proline-rich antibacterial peptide family and the most potent peptide, binds to the SBD of *E. coli* DnaK thereby inhibiting its ATPase activity (Kragol *et al.,* 2002). Recent studies have demonstrated that proteins and enzymes, alongside their remarkably studied functional sites and allosteric sites, contain a lot of practically unexplored potential target pockets (Suplatov and Švedas, 2015). Though the molecular mechanisms of allosteric regulation in Hsp70 have been studied extensively at each functional and structural levels, the current understanding of allosteric inhibition of its activities by small molecules is indeed missing (Stetz and Verkhivker, 2016).

In this study, we propose to explore the potentials of allosteric sites. We seek to discover druggable allosteric sites and subsequently screen these sites against a database of small molecules in a bid to find potential allosteric modulators.

1.7 Project Motivation

The multi-functional nature of Hsp70s predisposes them as promising therapeutic targets. Hsp70s play vital roles in various cell developments, and survival pathways, therefore targeting this protein will provide a new avenue towards the discovery of effective therapeutic agents for the treatment of a wide range of diseases. Allosteric sites of the full-length 3D structures of this protein in its multi-conformation states have not been explored for antibiotic development. We proposed to target the potential allosteric sites of Hsp70 for antibiotic development using *E. coli* as a model organism. We also view the effect of selected inhibitory compounds on the protein.

1.7.1 Knowledge gap

To our knowledge, this research is the first study that is focused on targeting allosteric binding sites of *E. coli* Hsp70 in functionally important conformations, including a unique allosteric intermediate state, against natural compounds from the South African Natural Compounds Database (SANCDB) (https://sancdb.rubi.ru.ac.za/).

1.7.2 Research Hypothesis

This research hypothesizes that targeting allosteric sites of *E. coli* Hsp70 and screening against compounds from the South African Natural Compounds Database (SANCDB) can be explored for inhibitory drug design purposes.

1.8 Aims and Objectives

1.8.1 Aims

This study aims at identifying allosteric binding sites of *E. coli* Hsp70 in its open, closed and intermediate conformations, as well as exploring the inhibitory potentials of compounds from SANCDB (https://sancdb.rubi.ru.ac.za/) against these sites.

1.8.2 Objectives

1   Identification of potential allosteric sites of *E. coli* Hsp70 protein.

2   Molecular docking of natural South African compounds from SANCDB (https://sancdb.rubi.ru.ac.za/) to identified allosteric sites.

3   Analysis of stability and conformational changes of protein-ligand complexes.

CHAPTER TWO: ALLOSTERIC SITES IDENTIFICATION

2.1 Chapter Overview

Allostery is a key process to regulate the functional activity of a protein. Targeting allosteric sites rather than orthosteric sites in reference to drug designs can offer certain advantages such as target specificity, selectivity and low toxicity (Huang *et al.,* 2017). Penkler *et al.,* 2017 identified allosteric hot residues on the open and closed conformation of *E. coli* Hsp70 structure using perturbation response scanning (PRS) method in combination with MD simulation. The structures used in this study were structures from Penkler *et al.,* 2017 containing PRS data. These structures were extracted through geometric clustering of the protein's RMSD. In this chapter, potential allosteric pocket prediction tools such as FTMAP (http://ftmap.bu.edu/login.php), SiteMap (https://www.schrodinger.com/sitemap), and Allosite (http://mdl.shsmu.edu.cn/AST) were used to predict allosteric sites that correlate to the results of Penkler *et al.,* 2017. These tools utilize different techniques such as pocket-based analysis, support vector machine (SVM), machine learning and energy based method to accurately predict ligand binding pockets other than those of known active site (Greener and Sternberg, 2018). Allosteric sites on the structures extracted from Penkler *et al.,* 2017 are explored for inhibitory design and to discover new functions. Each structure without any endogenous ligand is uploaded in these servers, and ligands binding within 10 Å of the identified sites are noted.

2.1.1 Allosteric sites

The word allostery, coined from the Greek words: allos which mean "other" and stereos which mean "shape" (Christopoulos, 2002), describes the functional alteration at one site on a protein resulting from the change at another site distinct from each other (Greener and Sternberg, 2018). Figure 2.1 describes the definition of allostery. Alteration in the conformation of the protein structure can either be positive (allosteric effectors) thereby enhancing its catalytic activity or negative (allosteric inhibitors), reducing its catalytic activity (Helmstaedt *et al.,* 2001). Allosteric communication between binding sites is necessary to transmit information (Schulze et al., 2016). The idea of allostery has evolved within the past century with improved experimental and computational technologies (Jin Liu and Nussinov, 2016). Experimental methods used in understanding allostery include high-

resolution X-ray crystallography (Motlagh *et al.,* 2014), nuclear magnetic resonance spectroscopy (NMR), high-throughput screening (HTS) (Sugiki *et al.,* 2018), fragment-based screening (FBS) and electrophysiology (Nys *et al.,* 2016). Computational methods such as normal mode analysis (NMA), machine learning methods, molecular dynamics (MD) methods, and evolutionary methods, have emerged to help in the understanding of allostery (Wagner *et al.,* 2016). The initial models of allostery were first described by Monod-Wyman-Changeux (MWC) and Koshland-Nemethy-Filmer (KNF) (Thayer *et al.,* 2017). According to the MWC model, also known as the concerted model, allosteric proteins are oligomers consisting of the identical monomers arranged in a symmetrical order (Cui and Karplus, 2008) and they exist in two conformational states, i.e. relaxed state (R) i.e. high-affinity state and tensed state (T) i.e. low-affinity state (Hilser *et al.,* 2012) . In this model, each monomer (subunit) must exist in the same state, either the R or the T state. These states are said to be in equilibrium in the presence or absence of the ligand (Rapp and Yifrach, 2017). Binding of a ligand to a site other than the active site shifts the equilibrium to either of the two states (Bellelli and Brunori, 2011). Koshland-Nemethy-Filmer (KNF) model, also known as the sequential model, states that the subunits do not exist in the same state. The equilibrium can be shifted from one state to another through the binding of a ligand to one of the subunits (Freiburger *et al.,* 2011). Other models of allostery such as Population Shift Model, Morpheein Model, and the Dynamically Driven Model extended the MWC and KNF models (Lu *et al.,* 2014).
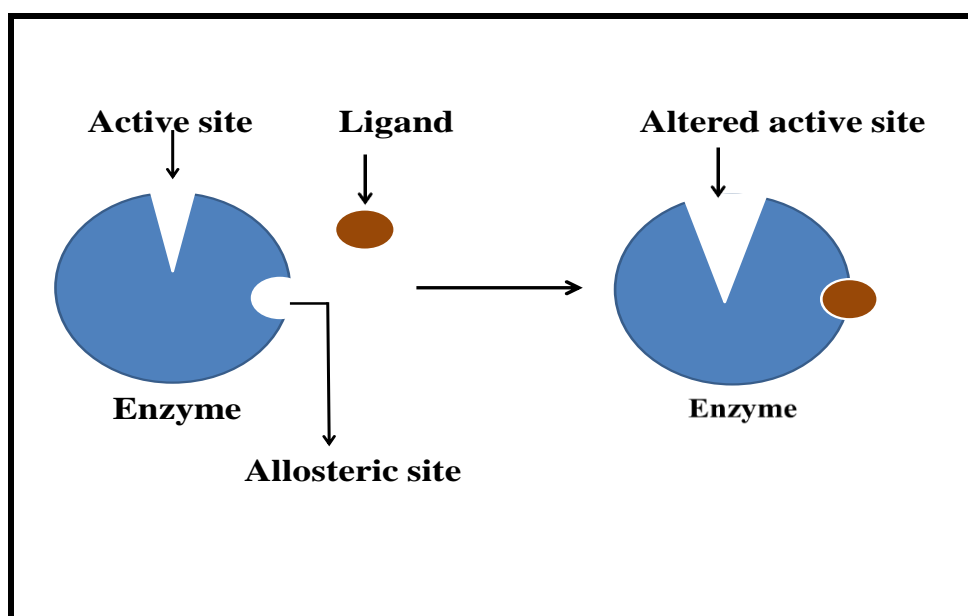
Figure 2.1: A schematic representation of allosteric inhibition (adapted from Thermodynamics and Reactions, n.d.)

2.1.2. Characteristics of allosteric proteins

Allosteric proteins are involved in biological processes control and regulation which are essential to the biological system. Distinctive features of allosteric proteins are that they comprise of different polypeptide chains with various active and allosteric sites (Einav *et al.,* 2016), and these proteins can react to a few distinct conditions in their surroundings. Allosteric proteins are larger and more complex than orthosteric proteins (Christopoulos *et al.,* 2004) and can either be activated or inhibited by allosteric ligand(s). They often control key reactions in major pathways that need to be regulated. In allosteric sites, greater stability and hydrophobicity is observed than in orthosteric sites since hydrophobic residues are highly enriched in these sites (Xiaobai Li *et al.,* 2013). Residues of allosteric sites are considerably less conserved than those of orthosteric sites which aid the specificity and selectivity of allosteric drugs (Shang *et al.,* 2016).

Despite the many potential benefits of allosteric therapy, identification of predictive approaches in the discovery of allosteric sites is one of the major hurdles in the development of allosteric drugs (Tee *et al.,* 2018).

2.1.3 Advantages of allosteric drugs

Allosteric drugs offer certain advantages over orthosteric drugs that target protein's functional site. Allosteric sites are evolutionarily less conserved and this confers higher selectivity and specificity on allosteric drugs. Allosteric drugs bind to non-conserved sites hence they are less toxic and have fewer side effects (Grover, 2013). Allosteric drugs do not compete with biological co-factors such as ATP and ADP that are present in the protein and can function in the presence of these ligands (Nussinov and Tsai, 2015).

2.1.4. Methods to identify and characterize allosteric sites

Identification and analysis of binding hot spots that contribute significantly to the binding free energy of ligands are critical steps in structural drug design (Landon *et al.,* 2009). Current known allosteric drugs have been identified by experimental methods such as X-ray crystallography and high-throughput screening (Blundell and Patel, 2004). While these experimental methods have been successfully used, the cost, time constraints, and poor performances of these experiments limits their utilization in routine investigations of protein targets (Ekins *et al.,* 2007). In this study, we use different allosteric pocket prediction tools that utilize different algorithms such as machine learning to identify druggable cavities.

These cavities are then compared with the results from previously identified PRS hotspots. We propose that these identified consensus regions are most likely allosteric sites.

2.1.4.1. FTMAP

FTMap is a computational tool that was created as an equivalent of the experimental tools used in predicting binding hot spots on protein structures (Kozakov *et al.,* 2015). This server predicts binding hot spots on protein structures that bind small organic molecules or probes by sampling the surface of the protein structure and finding residues with hydrophobic and polar pattern (Brenke *et al.,* 2009). FTMAP algorithm scores the identified hotspots based on their average energy using the Fast Fourier Transform (FFT) (Mirzaei *et al.,* 2015). FTMAP identifies these hot spots by screening the protein of interest against a set of 16 small organic molecules as probes (acetaldehyde, acetamide, acetone, acetonitrile, benzaldehyde, benzene, cyclohexane, dimethyl ether, ethane, ethanol, isobutanol, isopropanol, methylamine, N,N-dimethylformamide phenol, and urea) of different sizes, shapes and polarity (Ngan *et al.,* 2012). The regions in which several different probe clusters bind are identified as the hot spots and region with the largest number of these clusters is selected as the main spot (Beglov *et al.,* 2018).

2.1.4.2. SiteMap

Schrodinger's SiteMap is a computational tool for the identification and characterization of sites that contribute to the binding of small molecules in respect to drug discovery (T. A. Halgren, 2009). This tool provides a quick and efficient way of predicting potential binding sites with a high degree of certainty (T. Halgren, 2007). SiteMap uses an algorithm similar to the GRID algorithm of Goodford that uses interaction energies between the protein and grid probes to locate energy-efficient sites (Patschull *et al.,* 2012). SiteMap defines sites by a combination of physicochemical properties measured at each site such as the degree of protein enclosure, the degree of solvent exposure, the degree to which a ligand is able to donate or accept hydrogen bonds, druggability score (D-score), hydrogen-bonding possibilities, hydrophilic and hydrophobic properties, Site score (S-score), size, and volume (Gudipati *et al.,* 2018; Rodina *et al.,* 2013). SiteMap identifies sites by linking site points that contribute significantly to the interaction between either protein-ligand or protein-protein and ranks these sites based on their physicochemical properties (T. Halgren, 2007). The top 5 ranked potential ligand-binding sites are displayed.

2.1.4.3. Allosite

Allosite is a web server that offers an effective way in the identification of allosteric sites and can be a useful tool for the discovery of allosteric drugs (Song *et al.,* 2017). Allosite uses algorithms like pocked-based analysis and support vector machine (SVM) classifiers to predict the location of allosteric sites on protein (Huang *et al.,* 2013). Allosite provides physicochemical properties of identified sites such as druggability score, feature score, perturbation score, total solvent-accessible surface area (SASA), and volume. The druggability score represents protein cavity druggability based on a knowledge-based logistic model which is trained with the largest number of druggable and non-druggable cavities. It ranks the identified sites as druggable, difficult to drug and undruggable (Schmidtke and Barril, 2010). The feature score is derived based on the physicochemical and topological feature of the predicted allosteric site. The perturbation score evaluates the allosteric effect of pseudo ligands using normal mode analysis.

2.1.5. Structures used in this study

Structures used in this study were provided by Penkler *et al.,* 2017. Previous study by Penkler *et al.,* 2017 identified allosteric hot residues on the *E coli* Hsp70 in its open and closed conformations using perturbation response scanning (PRS) in combination with molecular dynamics (MD) simulation. A total of 12 unique configurations of *E coli* Hsp70 were prepared using crystal structures 4B9Q (open conformation) and 2KHO (closed conformation) from the Protein Data Bank (PDB) as starting structures for the experiment. Conformational changes between the various nucleotide and substrate-bound configurations of the protein were analysed by applying PRS. PRS uses the Linear Response Theory (LRT), in which a particular protein conformation can be defined by the Hamiltonian disturbance of an alternative conformation and the shift in coordinates as a result of an external force including ligand binding. LRT determines the allosteric effect each residue has on all other residues in a given protein upon external perturbation (Gerek & Ozkan, 2011). Each configuration was subjected to all-atom MD simulation for a minimum of 100ns and a maximum of 200ns in order to obtain a 20ns equilibrated trajectory wherein the RMSD fluctuation of the backbone converged at around 2Å. Finally, prediction of intra-protein contact residues was determined using the protein interactions calculator (Penkler *et al.,* 2017).

Of the 12 configurations generated, two structures (H_prime and B_prime) (Figure 2.2) used in this study were extracted by geometric clustering of protein RMSD's on trajectory frames spanning the 20ns equilibrated time periods used in PRS calculations (Table 2.1).

Table 2.1: MD Trajectories indicating the simulation length and the backbone RMSD fluctuation used to obtain the respective 20ns equilibrated region used in PRS calculations.

Configuration	Simulation length (ns)	Equilibrated regions (ns)	RMSD fluctuation (Å)
H_prime	200	180-200	~ 2.0
B_prime	200	116-136	~ 2.3

The objective of this study is to identify compounds interacting with these residues and to analyze the inhibitory effects of these compounds.
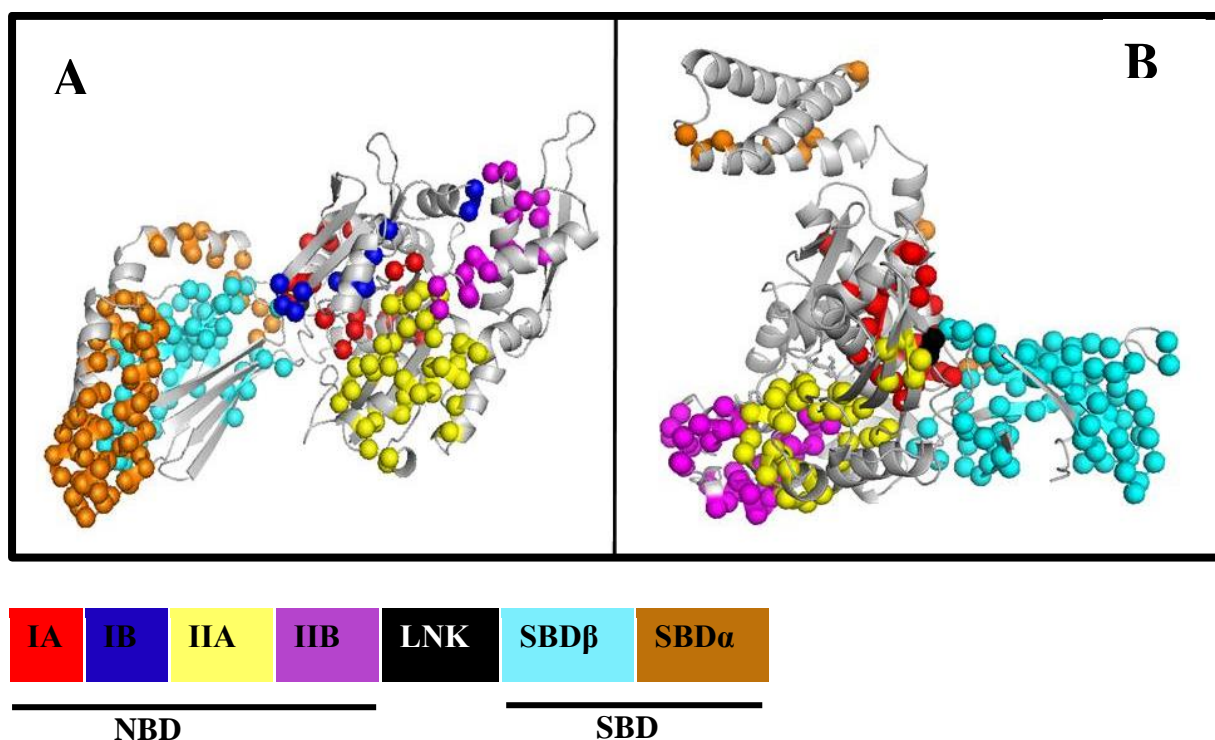
Figure 2.2: Structures retrieved from Penkler et al., 2017. Identified allosteric hot residues mapped to Cα atom and coloured by subdomain. **A:** Structure in its closed conformation (H_prime). **B:** Structure in its open conformation (B_prime).

2.2. Methodology

Identification of binding hotspots of small molecular fragments was done using computational tools like FTMAP, SiteMap, and Allosite. *E. coli* Hsp70 in its open (B_prime) and closed conformation (H_prime) was retrieved from Penkler *et al.,* 2017. Both structures were prepared using Discovery studio. Endogenous ligands, metal ions, and water molecules were removed from the structures.

2.2.1. FTMAP

Most crystallized structures usually contain co-crystallized ligands, metal ions, co-factors, and water molecules. Protein structures from Penkler *et al.,* 2017 were prepared using Discovery studio 2017. The protein preparation includes the removal of endogenous ligands, metal ions, water molecules, and other co-crystallized agents. The prepared structures are then uploaded to the FTMAP web server. FTMAP processes the PDB file by inspecting the structures to ensure that all bound ligands and water molecules are removed and adding any missing atoms, including polar hydrogens. The pre-docking minimization involves screening the entire surface of the protein structure using 16 small organic molecules as probes (acetaldehyde, acetamide, acetone, acetonitrile, benzaldehyde, benzene, cyclohexane, dimethyl ether, ethane, ethanol, isobutanol, isopropanol, methylamine, N,N-dimethylformamide phenol, and urea) (Ngan *et al.,* 2012). FTMAP utilizes the FFT method to sample these compounds while calculating accurate energies. After each probe is docked, the best 2000 poses are kept for each probe for further processing. The free energy of 2000 poses produced after the rigid body docking is minimized using the CHARM potential. The minimized probes are clustered according to proximity. Based on their Boltzmann averaged energies, the clusters are ranked and six clusters with the lowest average free energies are retained for each probe. Binding hotspots are determined by searching for clusters of different types of probes within several angstroms of each other. If several binding hot spots on the protein surface appear near each other, this is a strong indication of a potentially druggable binding site. The FTMAP algorithm steps used in generating the binding sites are outlined below (Figure 2.3).
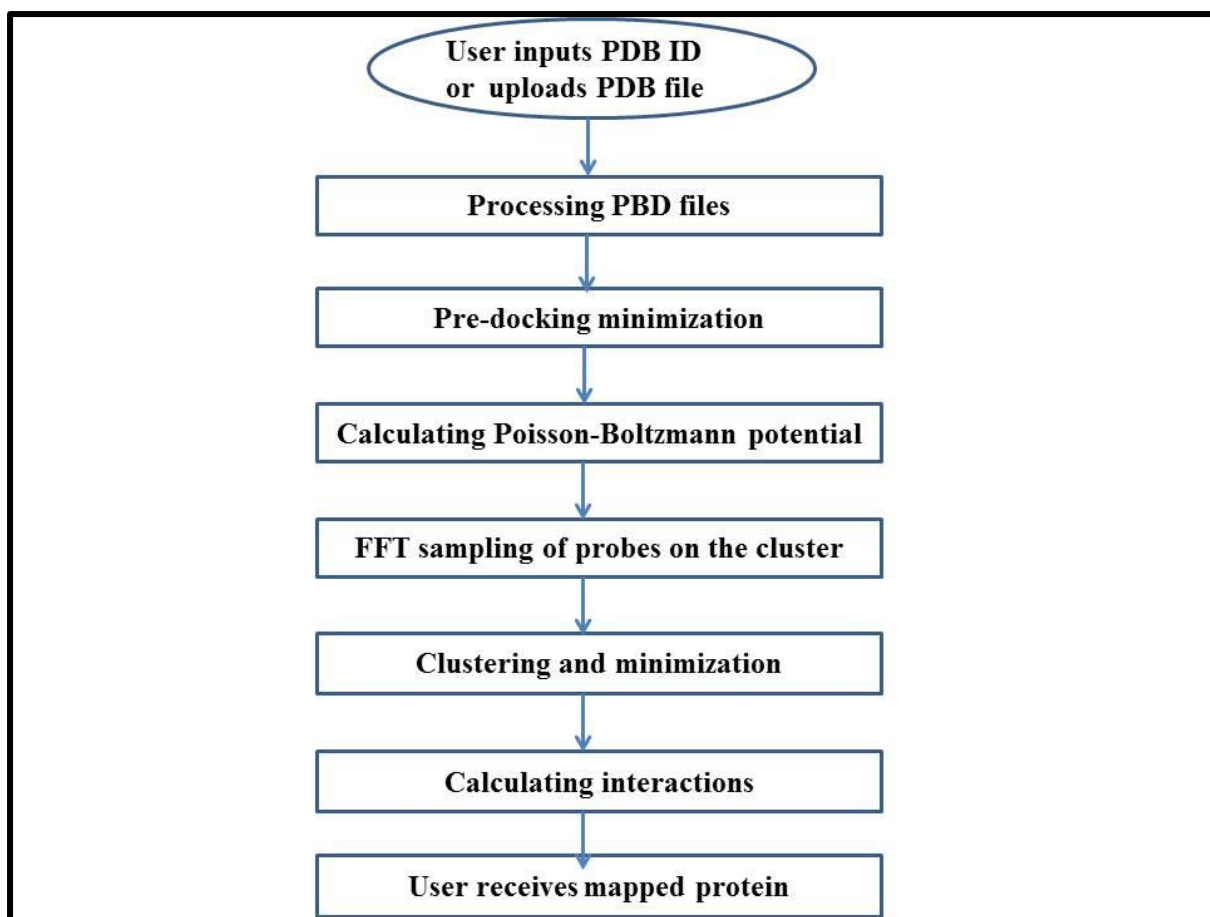
Figure 2.3: A diagrammatic flow of the steps involved in FTMap.

2.2.2. SiteMap

In order to ensure correct starting structures, the protein structures from Penkler *et al.,* 2017 was subjected to protein preparation. Each protein structure was prepared using the protein preparation wizard protocol available in the Schrödinger suite (Maestro version 11.5.0.11 MMshare version 4.1.011 Release 2018-1 Platform Linux-x86-64). Protein preparation wizard involves the pre-processing, optimizing and minimization of the protein structure. Pre-processing of the structure assigns bond orders, adds hydrogen atoms, fills in missing loops or side chains, creates zero-order bonds to metals, creates disulfide bonds, and deletes water molecules. Optimization involves the hydrogen-bond (H-bond) assignment. This is achieved by re-orientating the hydroxyl and thiol groups, water molecules, amine groups of Asparagine and Glutamine and the imidazole group in Histidine. This also predicts the protonation states of Histidine, Aspartic acid and Glutamic acid and the tautomeric state of Histidine (Schrödinger Suite 2012). This step eliminates atomic clashes, adds formal charges to the hetero groups and optimizes them at neutral pH using PROPKA (Olsson *et al.,*2011). The final step in the protein preparation step is the minimization of the structure. This is done

to restrain the heavy atoms in the structure so that the final result does not deviate so much from the starting structure. RMSD cut-off of 3.0 Å is specified and the OPLS3 force field is used (Schrödinger Suite 2012). Once the structures have been prepared, the binding site analysis was carried out. This was generated using SiteMap tool from Schrödinger suite. SiteMap finds, visualizes and evaluates protein binding sites (Schrödinger, 2016). SiteMap identified the top 5 ranked potential ligand-binding sites on structures from Penkler *et al.,* 2017. Detailed steps used in generating binding sites are outlined below (Figure 2.4).
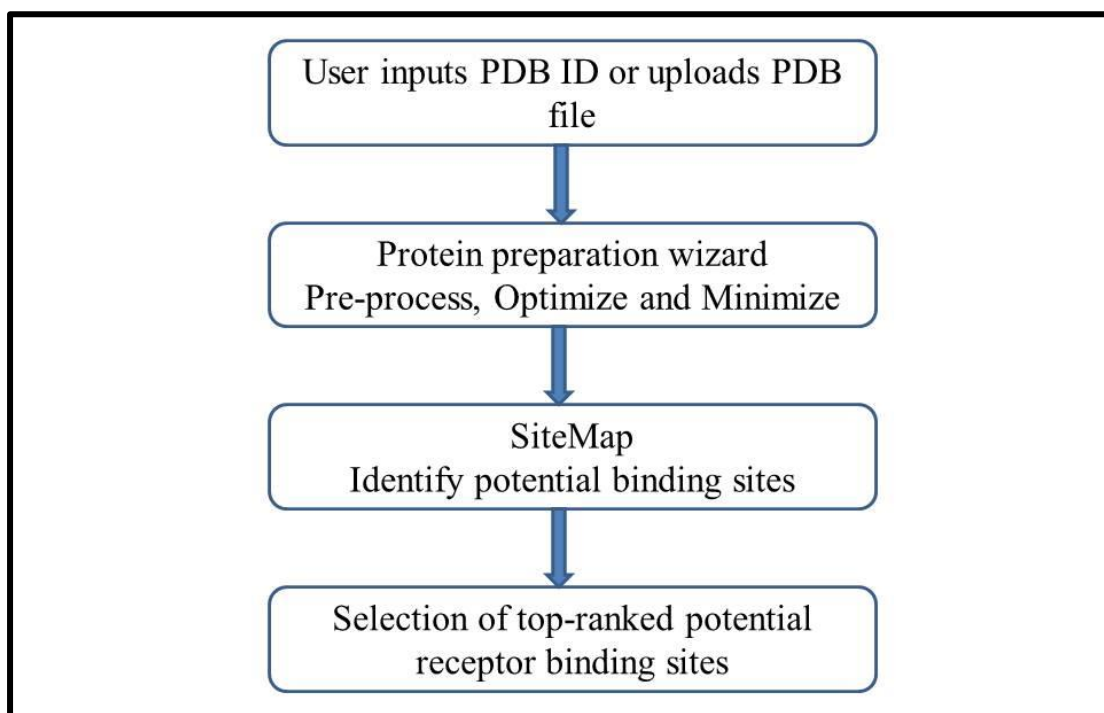
Figure 2.4: A diagrammatic flow of comparative SiteMap steps used in this chapter.

2.2.3. Allosite

The structures from Penkler *et al.,* 2017 were prepared using Discovery studio 2017. All endogenous ligands, co-factors, metal ions and water molecules were removed. The prepared structures were then uploaded to the Allosite server (http://mdl.shsmu.edu.cn/AST/). Allosite uses the feature-based and perturbation-based methods in the identification of allosteric sites. It then ranks the sites in terms of their allosteric characteristics such as druggability of the site, its volume, and SASA. Figure 2.5 shows the diagrammatic flow of comparative Allosite steps used in this chapter.
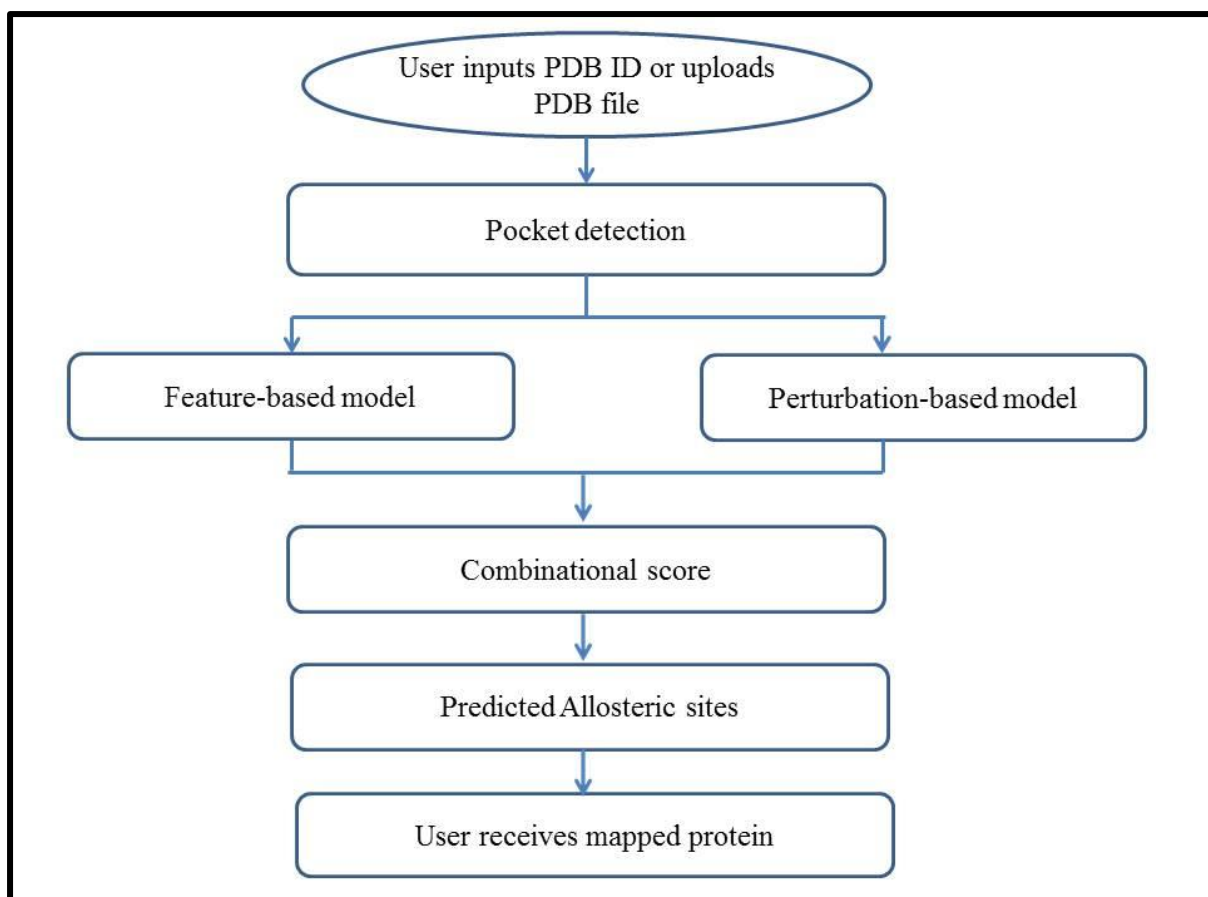
Figure 2.5: A diagrammatic flow of comparative Allosite steps used in this chapter.

2.3. Results and Discussions

2.3.1. FTMAP results of H_prime

FTMAP identified ten binding hot spots on H_prime (Figure 2.6). What is interesting to note is that most of these identified sites are clustered together. For example, Sites 4, 5, 6, 9 and 10 clustered together in the substrate binding domain (SBDα and SBDβ). These sites clustered around the sites occupied by the peptide substrate. Sites 1, 2, 3, and 7 assumed the same location in the nucleotide binding domain (IA, IIA, and IIB). These sites clustered around the sites occupied by the endogenous ligand (ATP). Site 8 (located in NBD subdomain IIA) distinctively stood out from clustering amongst the rest indicating a site of interest as it does not compete with the endogenous ligands present in the protein (Figure 2.6).
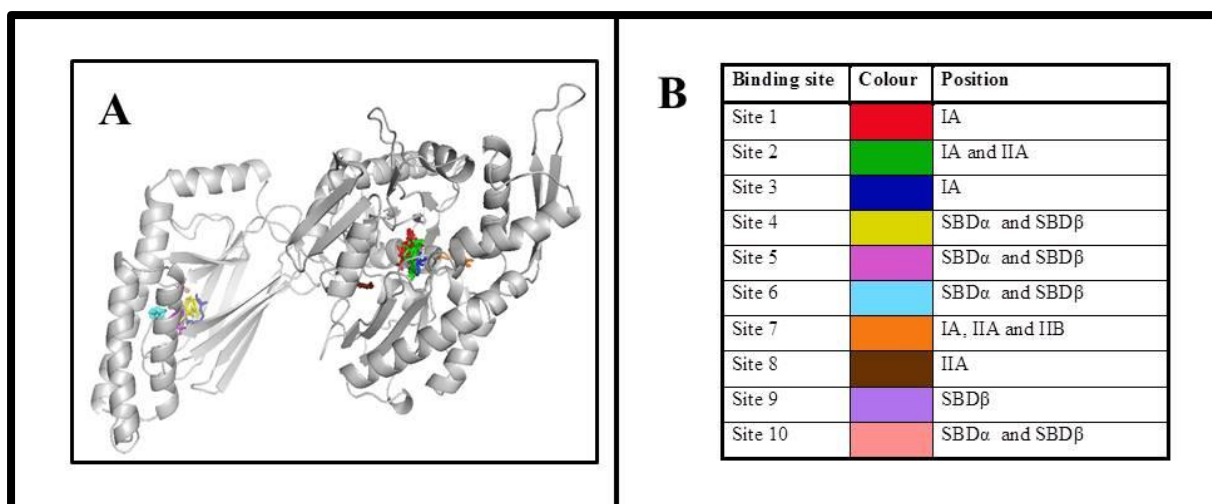
Figure 2.6: Sites identified by FTMAP on H_prime. **A:** shows the sites mapped on the protein structure. **B:** shows the colours representing each site.

2.3.2. FTMAP results B_prime

The number of sites identified by FTMAP on B_prime was significantly greater than that of H_prime. FTMAP identified 16 sites on this protein structure (Figure 2.7). Some sites clustered together while others stood apart. Sites 1, 8, 11 and 13 clustered together in the subdomains (IIA and IIB) of the nucleotide binding domain. Sites 2 and 3 clustered together in the subdomain IIA of the NBD. Sites 5 and 10 also clustered together in the subdomain SBDα. Sites 4, 6, 12 and 16 were seen to cluster together in the subdomain IA. The sites that stood apart were Sites 7, 9, 14, and 15.
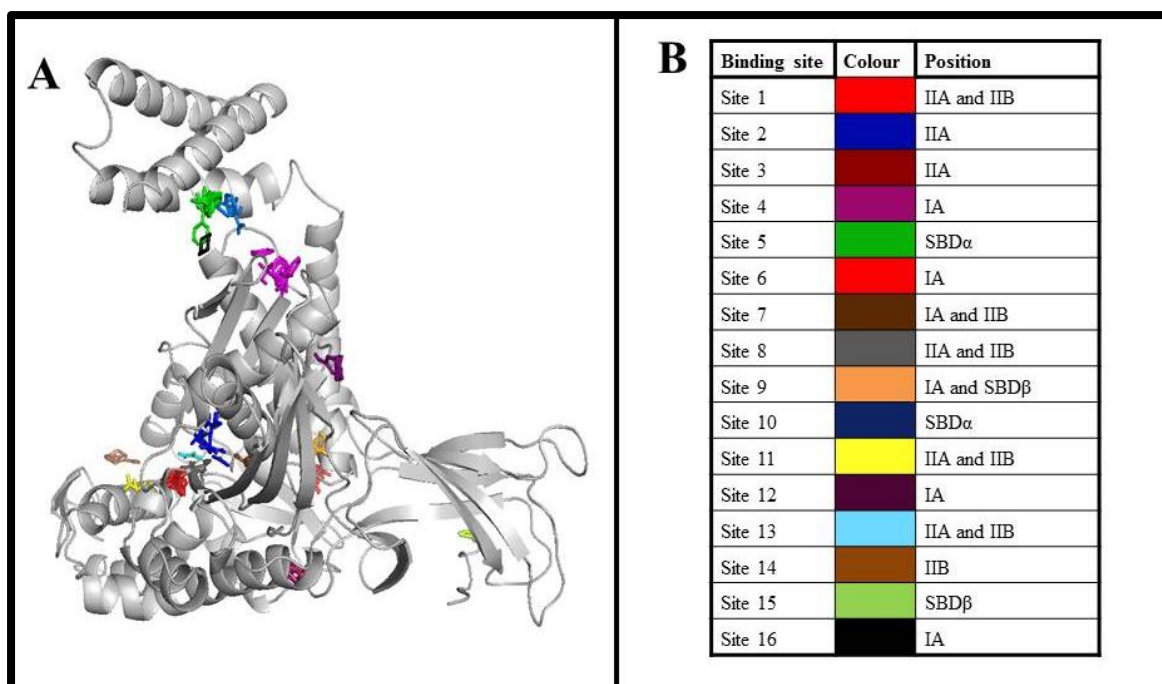
Figure 2.7: Sites identified by FTMAP on B_prime. **A:** shows the sites mapped on the protein structure. **B:** shows the colours representing each site.

2.3.3. SiteMap results of H_prime

SiteMap identified the five top sites on H_prime based on the S-score, D-score, size, and volume (Table 2.1). S-score greater than or equals to 0.80 could thus be potential ligand binding sites. The druggability of each site was measured as the D-score. The D-score is an important scoring function of potential ligand binding sites as it includes terms that propagate ligands binding such as adequate size and solvent isolation. Using the D-score criteria, sites as categorized as undruggable (highly hydrophilic sites that are comparatively smaller in size, having little or no hydrophilic characteristics and a D-score value lesser than 0.83), difficult to drug (adequately hydrophilic with less hydrophobic characteristics and having a D-score value greater than 0.83 but less than 0.98) and druggable (sites with favourable size, hydrophilic and hydrophobic characteristics and having a D-score value greater than 0.98) (Rodina *et al.,* 2013). Table 2.3 shows the top five sites identified by SiteMap. Sites 1 and 2 had significantly high S-score and D-score.

Table 2.2: Sites identified by SiteMap on H_prime.

Binding site	Position	S-score	D-score	Size	Volume
Site 1	IB, IIA, and IIB	1.050	1.037	257	723.040
Site 2	IA, IB and IIA	0.975	0.982	172	429.440
Site 3	IIA	0.933	0.906	60	194.480
Site 4	IA and IB	0.903	0.522	39	92.950
Site 5	IA, SBDα and SBDβ	0.724	0.627	45	154.350

2.3.5. SiteMap results of B_prime

The top five potential ligand binding sites on B_prime were identified by SiteMap (Table 2.2). All sites had an S-score of greater than 0.83 which signifies plausible ligand binding sites. However, only sites 1 and 2 had a druggable site as their D-score value was greater than 0.98. Site 3 had a D-score of 0.927 which is categorized as difficult to drug while sites 4 and 5 had D-scores less than 0.83 which is classified as undruggable.

Table 2.3: Sites identified by SiteMap on B_prime

Binding site	Position	S-score	D-score	Size	Volume
Site 1	IIA and IIB	1.160	1.058	103	171.500
Site 2	IA, SBDα and SBDβ	0.991	1.029	105	276.460
Site 3	IIA and IIB	0.906	0.927	74	368.040
Site 4	IA and SBDα	0.939	0.719	79	189.680
Site 5	IIA and SBDβ	0.895	0.748	65	213.690

2.3.6. H_prime results predicted by Allosite

Allosite uses the feature-based and perturbation-based methods in the identification of allosteric sites. The results from the allosteric site search on H_prime by Allosite are summarized in Table 2.3. A total of 9 potential ligand binding sites were predicted by Allosite. Allosite uses the physicochemical characteristics such as the hydrophobicity and the normalized polarity score of the amino acid residues lining the pock to predict its druggability score. These characteristics contribute positively to D-score which indicates that both hydrophobic and polar residues can significantly increase the druggability of the site. The prediction of druggability can then be based on the average score of 0.5 taking into consideration the standard deviation (a higher standard deviation value raises the cut-off to 0.7 and vice versa) (Schmidtke and Barril, 2010).

Table 2.4: Sites identified by Allosite on H_prime.

Binding pocket	Position	Volume	SASA	D-score	Logit Prob	Nma score	Hit score
Pocket 1	IB and IIB	1177.958	701.917	0.540	0.900	1.000	0.920
Pocket 2	IB and SBDβ	1438.481	640.068	0.049	0.606	1.000	0.685
Pocket 3	IIA	1035.300	528.973	0.330	0.720	1.000	0.776
Pocket 4	IA, IB and IIA	516.473	366.091	0.028	0.391	1.000	0.513
Pocket 5	IA	619.370	498.270	0.025	0.538	1.000	0.630
Pocket 6	IIA and IIB	437.706	248.777	0.086	0.493	1.000	0.595
Pocket 7	SBDα	148.114	37.050	0.884	0.599	1.000	0.680
Pocket 8	IB	386.517	275.921	0.171	0.480	1.000	0.584
Pocket 9	IA	212.944	67.047	0.147	0.541	1.000	0.633

2.3.7. B_prime results predicted by Allosite

Allosite predicted 9 potential ligand binding sites on B_prime as seen in Table 2.4. Allosite ability to estimate druggability certainly adds structural dimension and assesses the possibility that small drug-like molecules can bind to a specific target with the potency of altering its activity (Schmidtke and Barril, 2010).

Table 2.5: Sites identified by Allosite on B_prime.

Binding pocket	Position	Volume	SASA	D-score	Logit Prob	Nma score	Hit score
Pocket 1	SBDβ	819.628	417.703	0.587	0.841	1.000	0.872
Pocket 2	IIA and IIB	960.575	497.172	0.203	0.812	1.000	0.850
Pocket 3	IIA and SBDβ	1192.129	668.567	0.042	0.696	1.000	0.757
Pocket 4	IIA	277.559	121.206	0.036	0.482	1.000	0.585
Pocket 5	SBDα	580.689	241.483	0.115	0.543	1.000	0.635
Pocket 6	IA and SBDβ	445.796	180.843	0.016	0.468	1.000	0.574
Pocket 7	SBDβ	253.425	141.084	0.032	0.437	1.000	0.550
Pocket 8	SBDα	357.036	217.171	0.014	0.146	1.000	0.533
Pocket 9	SBDβ	441.009	176.161	0.067	0.452	1.000	0.562

2.3.8. Hotspot analysis

Allosteric sites are designed to develop more selective and specific enzyme activity inhibitors and to discover new functions (Roca *et al.,* 2018). Previous study by Penkler *at al.,* 2017 identified allosteric hot residues on *E. coli* Hsp70 using PRS in combination with MD simulation (Table 2.5). Conclusions drawn from this study suggests that allostery is solely determined by nucleotide-mediated conformational redistributions in the NBD and SBD and that these identified residues are critical for conformational stability and allosteric communication (Penkler *et al.,* 2017).

Table 1.6: Allosterically relevant residues identified by Penkler et al., 2017 using PRS in combination with MD simulation.

Structure	Position	Allosterically relevant residues
H_prime	IA	11, 12, 17-19, 113, 116, 117, 142, 143, 144, 146, 148, 149, 150, 156, 157, 167, 370, 373, 374, 376, 377, 380
	IB	59, 60, 66, 68, 70, 71, 73, 96-99
	IIA	195-201, 202, 203, 219-223, 225-227, 311-321, 324, 325, 329, 330, 333, 353, 356
	IIB	228, 229, 231-235, 252, 255, 256, 261, 264, 265, 281, 283, 284
	Linker	392
	SBDβ	396-414, 417, 429-444, 448, 450, 451-455, 457, 464-467, 478, 479, 485, 486, 498, 500-502
	SBDα	503-508, 511, 512, 514, 515, 516, 517, 520, 530-538, 540, 546-568, 571, 581-602
B_prime	IA	7, 113, 120, 127, 139-141, 143-146, 148, 149-153, 156, 164, 166, 167, 170, 171, 381
	IIA	192, 193, 195, 198-201, 205, 207, 216,-218, 224-227, 310-314, 316, 338-340, 344-347, 349-351, 354, 361, 362
	IIB	228-240, 243, 267, 270, 271, 273-281, 297-299, 303-309
	Linker	392, 393
	SBDβ	398, 400-410, 413, 414, 419, 425-440, 446, 449-451, 453-490, 493, 494, 497-502
	SBDα	503-505, 515, 520, 539, 540, 543, 550, 551, 554, 582

In this study, we utilized a combination of computational tools to obtain a consensus prediction of protein druggable sites. Each of the servers used in the identification of potential ligand binding sites on both protein structures utilizes different techniques. FTMAP samples the surface of the protein with 16 different probes. SiteMap identifies sites based on the D-score, size, S-score, and volume of the sites respectively. Allosite identifies sites based on the volume, druggability and the total solvent-accessible surface area of the predicted allosteric sites. Validation on known allosteric sites is important in this study. As seen in Figure 2.8, there was at least one consensus site between all three prediction tools on both protein structures. Ligands binding to these consensus sites will be selected and evaluated for inhibition purposes. We identified residues around the consensus sites and these residues were in agreement with results from PRS data, hence the reason for selection as seen in Table 2.6 and Table 2.7. Hence, these sites will be screened for potential allosteric modulators in Chapter 3.
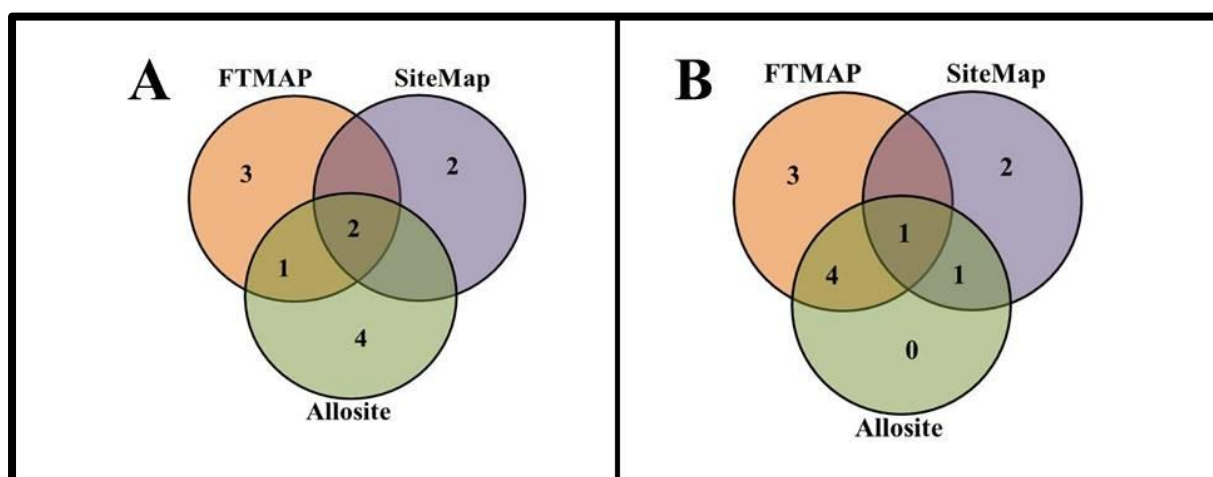
Figure 2.8: Venn diagram of results comparison by methods used in the identification of allosteric sites. A: FTMAP, SiteMap, and Allosite all identified 2 common sites in H_prime. These sites according to SiteMap have D-scores of 0.982 and 0.906 and S-scores of 0.975 and 0.933 respectively. B: FTMAP, SiteMap, and Allosite all identified 1 common site in B_prime having D-score of 0.927 and an S-score of 0.906 respectively.

Table 2.7: Residues of the consensus site found in H_prime.

Allosteric site	Position	Residues
Site 1	IA, IB and IIA	G6, I7, D8, L9, S14, V16, A17, I18, M19, P24, R25, V26, L27, E28, D33, R34, T35, R75, R76, T141, V142, P143, A144, Y145, F146, N170, E171, F200, D201, I202, T221, N222, G223, D224, P367, D368, E369, A370, V371, A372, I373, G374, A375, A376

Indicated in red are residues identified by Penkler et al., 2017 using PRS in combination with MD simulation

Table 2.8: Residues of the consensus sites found in B_prime.

Allosteric sites	Position	Residues
Site 1	IA, SBDβ and SBDα	I73, G74, R75, A95, A96, D97, N98, G99, D100, A101, D148, A149, Q150, R151, Q152, A153, T154, K155, D156, A157, K166, G167, I168, L382, T383, K452, S453, L454, G455, D479, A480, D481, G482, I483, L484, S504, S505, G506, L507
Site 2	IIA, Linker and SBDβ	D201, I202, S203, V218, L219, A220, T221, N222, G223, D224, T225, D311, L312, N314, R315, S316, I317, E318, P319, L320, L321, V322, A323, L324, Q325, D326, A327, G328, L392, D393, V394, T395, P396, L397, S398, A413, K414, N415, T416, T417, I418, P419, T420, K421, H422

Indicated in red are residues identified by Penkler et al., 2017 using PRS in combination with MD simulation

2.4. Conclusion

The identification of druggable sites in therapeutic studies is crucial for structure-based drug design in order to discover allosteric sites and to design small molecules with therapeutic effects that bind to these sites (Roca *et al.,* 2018). Two protein structures were extracted from Penkler *et al.,* 2017. These structures were subjected to allosteric site search using three different computational tools. Using a combination of methods may give better predictions rather than either method individually. Each tool utilizes different algorithms in the prediction of potential allosteric sites and scores these identified sites based on different physicochemical properties such as D-score, perturbation score S-score, size, total solvent-accessible surface area, and volume. A search for consensus sites identified by all three servers was done and this search revealed that these servers identified two consensus sites in H_prime and one consensus site in B_prime. The first consensus site found in H_prime was within the nucleotide binding domain IA, IB, and IIA, while the second site was also found in the nucleotide binding domain IIA. The only consensus site found in B_prime was located within the nucleotide binding domain IIA and IIB. Residues around one of the consensus site in H_prime (IA, IB and IIA) were in agreement with residues identified by Penkler *et al.,* 2017, hence this site can be considered as an allosteric site. Residues around the only consensus site in B_prime were not in agreement with PRS data; hence we selected two sites identified by SiteMap whose surrounding residues were in agreement with residues identified by Penkler *et al.,* 2017. These sites (IA, SBDβ, and SBDα) and (IIA, Linker, and SBDβ) were partly identified in other tools as SiteMap predicted druggable regions beyond FTMAP and Allosite. Ligands selection in Chapter three will be based on those binding to the selected sites whose surrounding residues correspond with PRS data.

In conclusion, these servers have displayed a level of competency in the prediction of potential allosteric binding sites. Further analysis will be carried out evaluate these predicted sites in the discovery of potential inhibitors.

CHAPTER THREE: HIGH THROUGHPUT VIRTUAL SCREENING

3.1. Chapter overview

Allosteric inhibitors bind to sites distinct from the active site and are capable of inducing conformational change on a protein structure. This chapter aims at the discovery of new allosteric modulators from compounds from SANCDB (https://sancdb.rubi.ru.ac.za). The entire surface of each structure will be screened against 623 small molecule compounds retrieved from SANCDB using Autodock Vina. Protein and ligand preparation will be done using AutoDock4 tools. The docking protocol was validated by redocking co-crystallized ATP to the protein structure. After that, exact docking parameters will be implemented in the virtual screening procedure. Ligands were selected based on their interactions with the allosteric sites identified in Chapter 2. Ligand poses interacting with low binding energies from each conformation will be identified as potential allosteric inhibitors and further analysis will be done on them.

3.1.1. High throughput virtual screening

The quest for a new means of drug discovery is a continuous effort as there is a global rise in various diseases afflicting the world (Truglio *et al.,* 2012). For many years, the development of small molecules as therapeutic agents for the prevention and treatment of diseases has played a crucial role in medical practice (Schiavone and Trabace, 2018). Previous methods used in the discovery of drugs were done through the random screening i.e. trial and error of medicinal plants. Although inefficient, this method yielded the identification of lead compounds (Giridhar, 2012). New experimental techniques such as high-throughput screening (HTS) have emerged in the discovery of novel potential molecules and have enhanced random screening efficiency (Michael *et al.,* 2008). The emergence of this new technique has caused an increase in the amount of available compound and biomedical data over the past decade and has become the standard method for drug discovery (Chen *et al.,* 2018). Amid its massive contributions to drug discovery, this standard method is expensive and unsatisfactory as it does not provides information such as the compound's toxicity and bioavailability prior to its development (Proudfoot, 2008). Virtual screening is an *in vitro* analogue of HTS that provides structural information of compounds and is cost, resource and time effective (Ballester *et al.,* 2012).

Virtual screening (VS) is a potent computational technique for the identification of lead molecules in drug discovery (Lavecchia and Giovanni, 2013). It is a structurally based method in the discovery of novel compounds as it relies on structural information of the target protein. This method screens large compound databases for compounds that complement protein targets with known structures and experimentally test those that are expected to bind well (Shoichet, 2004). For the purpose of the study, protein structures will be screened against compounds from SANCDB (Hatherley *et al.,* 2015). VS methods are divided into two distinct categories: ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS). LBVS is based on ligand similarity and it draws on the fact that ligands structurally similar to a biologically active ligand are more likely to be active than random ligands (Hamza *et al.,* 2012). This method is only suitable when there is no 3D target protein structure. Examples of LBVS include fingerprints, ligand-based pharmacophore modeling, and quantitative structure-activity relationship (QSAR). SBVS, unlike LBVS, is based on protein-ligand docking and it utilizes 3D structural information of the target protein to predict binding modes and affinities of ligands (Lionta *et al.,* 2014). The 3D information of proteins are mostly determined by experimental techniques such NMR, X-ray crystallography or by computational techniques such as homology modeling (Lavecchia and Giovanni, 2013). This chapter includes the SBVS approach to successfully identify potential allosteric inhibitors of Hsp70. A well-established approach used in SBVS is molecular docking (Drwal and Griffith, 2013). Other methods used in SBVS include *de novo* drug design and structure-based pharmacophore.

3.1.2. Molecular docking

The identification of new lead molecules that demonstrate pharmacological activity against natural targets and the knowledge of the potency of these molecules remains one of the most important stages in drug discovery (Lionta *et al.,* 2014). Molecular docking is a vital tool in SBVS that investigates the predominant binding modes of small molecules in protein pockets (protein with known 3D structure). A large number of algorithms and scoring functions are currently used to evaluate the protein-ligand interactions as well as the fundamental biochemical processes at an atomic level (Morris and Lim-Wilby, 2008). Molecular docking has become an extremely important part of the current efforts to discover and design drugs (Brooijmans and Kuntz, 2003). Over the past two decades, more than 60 different docking tools and programs with different algorithms have developed for both academic and commercial usage. These include AutoDock, Autodock Vina, DOCK, GLIDE, GOLD,

LigandFit, MOE-Dock, etc. (Pagadala *et al.,* 2017). Autodock Vina was used in this chapter. The process of molecular docking comprises two fundamental steps: ligand conformation, orientation and position prediction within the target sites and binding affinity assessment. It is important to note that the efficiency of the docking process increases when the location of the binding site is known (Meng *et al.,* 2011). In some cases, the binding site of a target protein is known before the ligand is docked into it. This type of docking is called targeted docking. Without knowledge of binding sites, cavity detection programs or online servers may be used for the identification of potential binding sites within the protein. This type of docking is called blind docking (Meng *et al.,* 2011). In this study, online servers such as FTMAP, Allosite, and Schrodinger's SiteMap was used to identify potential allosteric sites on the protein structures.

3.2. Methodology
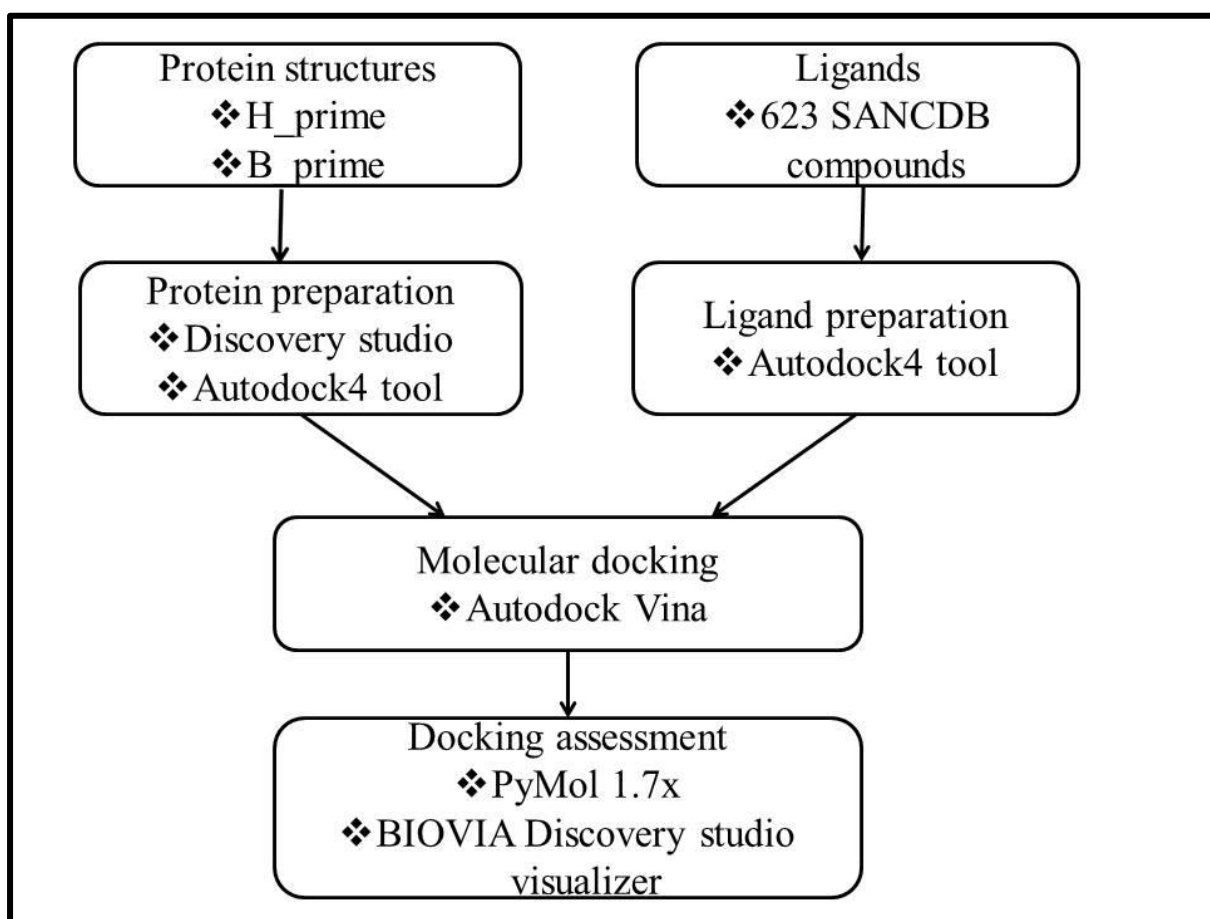
Figure 3.1: A diagrammatic representation of the different steps and tools used for docking studies.

The aim of molecular docking is to predict the position and orientation of a ligand when bound to protein structure (Ghersi and Sanchez, 2009). High throughput virtual screening was carried out on both conformations of protein structures obtained from Penkler *et al.,* 2017, using 623 novel compounds from SANCDB. These compounds were ready to dock as they have been previously minimized by Hatherley *et al.,* 2015. Autodock4 tool, Autodock Vina, Discovery studio visualizer, and PyMol was used to prepare, visualize and dock the ligand. Figure 3.1 shows a summary of the steps and tools used for docking studies.

3.2.1. Protein and ligand preparation

Protein structures H_prime and B_prime were extracted from Penkler *et al.,* 2017 as explained in Chapter 2. Both structures contained a co-factor (ATP) and a metal ion (Mg^{2+}). The structures were prepared using Discovery studio and Autodock4 tool. Polar hydrogen was added using Discovery studio. Gasteiger charges were assigned using Autodock4 tool prepare_receptor4.py for the protein structures (receptor) and prepare_ligand4.py for the ligand structures. This converted both the protein and ligand structures to a pdbqt format. The non-polar hydrogens were also merged.

3.2.2. Docking Parameter Setup

The proteins and ligands were prepared for docking as described above. Blind docking was carried out on both protein structures. The Autodock Vina plugin in PyMol (Seeliger and De Groot, 2010) was used to design a docking box centers prior to the docking simulation. Table 3.1 shows blind docking parameters covering the entire surface of the protein structures. The dimension of the box was calculated such that each side of the box had a clearance of 10Å and default 0.375Å grid spacing was used. An exhaustive of 320 was used.

Table 3.1: Docking parameters used for H_prime and B_prime

Protein	Box size	Center
H_prime	X = 110Å Y = 110Å Z = 110Å	X = 49.00 Y = 74.00 Z = 40.00
B_prime	X = 100Å Y = 100Å Z = 100Å	X = 48.88 Y = 60.67 Z = 39.55

3.2.3. Docking validation

Autodock Vina's ability to reproduce correct poses was first evaluated. Template structure (4B9Q) used in generating one of the structures (B_prime) from Penkler *et al.,* 2017 and its co-factor (ATP) was used to validate the docking parameters used in this study. 4B9Q was obtained from PDB. Water molecules and co-crystallized ligands were removed from 4B9Q using Discovery Studio Visualizer. The protein (4B9Q) and ligand (ATP) were prepared using the receptor and ligand preparation function in Autodock4 tool. ATP was docked to the prepared 4B9Q and its docking pose was compared. Docking protocol was done using Autodock Vina. Re-docked results were compared, and it assumed similar interactions with template structure.

3.3. Results and Discussions

3.3.1. Docking validation

Assessing docking procedures is crucial to drug discovery. 4B9Q was used to assess the docking quality of this study. 4B9Q is *E. coli* Hsp70 protein in its open conformation and contains a co-factor (ATP) and a metal ion (Mg^{2+}). This protein structure is the bacteria homologue of Hsp70. The co-factor ATP was redocked to the co-crystallised structure (4B9Q) which was the on of the template structures used in generating the structures used in this study. The redocked ATP assumed a similar orientation in comparison to its native state

as shown in Figure 3.2. Ligplot⁺ version v.2.1 was used to view and compare the interactions between the redocked poses. The redocked ATP had a binding energy of -8.7kcal/mol.
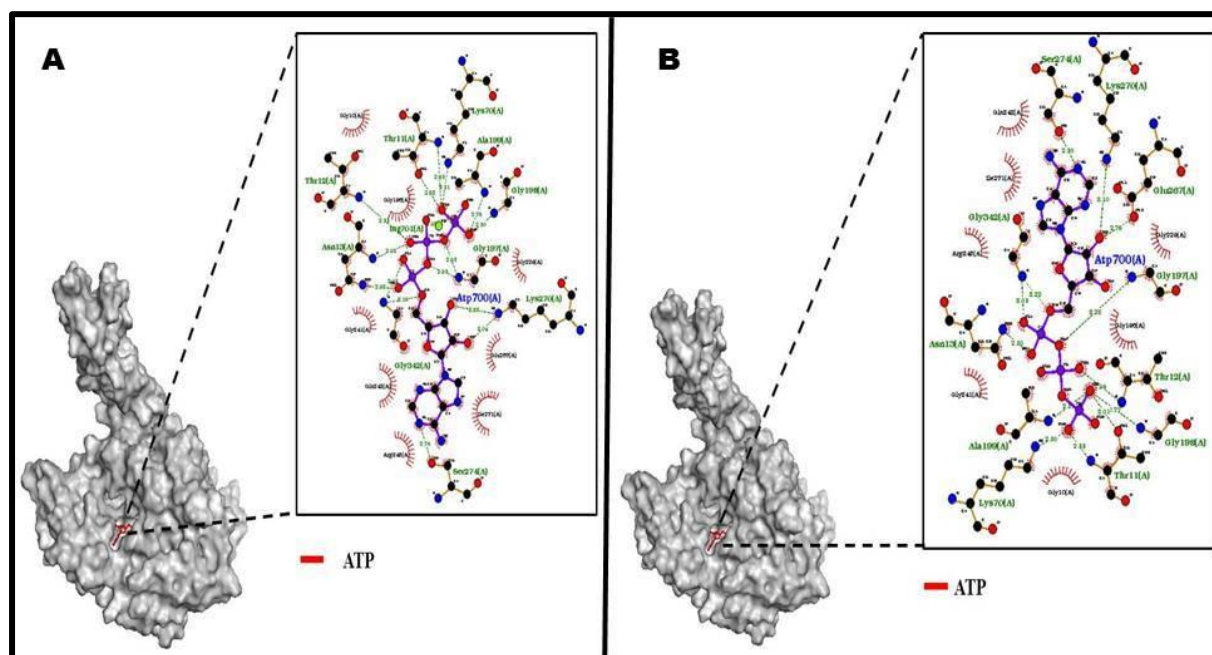
Figure 3.2: Docking validation results: 4B9Q was re-docked with its co-factor (ATP) It assumed similar binding sites in both protein and interacted with the same residues. A: shows the template structure (4B9Q) and its interactions with its co-crystallized co-factor (ATP). B: shows ATP redocked to 4B9Q.

3.2.2 Blind docking analysis

Blind docking involves the docking of ligands on the entire surface of a protein structure without prior knowledge of the catalytic binding sites. H_prime and B_prime were subjected to blind docking using 623 ready to dock compounds from SANCDB. Figure 3.3 shows the overall binding affinities of these compounds indicating the stability of ligands when docked to the respective structures. High specificity of the compounds from SANCDB was observed in both protein structures as 57% of these compounds had binding energies higher than that of ATP (used in validating docking parameters) while 6% of the compounds had similar binding energies with that of ATP. This could be attributed to the presence of many functional groups in the compounds. Interestingly, different binding patterns were observed in both H_prime and B_prime. Ligands with the highest binding affinity (lowest binding energies) observed in H_prime had a relatively high binding affinity in B_Prime. For example, SANC00662 had the highest binding affinity with binding energy of -12.5kcal/mol in H_prime, had binding energy of -7kcal/mol when bound to B_prime. Similarly, SANC00628 which had the lowest binding affinity with binding energy of -5.2kcal/mol in

H_prime, had binding energy of -7.3kcal/mol in B_prime. From Figure 3.3, it is evident that H_prime was more receptive to compounds from SANCDB as compared to B_prime.
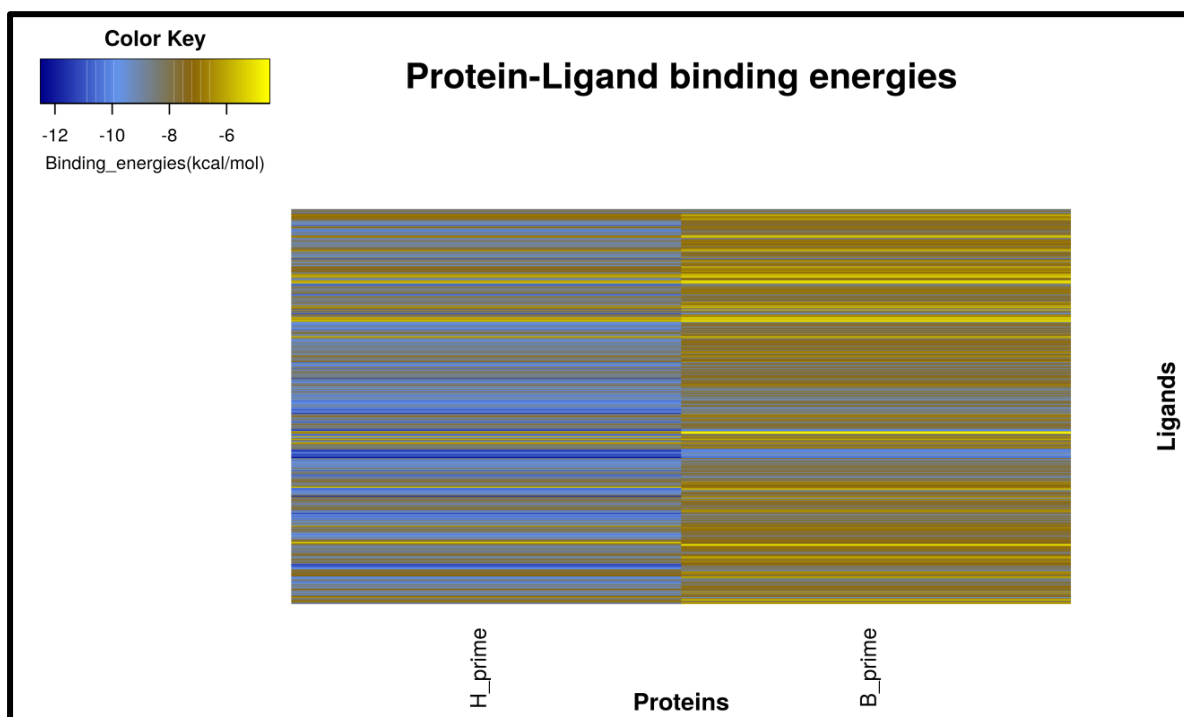
Figure 3.3: A heat map plot showing the binding energies across both protein structures. Ligands binding to H_prime had the best (lowest) binding energies.

3.2.2.1 H_prime

Figure 3.4 shows the docking results on this protein. Ligand binding was distributed across all domains and subdomain of the protein. Ligands binding to the ATP and substrate binding sites were discarded as the aim of this study is to identify potential allosteric modulators and not ligands competing with either ATP or peptide substrate. Other criteria used in the selection of ligands include ligands binding within 10Å of the identified allosteric sites (Chapter 2), ligands interacting with identified allosteric hot residues and ligands with binding energy less than -6kcal/mol. Ligands binding to H_prime had binding energies ranging from -5.2 to -12.5 kcal/mol (Figure 3.3).
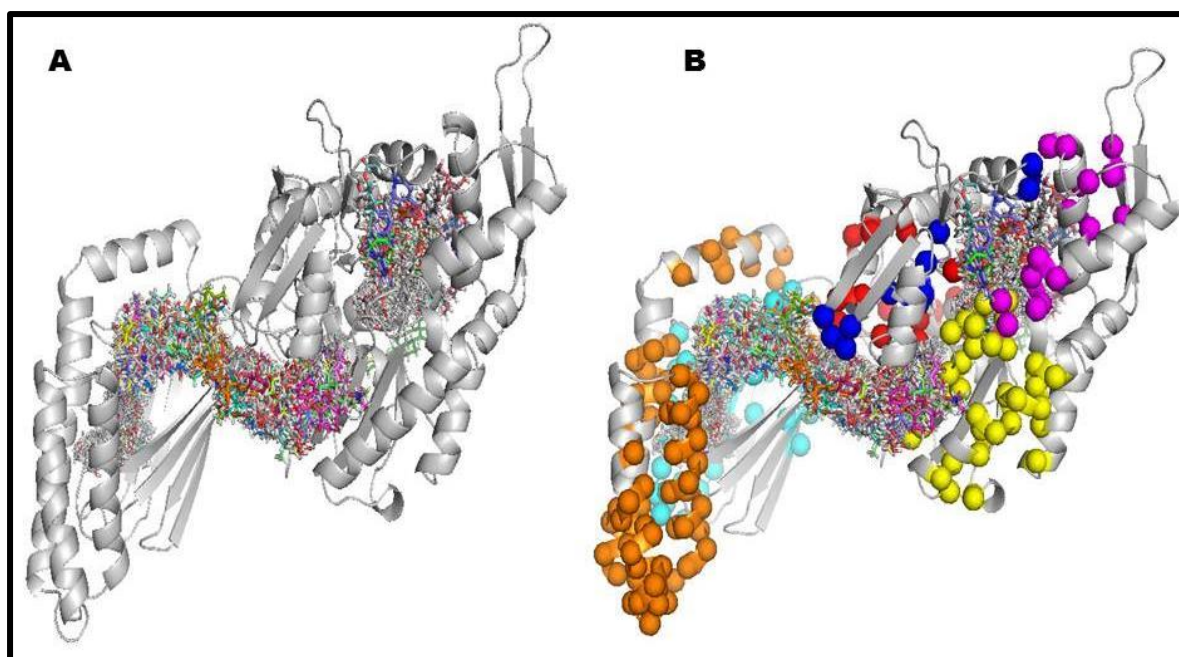
Figure 3.4: A: Blind docking results of SANCDB compounds on H_prime. Ligand binding was distributed across all domains and subdomains of the structure. **B:** Allosteric residues identified by Penkler et al., 2017 mapped by Cα atom

3.2.2.2. B_prime

Ligands binding in B_prime was not evenly distributed as seen in H_prime as the majority of the ligands scattered on the surface of the protein (Figure 3.5). Ligands were not observed in the ATP and substrate binding sites, indicating that compounds from SANCDB did not compete with ATP. Majority of the ligands can be seen binding to the NBD and just a few bound to the SBDα. Ligands binding to H_prime had binding energies ranging from -4.5 to -10.6 kcal/mol (Figure 3.3).
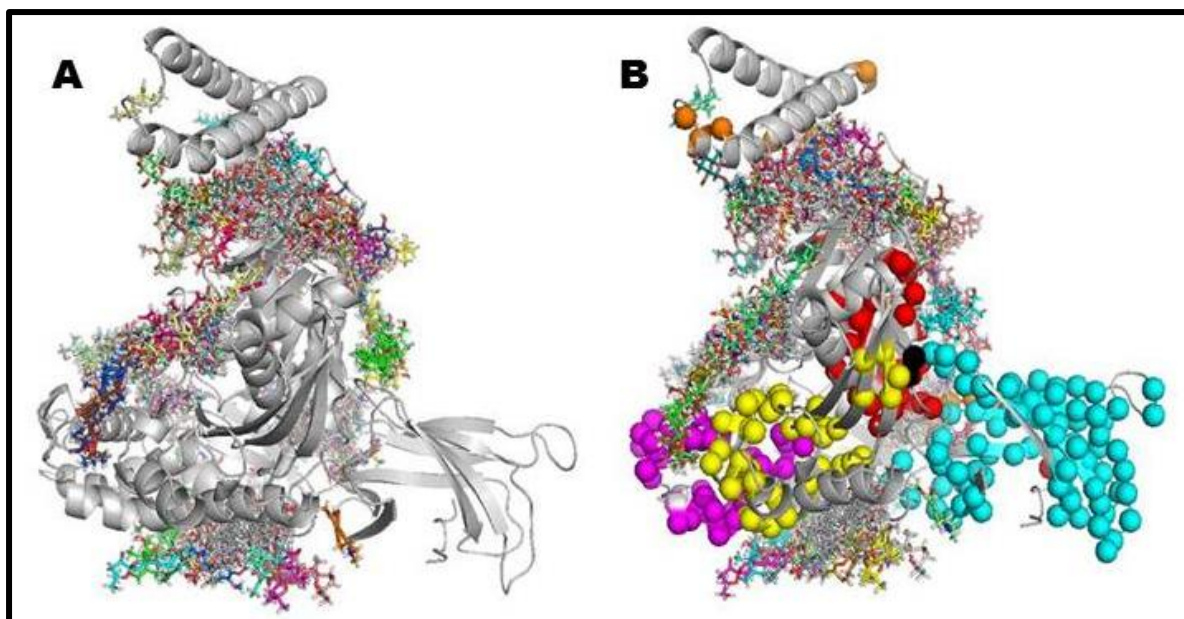
Figure 3.5: A: Blind docking results of SANCDB compounds on B_prime. Ligand binding was distributed across all domains and subdomains of the structure. **B:** Allosteric residues identified by Penkler et al., 2017 mapped by Cα atom

3.2.3 Criteria for ligand selection

Selection of ligands was done based on the criteria in Figure 3.6. These criteria include selecting ligands binding within 10Å of the allosteric sites identified in Chapter 2, ligands make significant interactions with allosteric residues identified by Penkler *et al.,* 2017, and ligands with binding energies of ≥-6kCal/mol. Discovery studio visualizer was to visualize the various interactions between the ligand-protein complexes. A bash script was used in selecting ligands binding within 10 Å of the identified allosteric sites. The x, y, z coordinates were determined from the log file provided by SiteMap. The following equation (Lambers, 2009) was used to calculate the distance between the allosteric site and the central point of each ligand, where x_1, y_1, and z_1 represent the coordinates of the allosteric site and x_2, y_2, and z_2 represent the central point of each ligand.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$
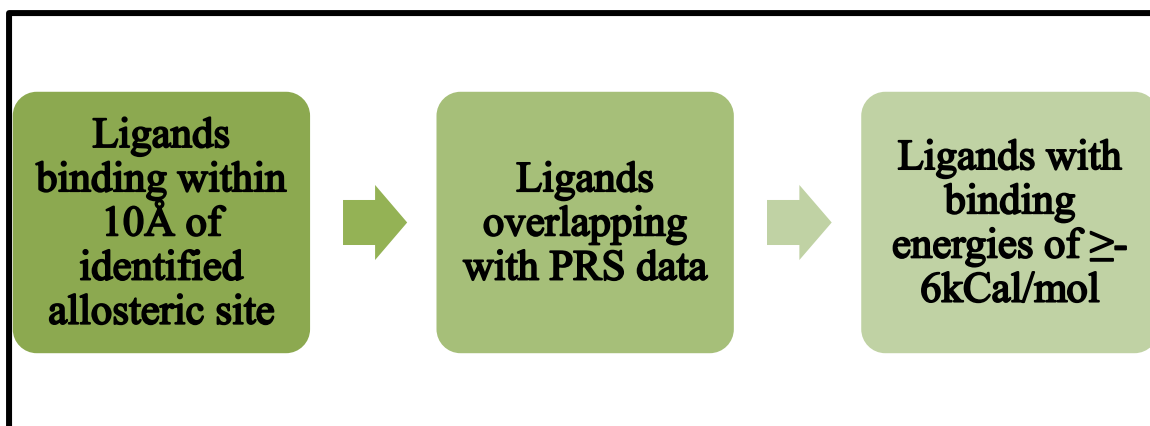
Figure 3.6: Criteria for selecting ligands.
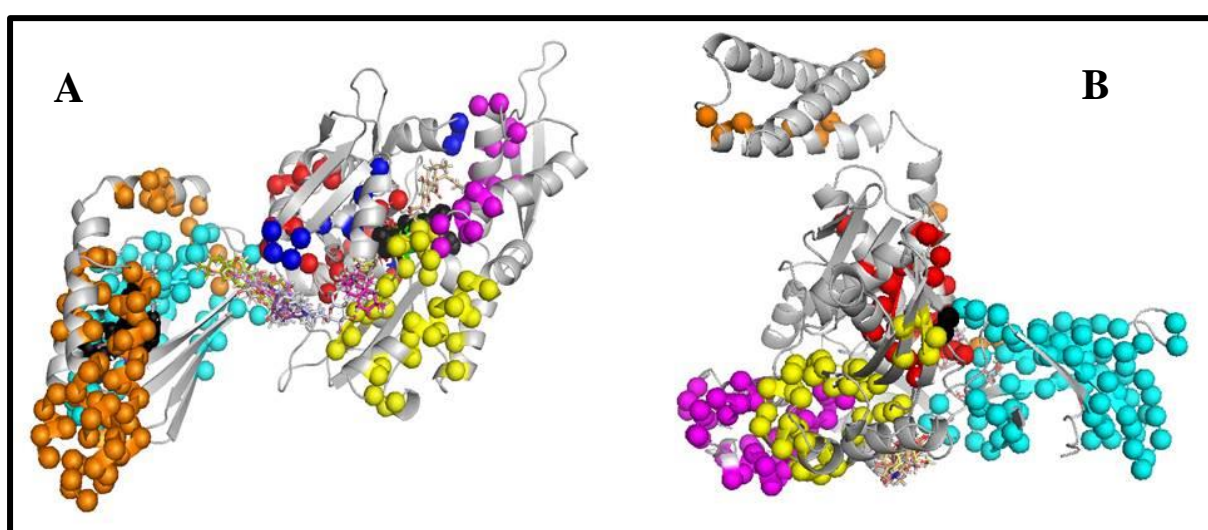
Figure 3.7: Selected ligands binding on their respective protein structures. **A:** H_prime and **B:** B_prime

3.2.3.1 H_prime

In order to identify allosteric modulators, we performed virtual screening on H_prime using 623 compounds from SANCDB. Results from Chapter 2 identified potential allosteric sites using three different prediction tools. A consensus site was identified between these tools and this site was further validated as an allosteric site as the residues surrounding the site were in agreement to the allosteric residues identified by Penkler *et al.,* 2017. Ligands binding to this site were isolated and analysed. Most of the sites identified by these tools clustered together, hence, some of the ligands notably bound to more than one allosteric region. This explains the reason why each ligand had different interactions with residues surrounding the allosteric site (Table 3.9).

Of the 623 compounds from SANCDB screened against this protein structure, 295 compounds bound partially or totally to the identified allosteric site. Seven of these ligands

selectively bound the identified allosteric site were selected based on the distance between the ligands and allosteric sites, their interactions with allosteric residues identified by Penkler *et al.,* 2017 and their binding energies (Figure 3.7A). These interactions include conventional hydrogen bonds, Van der Waals, Alkyl, Pi-Alkyl, Pi-Cation, Pi-Sigma, Pi-Pi T-shaped, Pi-Pi Stacked interactions (Figure 3.8). Due to the 10Å cut-off distance used in selecting ligands, majority of the ligands interacted with the same residues, while a few had unique interactions with identified allosteric residues.

Interestingly, SANC00378 and SANC00430 bound the allosteric sites identified in both protein structures and showed good interactions with identified allosteric hotspot as seen in Table 3.2 and Table 3.4. In H_prime, SANC00378 and SANC00430 had binding energies of -7.7kcal/mol and -7.6kcal/mol respectively, while in B_prime they had binding energies of -6.9kcal/mol and -7.4kcal/mol respectively. Although these compounds bound to the allosteric sites identified in H_prime and B_prime, they had different types of interactions with different allosteric residues as seen in Table 3.2 and Table 3.4. Residues highlighted in red in Table 3.2 are allosteric hot residues identified by Penkler *et al.,* 2017.

Table 3.3 shows the characteristics of the ligands in terms of its compounds name, 2D structure, source organism(s), classifications and uses. Although the uses of most of the ligands have not been recorded, SANC00499 have recorded cytotoxic activity on leukemia (Kuroda *et al.,* 2001; Kuroda *et al.,* 2002) while SANC00430 have recorded anticancer activity (Fouche *et al.,* 2008).

Table 3.2: Selected ligands binding to the allosteric site and making substantial interactions with identified allosteric residues in H_prime.

Ligands	BE (kCal/mol)	Hydrogen bonds interactions	Other interactions
SANC00499	-10.1	R76, N98, R151, T221, N222, T417, P419, K421	D97, A144, Y145, F146, N147, D148, K214, T215, F216, E217, V218, A220, D326, L391, T416, T420, D479, A480
SANC00685	-9.7	N98, Y145, F146, D148, R151, T221	R76, Q78, D79, D97, A144, N147, F216, E217, V218, L219, A220, N222, D326, P419, T420, K421
SANC00492	-8.8	N222, G223	G74, R75, R76, F77, Q78, D79, V82, P143, A144, Y145, F146, T199, F200, D201, I202, S203, T221, N224, T225
SANC00340	-7.8	N98, Y145, F146, D148, R151, K421	A144, N147, F216, E217, V218, T416, P419, T420, D479
SANC00357	-7.8	F200, N222, G223	R75, R76, F77, Q78, D79, V82, T199, D201, T221, D224, T225
SANC00378	-7.7	R76	G74, R75, F77, Q78, D79, V82, P143, A144, Y145, F146, T199, F200, D201, T221, N222, G223, D224, T225
SANC00430	-7.6	R76, G223	R75, F77, Q78, D79, V82, Y145, T199, F200, D201, I202, T221, N222, D224, T225

Other interactions include Van der Waals, Alkyl, Pi-Alkyl, Pi-Cation, Pi-Sigma, Pi-Pi T-shaped, Pi-Pi Stacked interactions. Indicated in red are residues identified by Penkler *et al.,* 2017 using PRS in combination with MD simulation
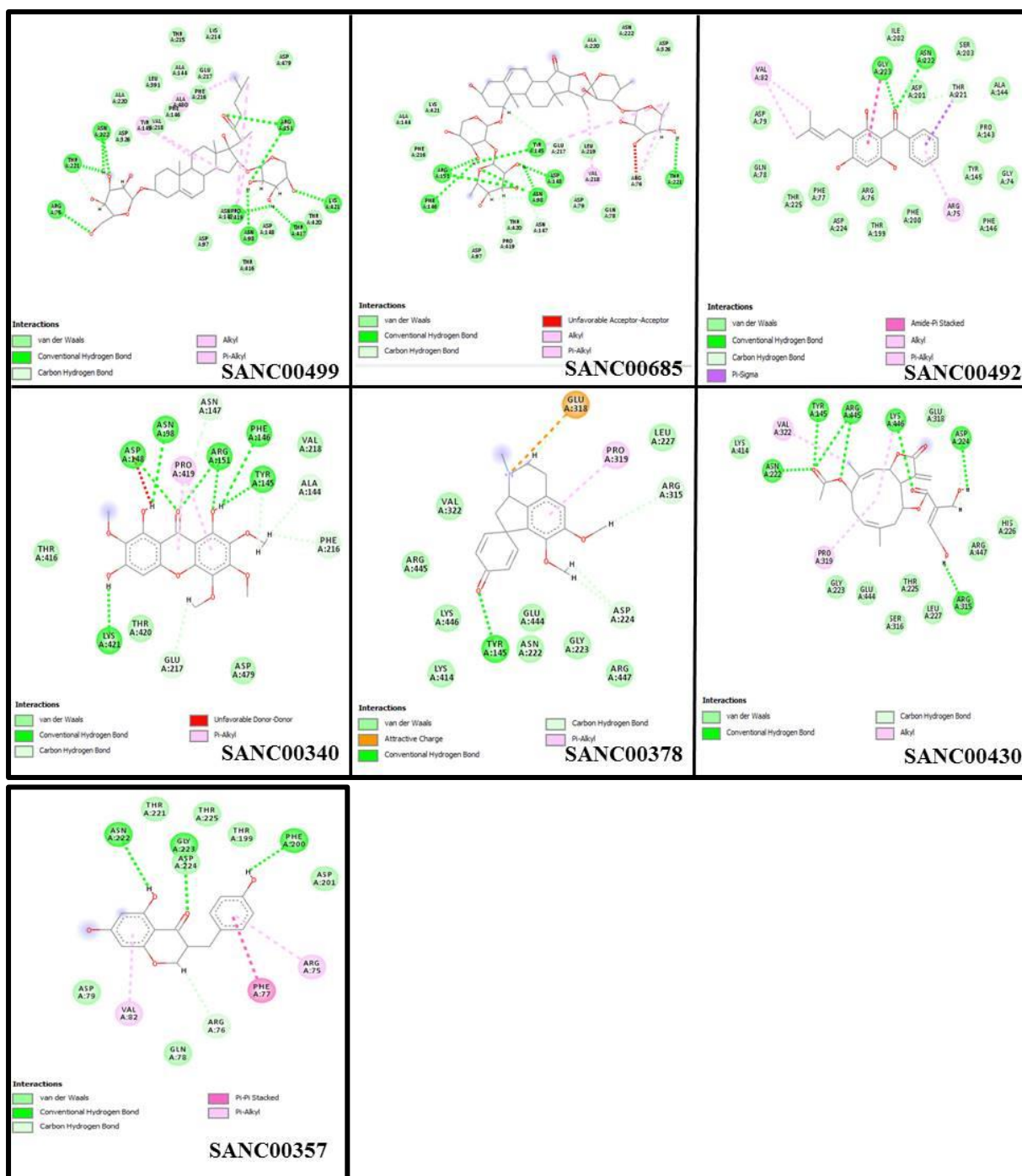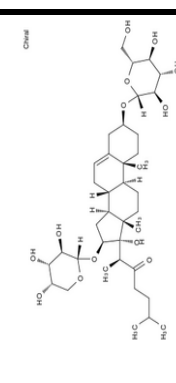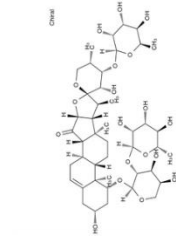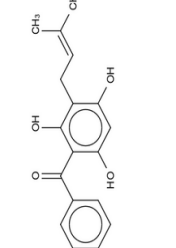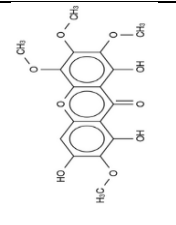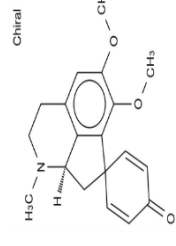
Figure 3.8: Protein-ligand interactions in H_prime as listed in the table above using Discovery studio visualizer.

Table 3.3: Characteristics of ligands selected in H_prime.

Ligand	Compound Name	2D Structure	Source organisms	Classifications	Uses	References
SANC00499	16β-[(α-L-Arabinopyranosyl)oxy]-3β-[(β-D-glucopyranosyl)oxy]-17α-hydroxycholest-5-en-22-one		*Galtonia candicans*	Cholestane, Glycoside	Cytotoxic activity on HL-60 leukemia cells	(Kuroda *et al.,* 2001) (Kuroda, *et al.*, 2002)
SANC00685	Ornithosaponin B		*Ornithogalum thyrsoides*	Unclassified	None recorded	(Kuroda *et al.,* 2006)
SANC00492	Methanone, phenyl[2,4,6-trihydroxy-3-(3-methyl-2-buten-1-yl)phenyl]		*Helichrysum asperum*	Acylphloroglucinol	None recorded	(Jakupovic *et al.,* 1989)
SANC00340	1,6,8-Trihydroxy-2,3,4,7-tetramethoxyxanthone		*Securidaca longepedunculata*	Xanthone	None recorded	(Meyer *et al.,* 2008)
SANC00378	Pronuciferine		*Cissampelos capensis*	Alkaloid, Aporphine	None recorded	(De Wet *et al.,* 2011)

SANC00430	Eucannabinolide		*Schkuhria pinnata*	Sesquiterpene, Terpenoid	Anticancer activity	(Fouche *et al.,* 2008)
SANC00357	4'-Demethyl-3,9-dihydroeucomin		*Resnova humifusa, Eucomis montana, Drimiopsis maculata*	Flavonoid, Homoisoflavonoid	None recorded	(Koorbanally *et al.,* 2006) (Koorbanally *et al.,* 2006) (Koorbanally, *et al.,* 2006) (Du Toit *et al.,* 2005)

3.2.3.2 B_prime

Results from Chapter 2 identified potential allosteric sites using three different prediction tools. Two consensus sites were identified between these tools and these sites were further validated as an allosteric site as the residues surrounding the site were in agreement to the allosteric residues identified by Penkler *et al.,* 2017. Of the 623 compounds from SANCDB screened against this protein structure, 163 ligands bound partially or totally to Site 1 while 143 ligands bound partially or totally to Site 2. A total of six ligands were selected, three ligands were selected from each site based on the distance between the ligands and allosteric sites, their interactions with allosteric residues identified by Penkler *et al.,* 2017 and their binding energies (Figure 3.9). These ligands were isolated and analysed. Residues highlighted in red in Table 3.2 are allosteric hot residues identified by Penkler *et al.,* 2017. Table 3.5 shows the characteristics of the ligands in terms of its compounds name, 2D structure, source organism(s), classifications and uses. Antiviral activity was recorded on SANC00284 (Bessong *et al.,* 2005), while SANC00676 have recorded Murine P388 lymphocytic leukemia (PS) cell-growth inhibition, and tubulin inhibitory activity (Singh and Pettit, 1987). SANC00477, SANC00556, SANC00378, and SANC00430 have no recorded uses.

Table 3.4: Selected ligands binding to the allosteric sites and making substantial interactions with allosteric residues identified by Penkler *et al.,* 2017 in B_prime.

Ligands	BE (kCal/mol)	Hydrogen bond interactions	Other interactions
SANC00284	-7.1	R76, N98, D100, S505	D97, M404, K452, S453, L454, G455, A503, S504
SANC00477	-8.4	G74, R75, D100, G506	L72, I73, G75, A101, W102, A111, P112, P113, A149, Q150, A153, K452, S505, N508, E509, D510, E511
SANC00556	-7.3	R76, N98, A503	D97, D100, M404, H439, K452, S453, L454, G455, S504, S505
SANC00676	-6.6	D224, R445	Y145, N222, G223, L227, R315, E318, P319, V322, K414, E444, K446, R447
SANC00378	-6.9	Y145	N222, G223, D224, L227, R315, E318, P319, V322, K414, E444, R445, K446, R447
SANC00430	-7.4	Y145, N222, D224, R315, R445, K446	G223, T225, H226, L227, S316, E318, P319, V322, K414, E444, R447

Other interactions include Van der Waals, Alkyl, Pi-Alkyl, Pi-Cation, Pi-Sigma, Pi-Pi T-shaped, Pi-Pi Stacked interactions. Indicated in red are residues identified by Penkler *et al.,* 2017 using PRS in combination with MD simulation.
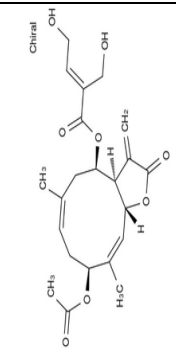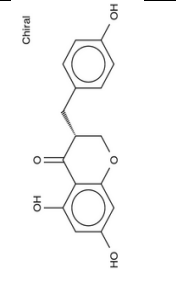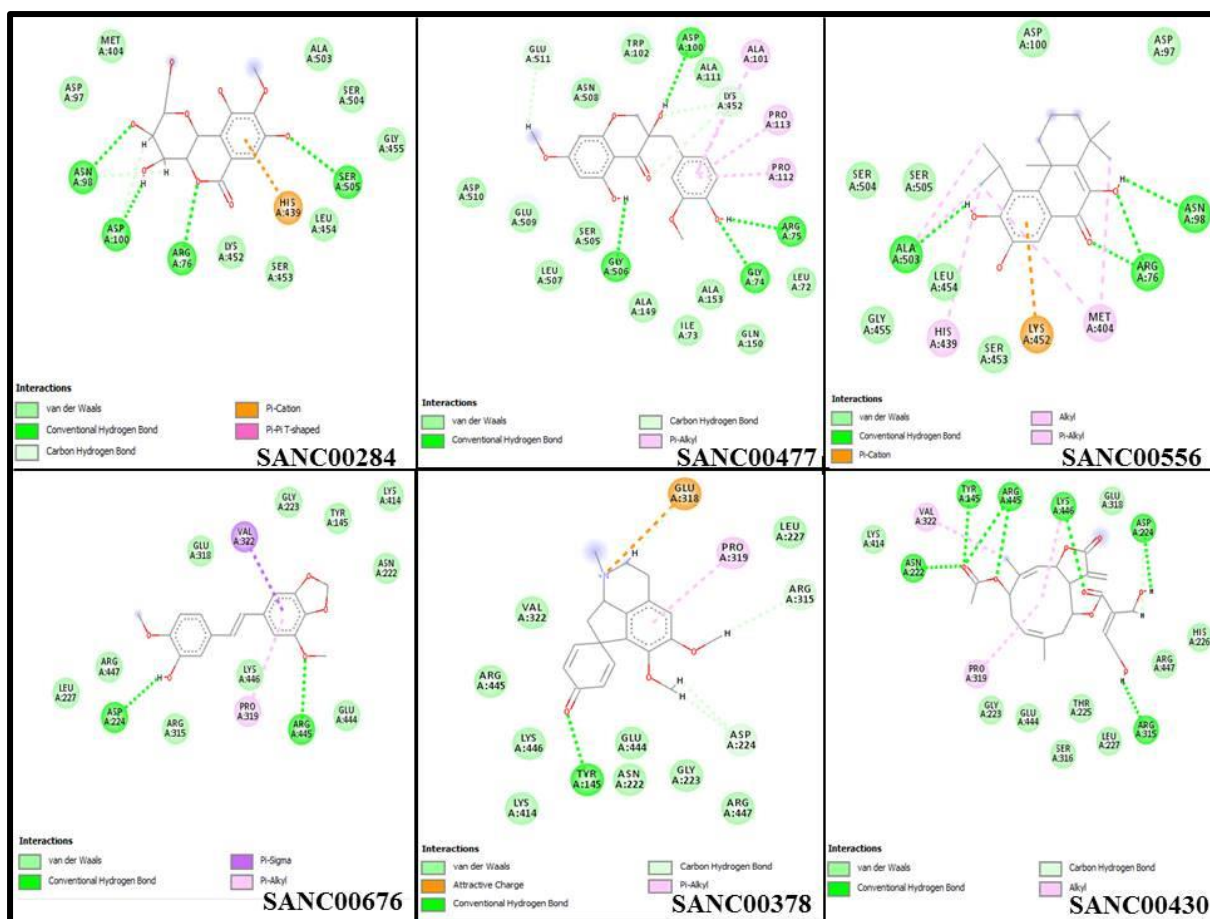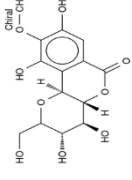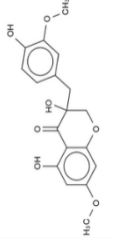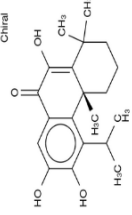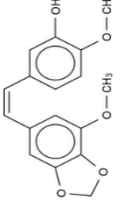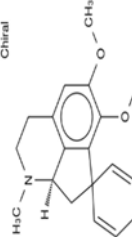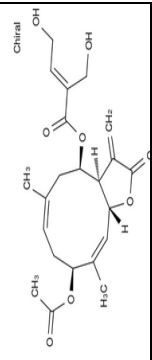
Figure 3.9: Protein-ligand interactions in B_prime as listed in the table above using Discovery Studio Visualizer.

Table 3.5: Characteristics of ligands selected in B_prime.

Ligand	Compound Name	2D Structure	Source organisms	Classifications	Uses	References
SANC00284	Bergenin		*Peltophorum africanum*	Glycoside	Antiviral	(Bessong *et al.,* 2005)
SANC00477	3,5-Dihydroxy-7-methoxy-3-(4'-hydroxy-3'-methoxybenzyl)chroman-4-one		*Albuca fastigiata*	Flavonoid, Homoisoflavonoid	None recorded	(Koorbanally *et al.,* 2005) (Du Toit *et al.,* 2005)
SANC00556	6,12,13-Trihydroxychina-5,8,11,13-tetraen-7-one		*Harpagophytum procumbens*	Terpenoid, Triterpene	None recorded	(Clarkson *et al.,* 2006)
SANC00676	Combretastatin A-2		*Combretum caffrum*	Unclassified	Murine P388 lymphocytic leukemia (PS) cell-growth inhibition, Tubulin inhibitory activity	(Singh and Pettit, 1987)
SANC00378	Pronuciferine		*Cissampelos capensis*	Alkaloid, Aporphine	None recorded	(De Wet *et al.,* 2011)

| SANC00430 | Eucannabinolide | | *Schkuhria pinnata* | Sesquiterpene, Terpenoid | Anticancer activity | (Fouche *et al.,* 2008) |

3.4 Conclusion

The use of computational tools in drug discovery can speed up the challenging design and optimisation process for a new drug candidate. Due to the rapid development of faster architectures and better algorithms for high - level computations, the impact of computer structure-based drug design (SBDD) on drug discovery has increased in the last decade (De Vivo *et al.,* 2016). In this chapter, 623 small molecules from SANCDB were screened against both conformations of *E. coli* Hsp70 with the aim of searching for allosteric modulators. Potential allosteric ligand binding sites were identified in Chapter two using three different allosteric prediction tools and consensus sites were identified among these tools. Consensus sites whose surrounding residues where in agreement with allosteric residues identified by Penkler *et al.,* 2017 were isolated and ligands binding to these sites were isolated for inhibitory designs. Ligands binding to H_prime displayed good binding energies ranging from -12.5 to -5.2 kcal/mol. Ligands binding to B_prime had binding energies ranging from -10.6 to -4.5kcal/mol. A heat map was plotted to show overall binding energies of the compound when bound to the respective protein structures. Seven ligands binding to the consensus sites identified in H_prime was isolated. These ligands conformed to the criteria which include binding within 10Å of the identified sites, overlapping with PRS data and with binding energies of $\geq$ -6kcal/mol. Selection of ligands in B_prime was quite difficult as most ligands scattered on the surface of the protein structure. However, six ligands were finally selected, and these ligands made reasonable interactions with identified allosteric residues. These ligands were taken to molecular dynamics simulation to assess for ligand stability complex.

CHAPTER FOUR: MOLECULAR DYNAMICS SIMULATION

4.1 Chapter overview

The knowledge of the interactions between the protein-ligand complexes is central to many biological processes as well as modern drug development processes (Arcon *et al.,* 2017). Protein activity and interactions are dependent on their structural stability, dynamics, and flexibility at their atomic level (Hospital *et al.,*2015). In this chapter, all-atom molecular dynamics (MD) simulations will be performed to examine the stability and folding behaviour of the protein-ligand complexes and also investigate the structure, dynamics, and thermodynamics of these complexes. GROMACSv1.5.1 simulation tool will be used. Each protein conformation, including inhibitor-free and inhibitor-bound systems, will be subjected to energy minimization, equilibration of the constant number of particles, Volume, and Temperature (NVT), and constant number of particles, Pressure, and Temperature (NPT), and 100ns production runs. In total, 15 MD simulations will be performed, 8 simulations for H_prime and 7 simulations for B_prime. Importantly endogenous ligands will be included in all our systems to retain the allosteric intermediate state, which, based on the functional mechanism, only exists in the presence of bound ATP. Using Gromacs tools *gmx rmsd, rmsf, gyrate,* trajectories will be analyzed, and conclusions will be drawn. Further, the free energy of binding of the ligands to the protein structures will be estimated using the MMPBSA tool in GROMACS to assess the robustness of protein-ligand associations.

4.1.1 Molecular dynamics simulation

Protein structures undergo conformational changes due to their dynamic nature and the study of these changes can provide vital information in drug design (Teague, 2003). Dynamic occurrences at the molecular-level can play a key role in controlling processes that affect biomolecular functional properties. Several experimental and computational techniques have been created to understand the conformational dynamics of protein structures (Srivastava *et al.,* 2018). Computational techniques are very common in the pharmaceutical industry and used to test the properties of a molecule without synthesizing it. This method is less expensive than experimental techniques and has provided spatial and temporal resolution that experimental methods do not provide (Abraham *et al.,* 2015). Molecular dynamics (MD) simulations is a computational technique in which classical equations of motion (Newton's

equations) are solved based on the force between atoms in the initial configuration to find the next configuration. The laws of Newton tell us how these forces affect the movement of atoms (F = ma, where F is the force applied on the atom, m is the mass of the atom, and a is the acceleration of the atom). MD calculates the time-dependent behaviour and movement of a molecular system (Llorach-Pares *et al.,* 2017). Simply put, MD assesses the movements of atoms due to their interactions. These interactions could be as a result of bond extension, angle bending and dihedral torsions of the bonded atoms. Non-bonded atoms interactions are as a result of the interactions between Van der Waals and electrostatic forces. In MD, atoms move at a constant speed between perfectly elastic collisions and the dynamic problem can be solved without making approximations within the limits imposed by the precision of the machine. It provides quantitative and/or qualitative information on the fluctuations and conformational changes of large macromolecules such as proteins, nucleic acids (DNA, RNA) and membranes. Thermodynamics depicts the driving force behind chemical processes and kinetics describes the mechanism through which these chemical processes are carried out (Win & Budhraja, 2007). The enormous application potential of MD simulation in understanding the dynamics of protein structures has led to implementations of MD in many software packages, including GROMACS (Berendsen *et al.,* 1995), AMBER (D.A. Case *et al.,* 2018), NAMD (Phillips *et al.,* 2005), CHARMM (Brooks *et al.,* 1983), LAMMPS (Plimpton, 1995) and Desmond (Bowers *et al.,* 2006). This different software has distinct features. For this study, we utilize GROMACS v1.5.1 simulation tool. The force field is a collection of equations that describe the time evolution of bond lengths, bond angles, and torsions, including the non-bonding Van der Waals and electrostatic interactions between atoms. It is designed to replicate molecular geometry and selected structural properties and also to calculate the system's total energy.

4.2. Methodology

4.2.1. Topology generation

4.2.1.1. Protein topology

Visualization of protein structure was done using Discovery studio visualizer. Protein topology file was generated using pdb2gmx program of the GROMACS 5.1.2 package employing parameters from the GROMOS96 43a1 force field. This file contains all the information necessary to define the molecule within a simulation such as non-bonded parameters (atom types and charges) and bonded parameters (bonds, angles and dihedrals). A

compatible configuration (.gro) file that contains the coordinates of the molecule in the system and its corresponding parameter file (.itp) which shows the connection of atoms in the molecules were also generated.

4.2.1.2. Ligand topology

Ligand preparation is one of the most difficult MD simulation tasks. Ligand topology file was prepared with an external tool that is compatible with the force field chosen in the protein topology generation. This tool PRODRG 2.5 (http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg) is an automated server for ligand topology generation (Schüttelkopf and Van Aalten, 2004). Each molecule in pdb format was uploaded to the server. A compatible configuration (.gro) file that contains the coordinates of the molecule in the system and its corresponding parameter file (.itp) which shows the connection of atoms in the molecules were also generated.

4.2.2. Definition of box and Solvation

Once the protein and ligand topology files have been generated, they are merged as a single file to obtain the final starting structure and topology file for each protein-ligand complex. All MD simulations were carried out under periodic border conditions (PBC) using a triclinic box to avoid problems with boundary effects caused by the finite size and to create an infinite simulation environment. A box dimension of 1.75nm was used in order to ensure that the box was large enough to hold each system. This places the protein 1.75nm from the edge of the box. The systems were solvated using SPC water model (spc216). Solvation of the system ensures easy movement and interactions between the protein and ligand. VMD was used to visualize the shape and size of the solvent box surrounding the whole protein. Selection of an appropriate box is important as a wrong box has an impact on the simulation results.

4.2.3. Addition of ions

The solvated system now contains a charged protein; therefore, a certain number of Na^+ (Sodium) and Cl^- (Chloride) ions were added to neutralize the system. The addition of ions is dependent on the total net charge of the system.

4.2.4. Energy minimization

The aim of minimizing the system is to ensure that the system has no steric clashes that resulted from the addition of ions and water molecules or inappropriate geometry. This is

done using the GROMACS MD engine mdrun. A successful energy minimization generates the potential energy and the maximum force.

4.2.5. Equilibration

The solvent and the ions surrounding the protein were equilibrated to stabilize the system. The temperature and pressure of the system were equilibrated using the typical canonical NVT at 293K and the isothermal-isobaric NPT ensemble for 100ps.

4.2.6. Production MD

Once the equilibration processes have been completed, the system is now equilibrated at the desired temperature and pressure. The next step is to release the position restraints and run production MD. The equilibrated system was then subjected to the production MD for 100ns.

4.2.7. Analysis

Now that our protein has been simulated, we should carry out some system analysis. The first analysis is the MD trajectory analysis (trjconv) and this is a post-processing tool used to remove the periodic boundary conditions generated during the simulation and also to center the system in the box for further analysis. The second analysis is the root mean square deviation (RMSD) and this analysis is done to visualize the structural stability of the protein. The third analysis is the root mean fluctuation which is a measure of the local chain flexibility. The final analysis is calculating the radius of gyration (Rg) and this is a measure of the structural compactness. Each of the analyses excluding the first was calculated for both structures and plotted using R studio (https://www.rstudio.com/). Interactions between each ligand and corresponding protein structure were visualized using Discovery studio visualizer and PyMOL.
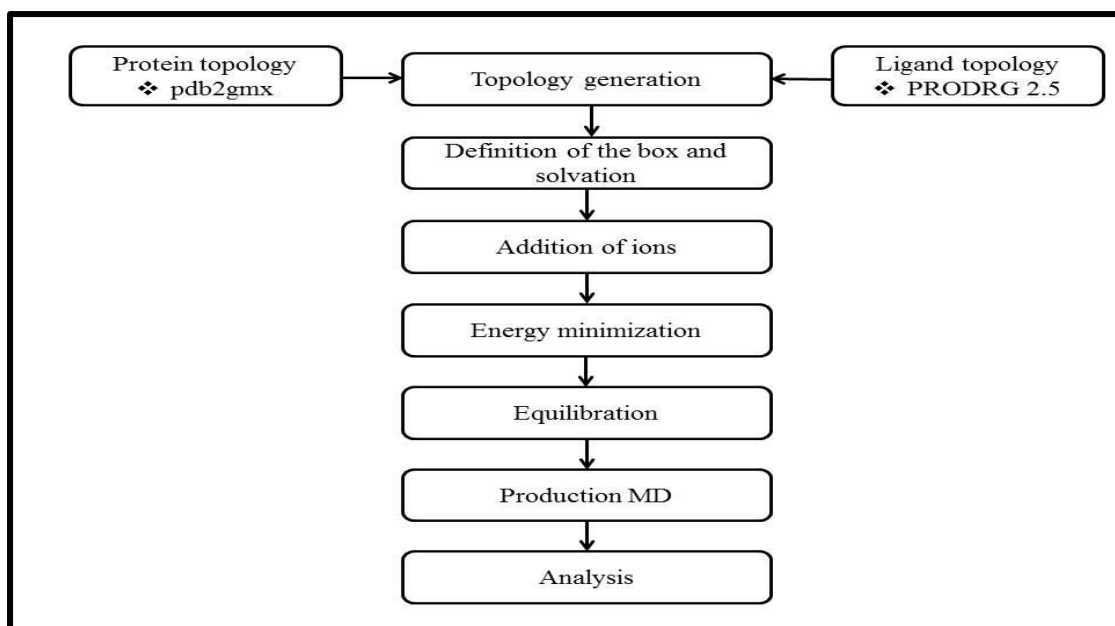
Figure 4.1: Summary of MD work-flow.

4.3. Results and Discussion

MD simulations were performed to probe the conformational dynamics of both structures in the Apo state and when bound to their respective ligands at atomic resolution. The Apo structures were simulated in complex with its co-factor ATP and Mg^{2+} without any ligand. Structural mobility analysis in MD plays an important role in data interpretation, especially in the simulation of biomolecules. MD simulation results were analyzed in terms of their backbone root mean square deviation (RMSD), root mean square fluctuation (RMSF) and radius of gyration (Rg). The RMSD and RMSF of the structures are the most common mobility measures calculated from simulations (Martínez, 2015). In this chapter, a total of 15 MD simulations will be performed, 8 simulations for H_prime and 7 simulations for B_prime. This also includes simulating the protein structures (Apo) without any ligand. Importantly endogenous ligands will be included in all our systems as they are important in the functionality of the structures. Each simulation was submitted to MD simulation of 100ns.

4.3.1. H_prime

4.3.1.1. RMSD

The root mean square deviation (RMSD) in MD simulation is an invariably used measurement of the differences between two set of values, and in this context between the ligand-free (Apo) and ligand-bound. It is often used to determine the structural stability of a structure (ligand-bound) in the time scale of the simulations and to discern its differences from the initial coordinates (Apo). 100ns MD simulations of the Apo and respective ligands

bound were conducted to provide a direct comparison of conformational deviation. Figure 4.2 shows the deviation of the protein structure when bound to the respective ligands. One thing that is common is the stability of the complexes at the beginning till the end of the simulation. The various protein-ligand complexes converged early evenly which indicates a stable system with exception to SANC00357 and SANC00492. SANC00357 had the largest RMSD value. Backbone RMSD analysis of the protein structure, when bound to its respective ligand converged at different values. When bound to SANC00499, SANC00685, SANC00340, SANC00378, and SANC00430, the RMSD of the protein backbone converged at around 0.30nm at the beginning of the simulation with the highest convergence at 0.40nm as seen in SANC00685 and 0.420nm as seen in SANC00499. SANC00492, when bound to the structure converged at around 0.30nm during the first 30ns of simulation, but as seen in Figure 4.2, there was a shift to 0.50nm which is visible after 35ns of simulation. When in complex with SANC00357, there was a noticeable difference between the protein-free and protein-bound complexes as the average RMSD value of the complex is 0.40nm while the average RMSD value of the ligand-free system is 0.23nm.
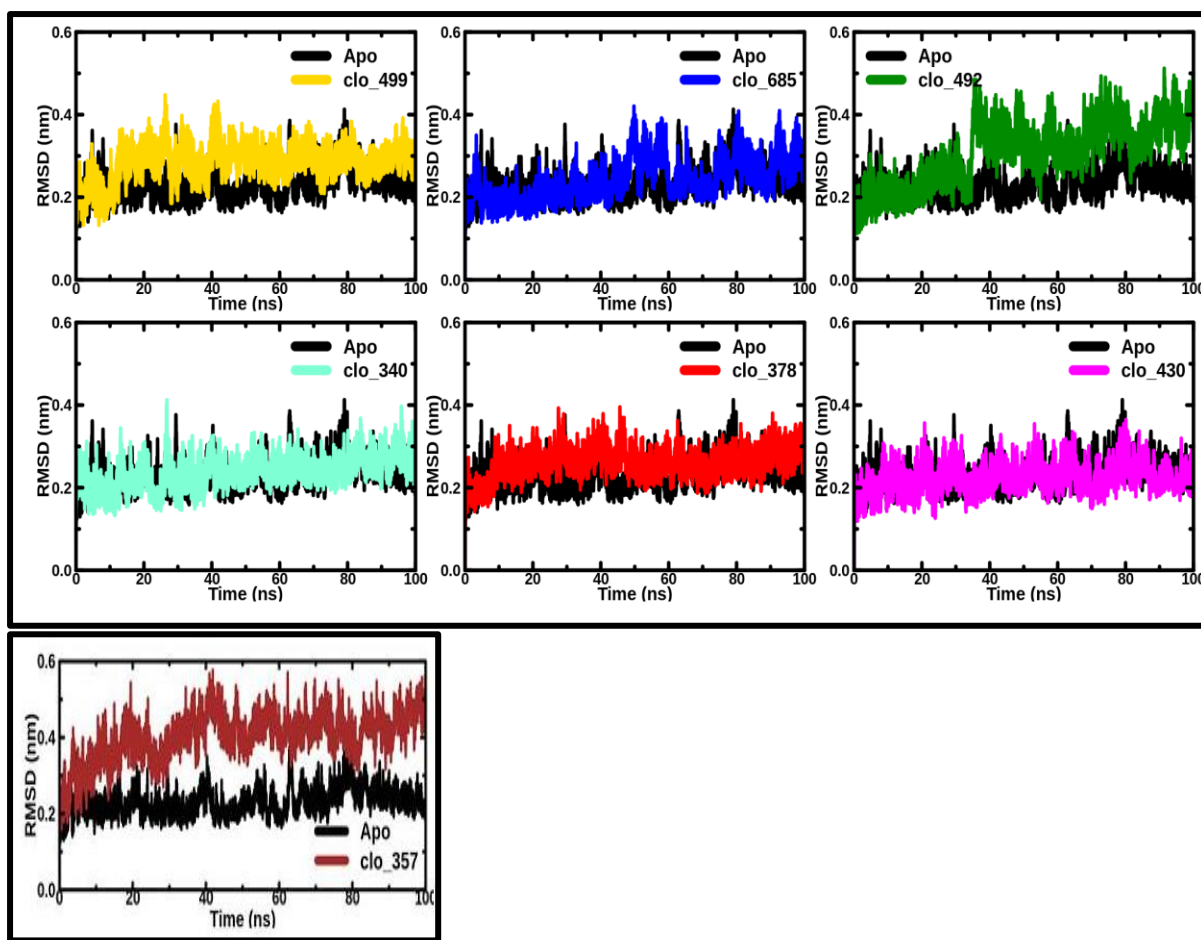
Figure 4.2: RMSD plot of ligands bound to H_prime.

4.3.1.2. RMSF

The root mean square fluctuation (RMSF) is a measure of the local chain flexibility. RMSF measures of the displacement of a certain atom or group of atoms over the number of atoms in relation to the reference structure (Martínez, 2015). Residues in protein-ligand complexes displayed higher RMS fluctuation values as compared to the ligand-free complex which indicates that ligand binding affects the structure. Conformational fluctuations in the NBD subdomains were observed with particularly large movement in the subdomains IIA and IIB (residues 250 – 300) in all ligand-bound state (Figure 4.3). Protein-SANC00499 displayed conformational fluctuations in the subdomain IIA (residues 300 – 340). Protein-SANC00685 displayed conformational fluctuations in the subdomain IIB (residues 230 – 260). Protein-SANC00357 also displayed some conformational fluctuations in the subdomains IA and SBDβ (residues 370 – 400).
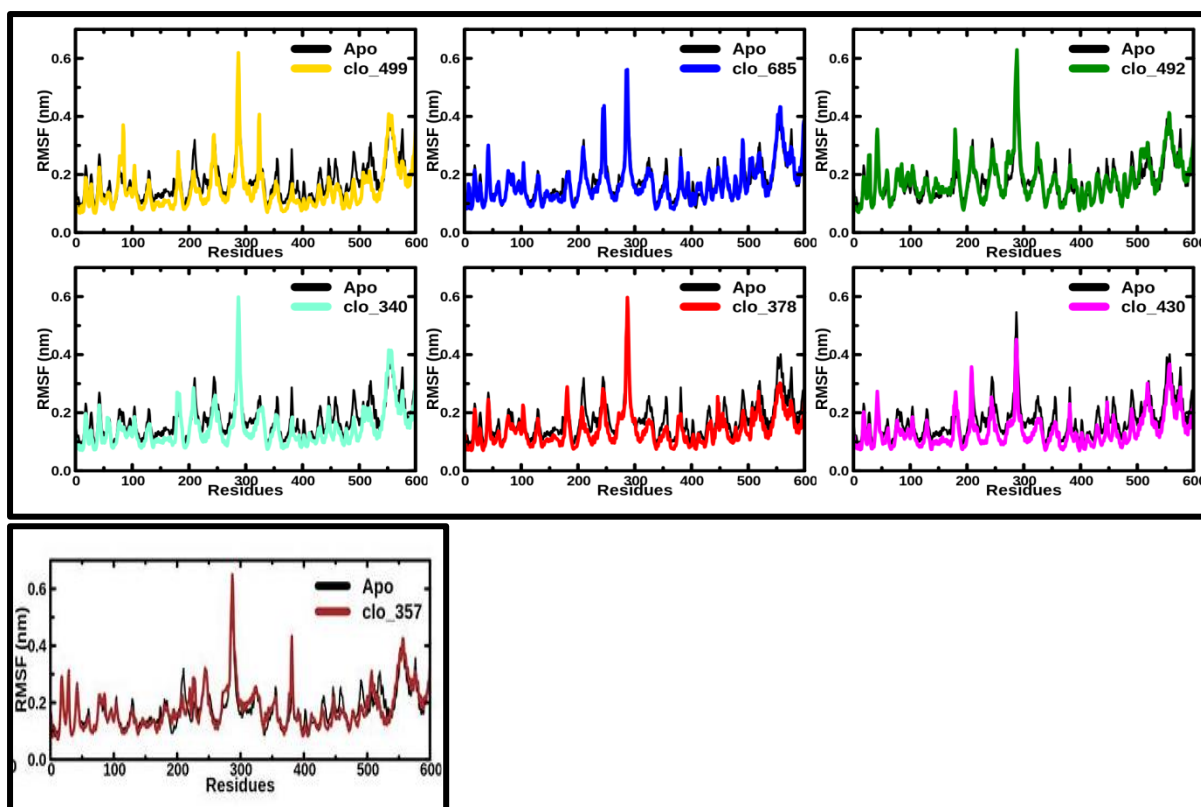
Figure 4.3: RMSF plot of ligands bound to H_prime.

4.3.1.3. The Radius of gyration (Rg)

The overall shape of the entire protein structures in its ligand-free (Apo) and ligand-bound states have been monitored for each MD run by calculating the corresponding radius of gyration. Rg is defined as the root mean distance from the atom collection from its common center of gravity and this describes the overall spread of a molecule. As shown in Figure4.4, the Rg values for the protein structures when bound to its respective ligands was 2.9±0.1nm. Rg analysis shows a decrease in value when the structure is in complex with its respective ligand indicating that the various ligands induce a more relaxed state when bound to the H_prime.
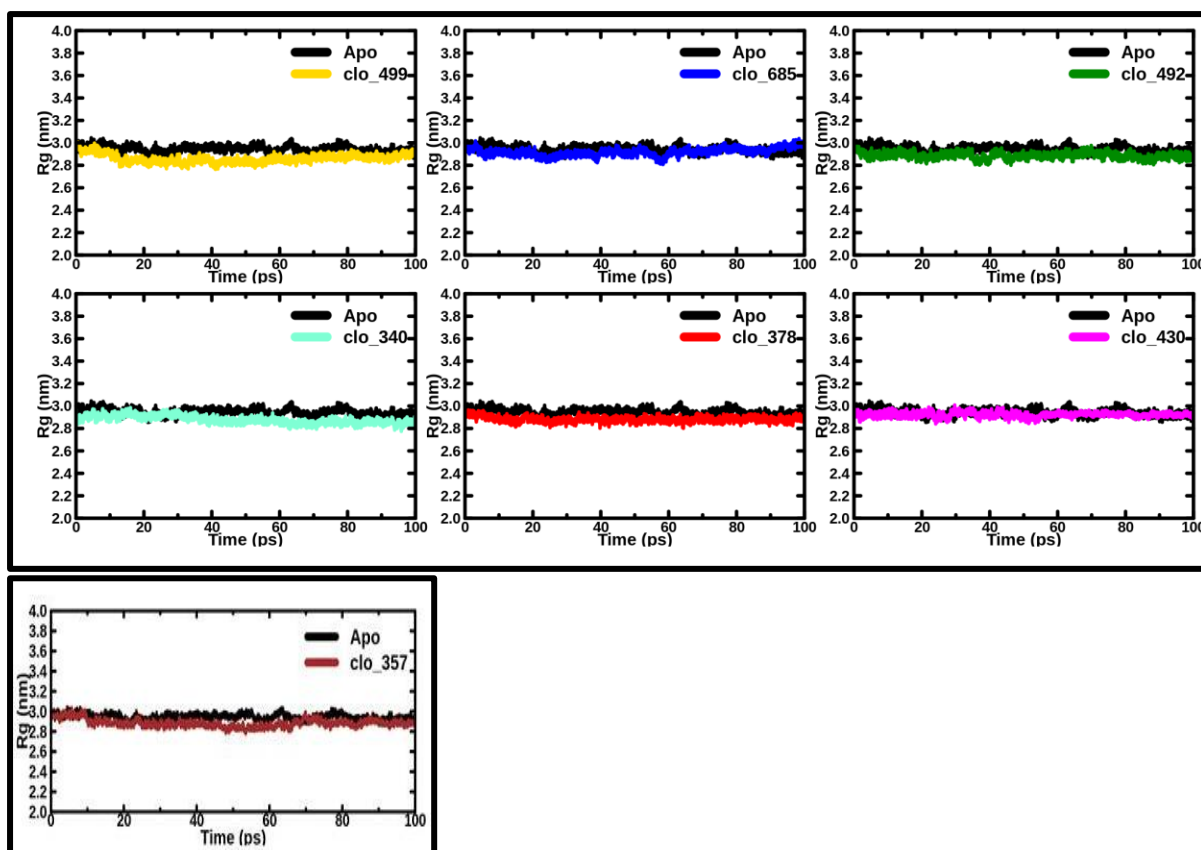
Figure 4.4: Rg plot of ligands bound to H_prime.

VMD was used in visualizing the trajectories after MD simulation. Binding of some of the ligands to the allosteric site triggered some form of structural rearrangements in the various subdomains in H_prime. Video S1 shows the movement of H_prime in its ligand-free states while Video S2, S3, S4, and S5 show the movement of H_prime when bound to SANC00499, SANC00685, SANC00492, and SANC00357 respectively. When bound to SANC00499 and SANC00685, the alpha-helical lid (SBDα) appears to block the SBDβ. SANC00357 triggers a slight docking of the SBD to the NBD. When bound to SANC00492, structural rearrangements were observed in the SBD, most especially the SBDα. Also, the subdomains IA and IB which make up the ATP binding cleft appears to close up when bound to SANC00492 as opposed to its open nature in the ligand-free state.

4.3.2. B_prime

4.3.2.1. RMSD

The RMSD is useful in determining the global stability of the protein-ligand complex over the entire simulation period. It is also used in analysing time-dependent structure motions. As seen in Figure 4.5, when bound to its respective ligands, each of the protein structure showed

a significant deviation from its initial structure (Apo). Ligand-free structure (Apo) appears to be very unstable as the RMSD flips throughout the simulation and converged towards the end of the simulation. It was stable during the first 40ns and then jumps at about 45ns with RMSD of 0.57nm and finally converged at 0.5nm at about 75ns. This indicates the highly flexible nature of the structure. The inclusion of the ligands seems to improve the stability of the structure as seen in Figure 4.5. The various ligands showed different RMSD values, the largest being SANC00284 (gold). Open_430 (blue) and Open_477 (red) showed the most stability. The RMSD of the protein structure when bound to SANC00477 (red), SANC00556 (magenta), SANC00676 (aquamarine) and SANC00430 (blue) converged almost evenly and early. When bound to SANC00284, the RMSD converged early to 0.25nm, but a jump was observed after 10ns to about 0.41nm. SANC00378 (green 4) showed the most irregular pattern of stability as the RMSD showed different shifts in its values.
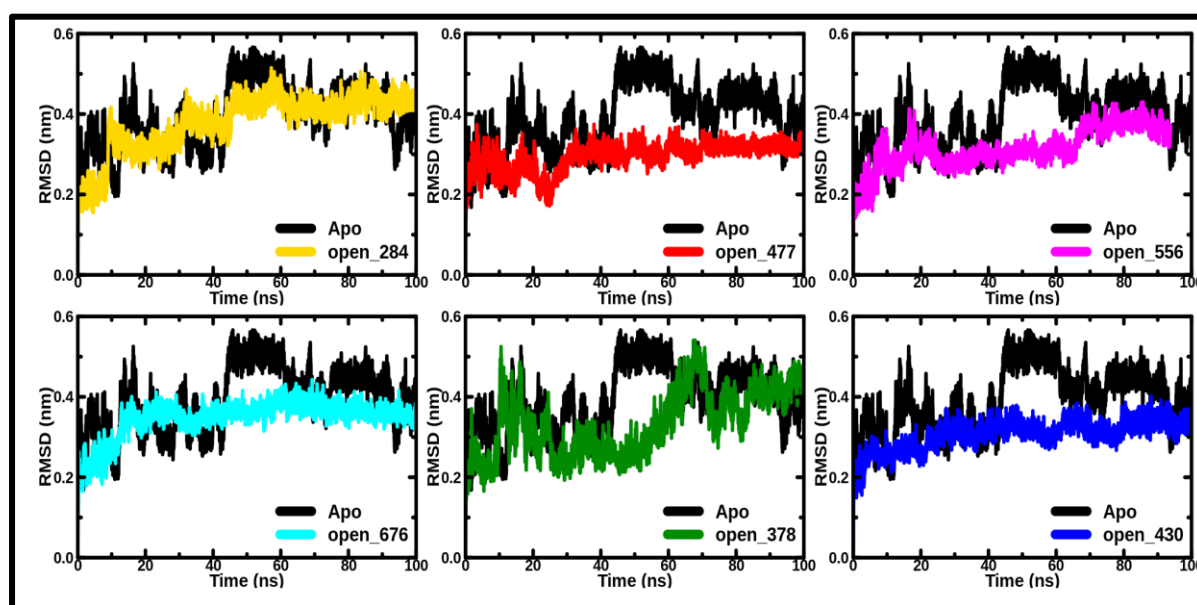
Figure 4.5: RMSD plot of ligands bound to B_prime.

4.3.2.2. RMSF

RMSF was calculated to understand protein flexibility. Conformational fluctuations in the SBD subdomains were observed with particularly large movement in the subdomains SBDα also known as the alpha-helical lid in all ligand-bound state except when bound to SANC00430. The protein structure, when bound to SANC00378 showed the most fluctuations in the SBDα subdomain. When bound to SANC00676 and SANC00430, conformational fluctuations were observed in the hydrophobic linker and SBDβ also known as the Beta substrate binding domain.
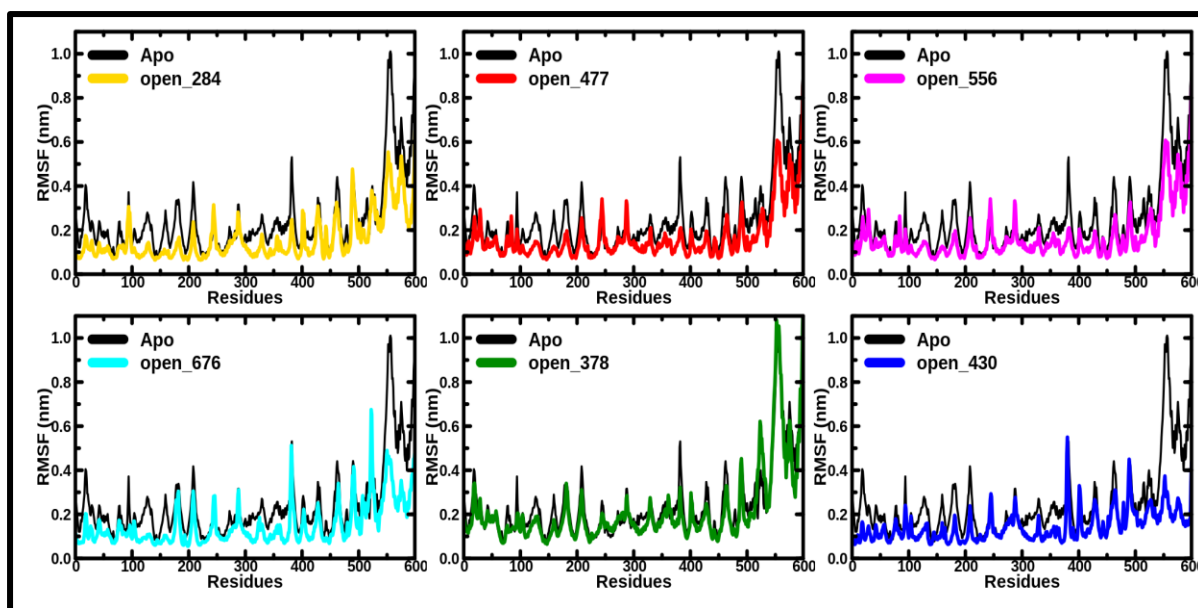
Figure 4.6: RMSF plot of ligands bound to H_prime.

4.3.2.3. Radius of gyration

The system, when bound to its respective ligand remained compact throughout the entire simulation as shown in the Rg plot (Figure 4.7). SANC00284, SANC00676, and SANC00430, when bound to the protein structure had very similar Rg values. When bound to all compounds except SANC00378, the complex experiences a decrease in Rg value as compared to its ligand-free state. This suggests a more relaxed state.
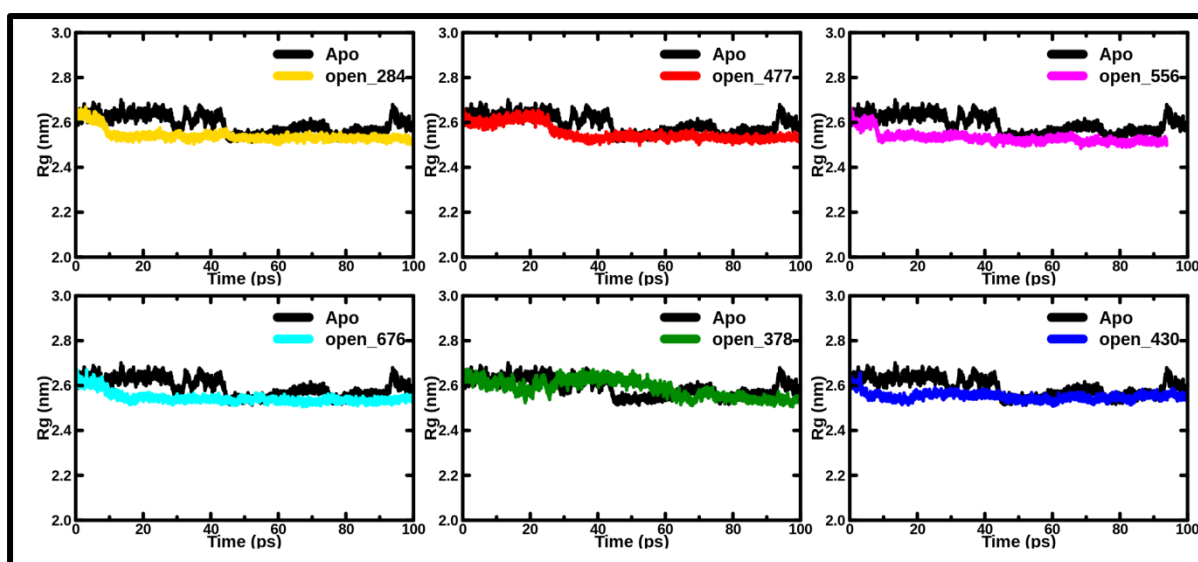
Figure 4.7: Rg plot of ligands bound to B_prime.

VMD was used in visualizing the trajectories after MD simulation. Binding of some of the ligands to the allosteric site triggered structural rearrangements in B_prime. Video S6 shows the movement of B_prime in its ligand-free states while Video S7, S8, S9, and 10 shows the movement of H_prime when bound to SANC00284, SANC00477, SANC00556, and SANC00676 respectively. In its ligand-free state, the SBDα and SBDβ are detached from each other and docked to different elements of the NBD (Kityk *et al.,* 2012). SANC00556 triggers some conformational changes when bound to the structure especially in the subdomain SBDα. The SBDα appears to undock itself from the subdomain IA and reattach itself to the SBDβ. SANC00477 and SANC00685 also triggered some structural changes in the subdomain SBDα.

4.3.3. Ligand stability

Assessing the stability of the binding pose of a ligand to a protein is important in evaluating the drug-likeness of the ligand. In order to determine the stability of the various ligands when bound to its respective protein structures, the RMSD of their Cα were plotted. Figures 4.8 and 4.9 showed the ligand stability in H_prime and B_prime respectively. All ligands remained stable throughout the entire simulation period, thus, this MD run was regarded as a qualified run.
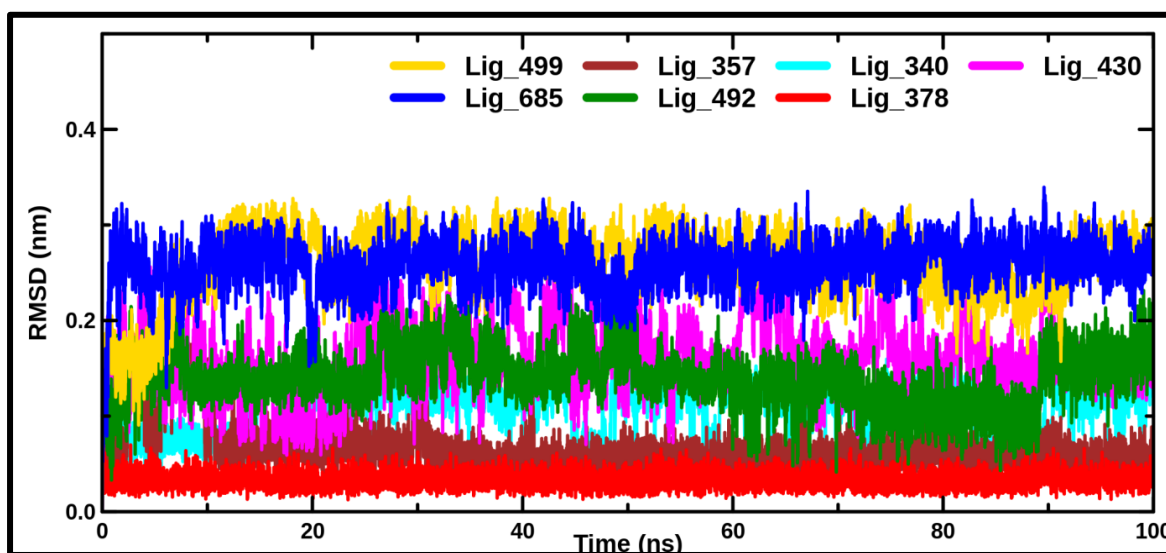
Figure 4.8: RMSD of ligand poses during MD simulations for H_prime. This shows the stability of the ligand when bound to H_prime
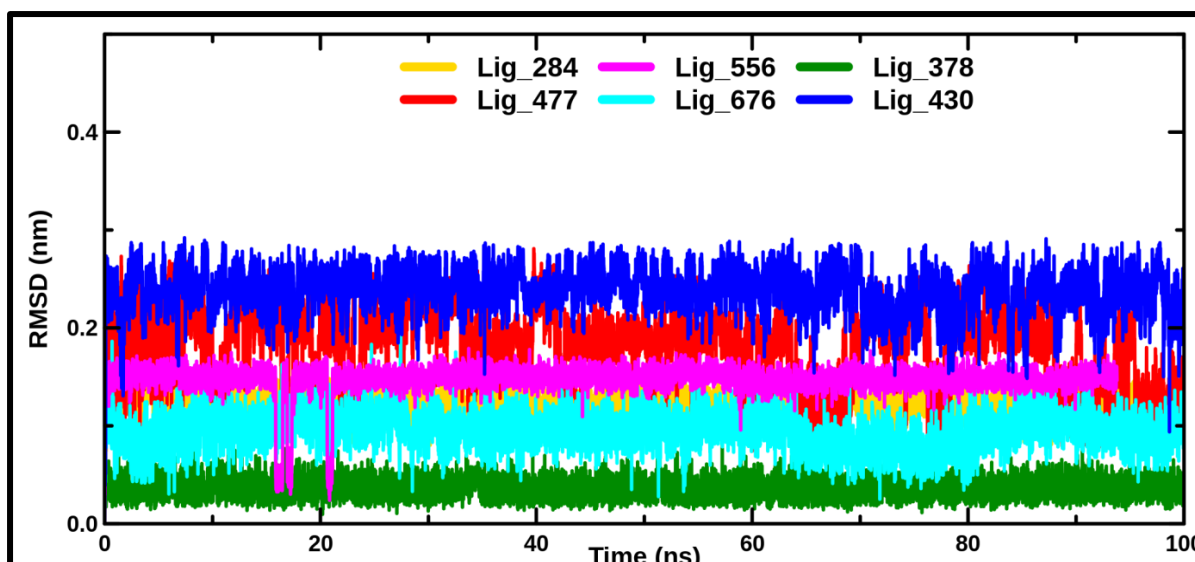
Figure 4.9: RMSD of ligand poses during MD simulations for B_prime. This shows the stability of the ligand when bound to B_prime.

4.3.4. MM-PBSA

The strength of a protein-ligand complex interaction involved in catalysis can be measured in terms of its binding free energy (Kumari *et al.,* 2014). Molecular Mechanics energy combined with the Poisson-Boltzmann and Surface Area continuum solvation (MM-PBSA) is a method for estimating the binding free energy of small molecules (ligands) to protein structures (Genheden and Ryde, 2015). Previous methods such as free energy perturbation (FEP) and thermodynamic integration (IT) have been explored in free energy calculations. Compared to these methods, MM-PBSA is less rigorous and more computationally highly efficient (Hou *et al.,* 2011). MM-PBSA is usually based on molecular dynamics simulations of the protein-ligand complex and is very successful in approximating the solvent contribution to the free energy (Wang *et al.,* 2018). MM-PBSA uses the Poisson-Boltzmann equation to calculate the Van der Waals, electrostatic, polar and non-polar solvation contributions to the binding free energy of the ligand to the protein structure.

In MM-PBSA, binding free energy (ΔG_{bind}) between a ligand and a protein structure to form a protein-ligand complex is calculated as (Wang *et al.,* 2018):

$$\Delta G_{bind,aq} = \Delta H - T\Delta S \approx \Delta E_{MM} + \Delta G_{bind,solv} - T\Delta S, \tag{1}$$

$$\Delta E_{MM} = \Delta E_{covalent} + \Delta E_{electrostatic} + \Delta E_{vdW}, \tag{2}$$

$$\Delta E_{covalent} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{torsion}, \tag{3}$$

$$\Delta G_{bind,solv} = \Delta G_{polar} + \Delta G_{non-polar}, \tag{4}$$

Where ΔE_{MM}, $\Delta G_{bind, solv}$ and $-T\Delta S$ represents a change in molecular mechanical energy during the gas phase, a change in solvent-free energy and a change in conformational entropy following binding, respectively. All of these changes are calculated by averaging the ensemble over a wide range of sampled conformations. ΔE_{MM} consists of three energetic terms calculated using molecular mechanics (MM): the covalent energy change ($\Delta E_{covalent}$), the electrostatic energy change ($\Delta E_{electrostatic}$), and the Van der Waals energy change (ΔE_{vdW}). $\Delta E_{covalent}$ comprises changes in the bond terms (ΔE_{bond}), the angle terms (ΔE_{angle}), and the torsion terms ($\Delta E_{torsion}$). The change in solvation free energy ($\Delta G_{bind, solv}$) is usually separated into polar and non-polar contributions (ΔG_{polar} and $\Delta G_{non-polar}$). The conformation entropy change $-T\Delta S$ is calculated on a set of snapshots taken from MD simulations by normal mode analysis (Wang *et al.,* 2018).

4.3.4.1. MM-PBSA calculations

MM-PBSA calculations were performed for 13 ligands bound to both conformational of the protein structure; 7 ligands bound to H_prime and 6 ligands bound to B_prime. GROMACS (*g_mmpbsa*) version 1.6 tool was used for this calculation. The last 15ns of the equilibrated MD trajectories were used to perform binding free energy calculations. *g_mmpbsa* tool calculates binding energy components using the MM-PBSA approach (Kumari *et al.,* 2014). This tool requires four input files; a trajectory file (*trr or xtc*) containing system snapshot during MD simulation, a topology file (*tpr*) containing all the information necessary to define the molecule within a simulation., an index file (*ndx*) containing the protein and ligand atoms in the system, and a file with solvation parameters (*mdp*) containing the necessary analytical conditions and controls. The complete *g_mmpbsa* tool includes many python scripts for the final statistical analysis using the energy terms obtained and also for the estimation of per-residue energy contributions in the protein-ligand complex (Kumari *et al.,* 2014).

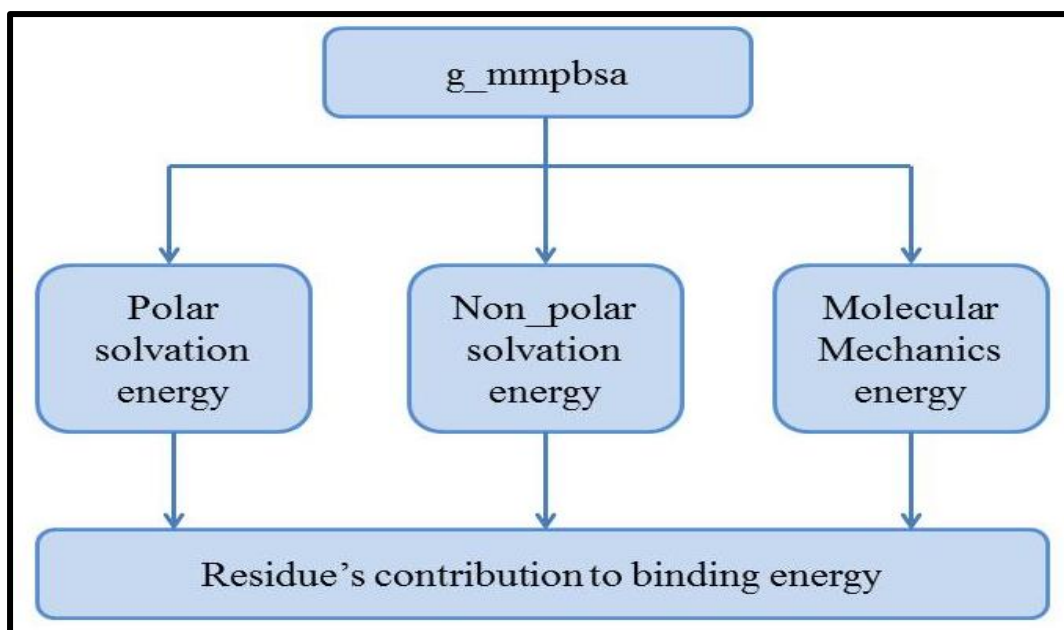**Figure 4.10:** MM-PBSA workflow utilized in binding free energy calculation.

4.3.4.2. MM-PBSA results of H_prime

In all the protein-ligand complexes studied, Van der Waals energy contributed significantly to the binding free energy of each compound except Cl_378 and Cl_430 which had positive values. A summary of the binding energy components is shown in Table 4.1. Figure S1 shows a bar plot of the binding energy components.

Table 4.1: Binding energy components (kJ/mol) obtained from GROMACS g_mmpbsa tool for the seven protein-ligand systems of H_prime.

Protein-ligand complex	ΔE_{vdW} (kJ/mol)	$\Delta E_{electrostatic}$ (kJ/mol)	ΔG_{polar} (kJ/mol)	$\Delta G_{non-polar}$ (kJ/mol)	$\Delta G_{binding}$ (kJ/mol)
Cl_499	-252.976±0.397	-4.944±0.122	79.342±0.291	-20.697±0.029	-199.276±0.373
Cl_685	-201.095±0.434	5.427±0.173	39.005±0.360	-17.034±0.038	-173.713±0.391
Cl_492	-169.589±0.369	-14.757±0.082	73.447±0.290	-13.183±0.022	-124.094±0.303
Cl_340	-126.142±0.289	-3.060±0.073	37.252±0.227	-10.428±0.026	-102.391±0.309
Cl_378	-0.000±0.000	0.803±0.053	4.366±0.960	-0.039±0.086	5.102±0.960
Cl_430	-0.000±0.000	0.803±0.025	4.372±0.956	-0.043±0.088	5.120±0.973
Cl_357	-197.281±0.210	-3.721±0.053	61.888±0.243	-14.274±0.019	-153.390±0.286

4.3.4.3. Per-residue contribution of H_prime

In order to determine the significant residues that influence the strength of the interactions in each protein-ligand complexes, the per-residue contribution was plotted for each protein-ligand complexes using *g_mmpbsa*. Protein-ligand complexes with positive binding free energies were excluded as this implies a false experiment. The per-residue contribution plots were done for protein-ligand complexes with negative binding energy ($\Delta G_{binding}$) (Figure S3). Figure 4.11, Figure 4.12 and Figure 4.13 show important residues contributing either positively or negatively to the binding free energies of the protein-ligand complexes.

Figure 4.11: A histogram showing the per-residue contribution of SANC00499 and SANC00685 when bound to H_prime.

Figure 4.12: A histogram showing the per-residue contribution of SANC00492 and SANC00340 when bound to H_prime.

Figure 4.13: A histogram showing the per-residue contribution of SANC00357 when bound to H_prime.

4.3.4.4. MM-PBSA results of B_prime

Van der Waals energy contributed significantly to the binding free energy of each compound except for Open_430 in all the protein-ligand complexes. Table 4.2 shows a summary of the binding energy components and Figure S2 shows a bar plot of the binding energy components.

Table 4.2: Binding energy components (kJ/mol) of B_prime obtained from GROMACS g_mmpbsa tool.

Protein-ligand complex	ΔE_{vdW} (kJ/mol)	$\Delta E_{electrostatic}$ (kJ/mol)	ΔG_{polar} (kJ/mol)	$\Delta G_{non-polar}$ (kJ/mol)	$\Delta G_{binding}$ (kJ/mol)
Open_284	-217.848±0.262	-14.491±0.124	65.352±0.205	-16.887±0.021	-183.862±0.299
Open_477	-223.546±0.261	-8.046±0.094	78.213±0.252	-16.707±0.20	-170.080±0.323
Open_556	-168.873±0.448	-1.737±0.290	62.628±0.619	-13.763±0.034	-121.718±0.508
Open_676	-134.384±0.448	-0.460±0.057	35.041±0.269	-12.797±0.027	-112.590±0.320
Open_378	-119.927±0.247	-18.381±0.126	49.811±0.478	-9.021±0.031	-97.497±0.493
Open_430	0.000±0.000	-0.856±0.009	61.720±2.219	-0.007±0.148	60.969±2.256

4.3.4.5. Per-residue contribution of B_prime

In addition, per residue contribution plot was performed using *g_mmpbsa* tool to determine the energy contribution of each protein residue that binds to the ligand. The per-residue contribution plots were done for protein-ligand complexes with negative binding energy ($\Delta G_{binding}$) (Figure S4). Figure 4.14, Figure 4.15, and Figure 4.16 show the residues contributing significantly to the binding free energy of the protein-ligand complex.
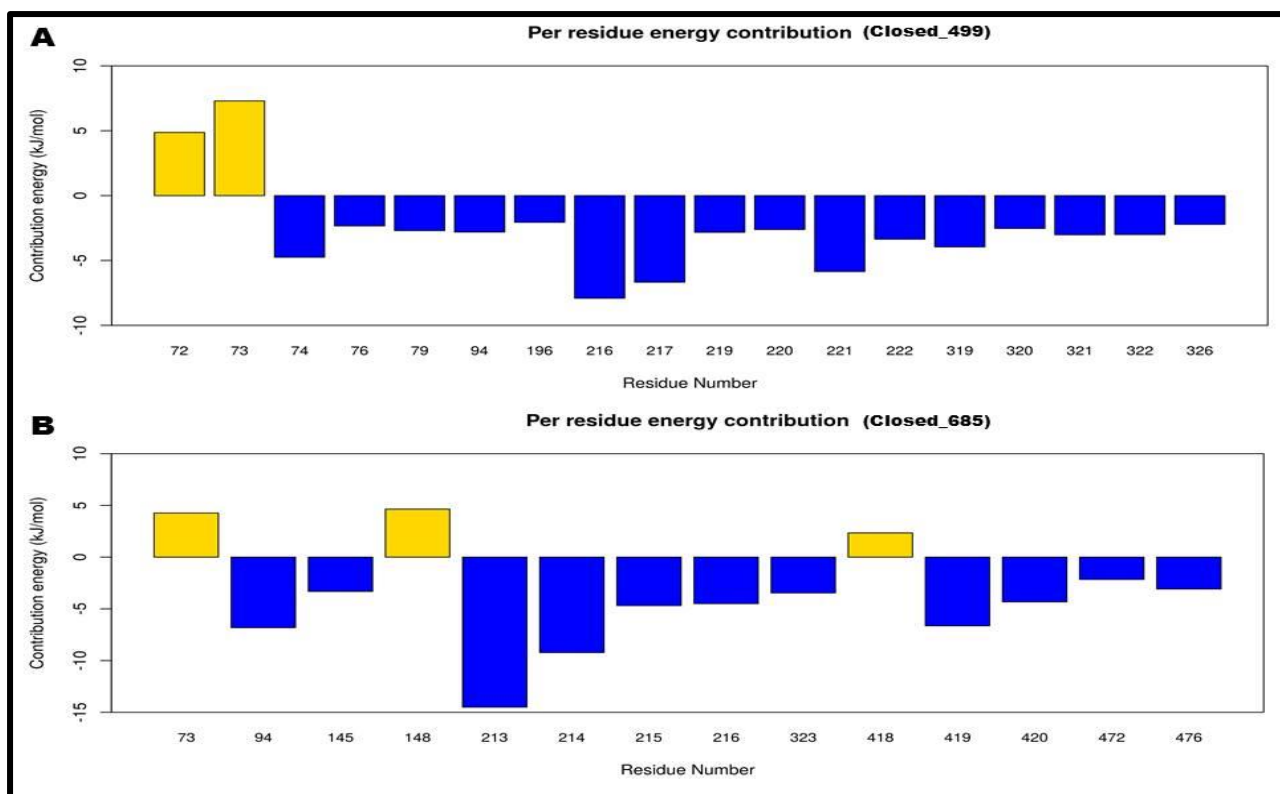
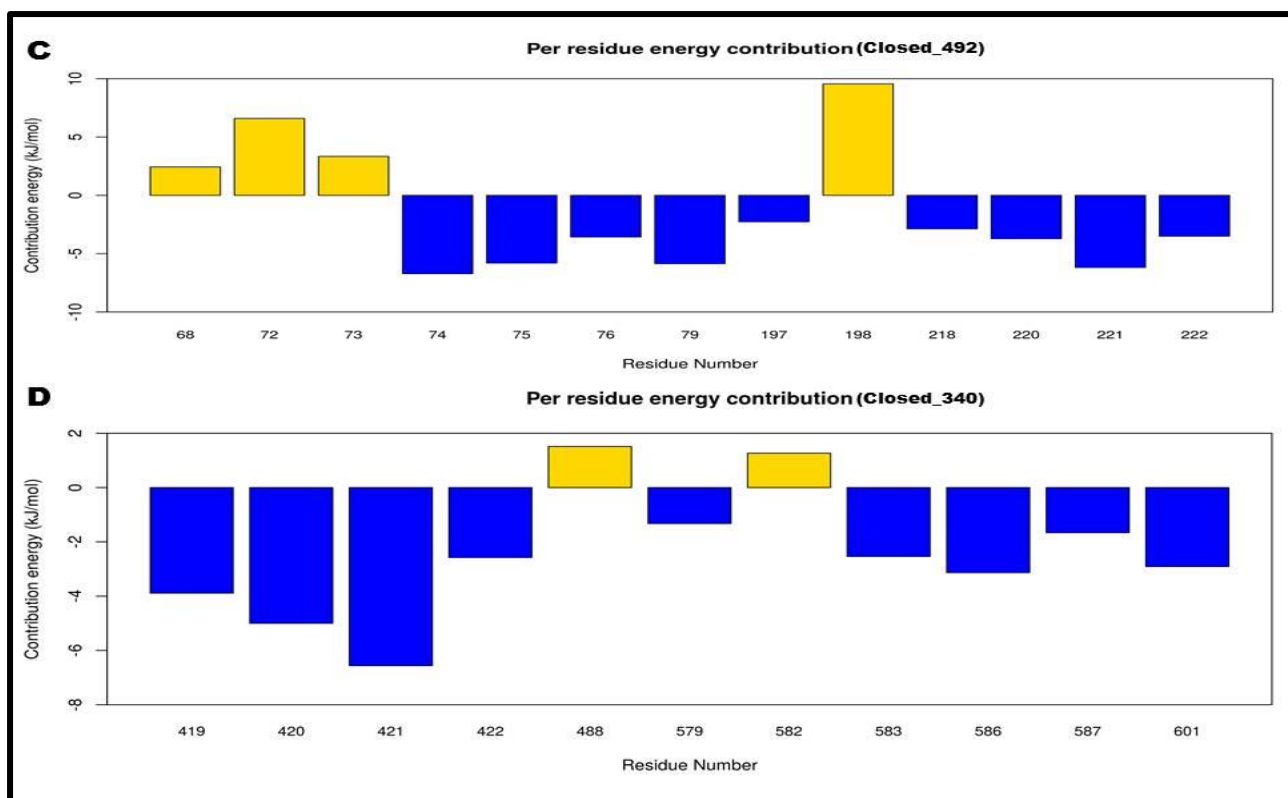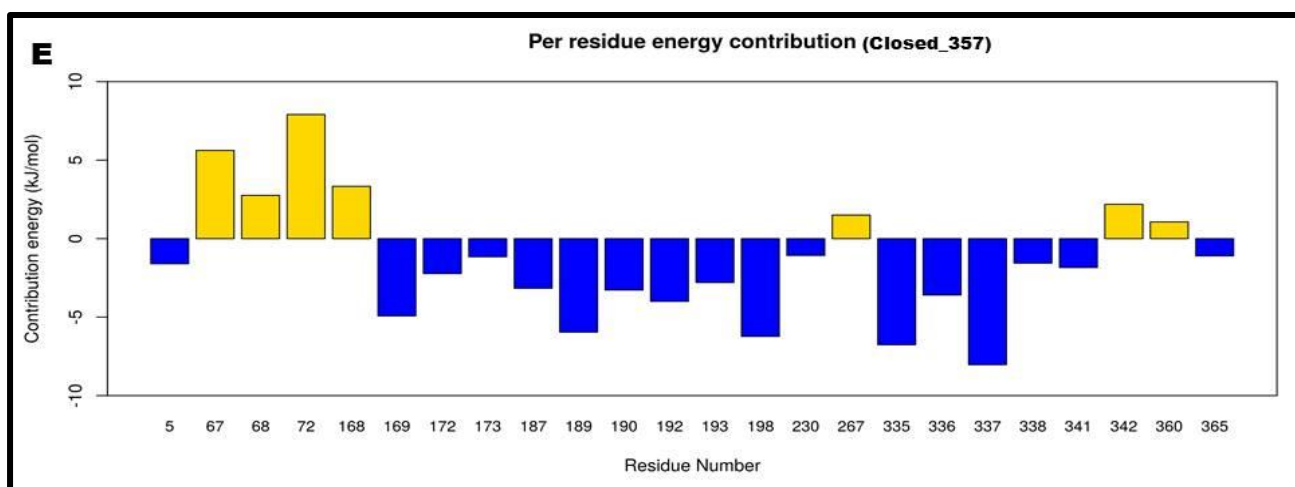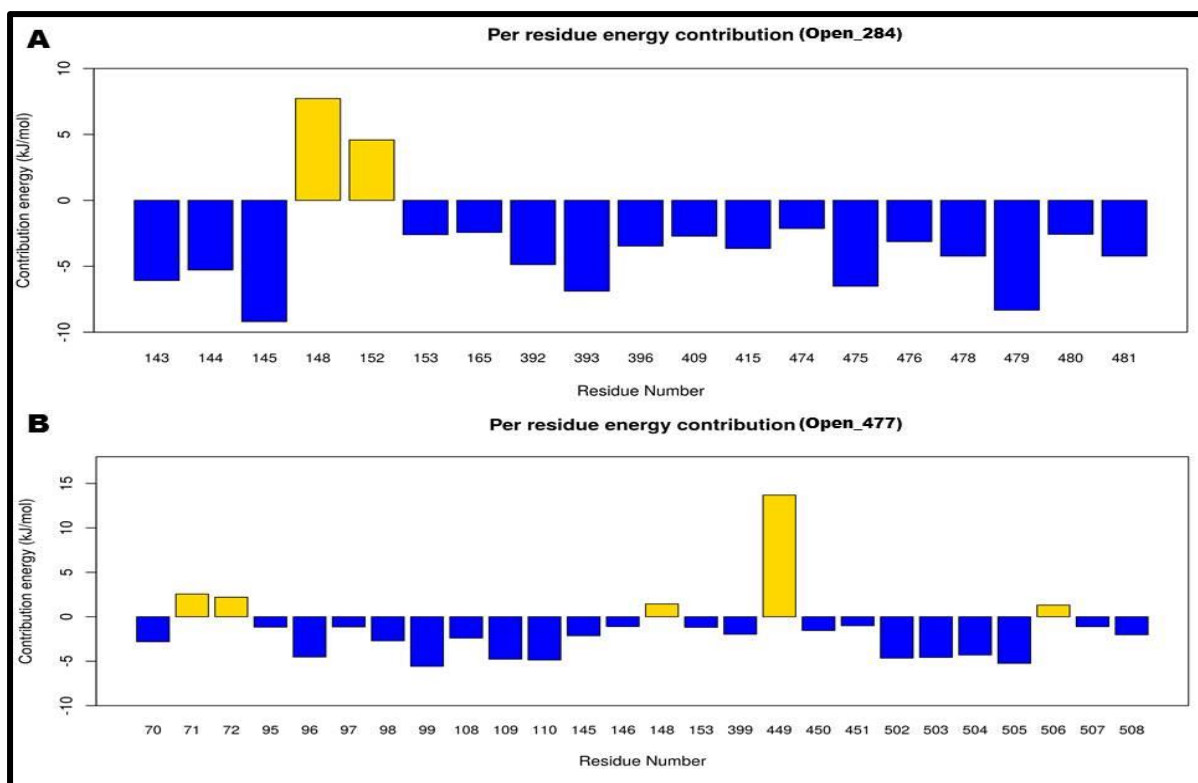**Figure 4.14:** A histogram showing the per-residue contribution of SANC00284 and SANC00477 when bound to B_prime.
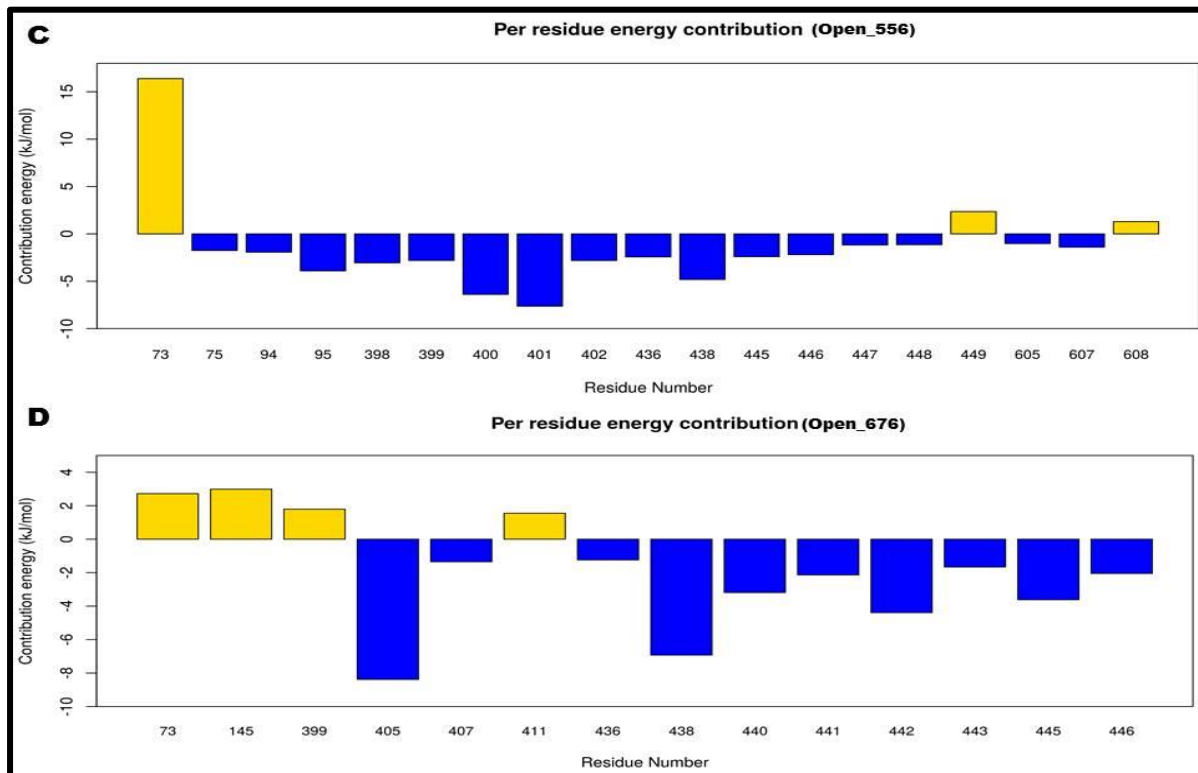

Figure 4.15: A histogram showing the per-residue contribution of SANC00556 and SANC00676 when bound to B_prime.
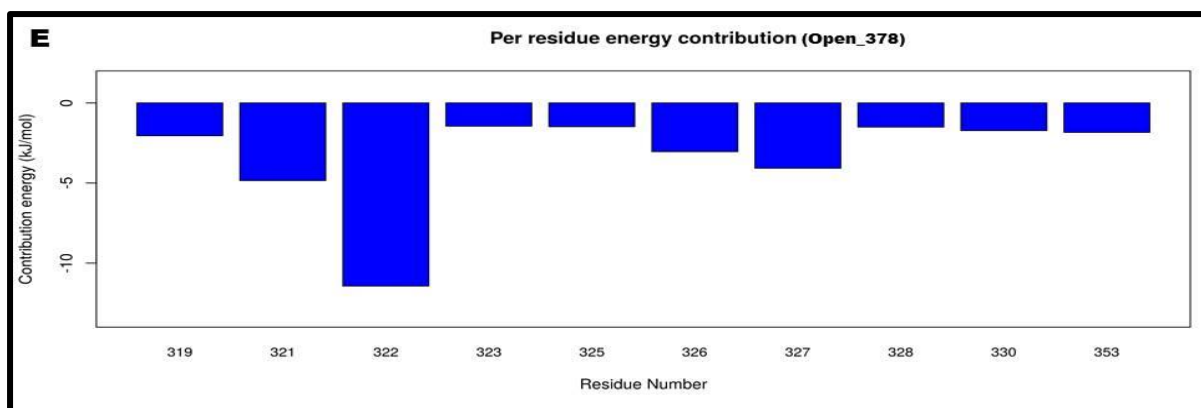
Figure 4.16: A histogram showing the per-residue contribution of SANC00378 when bound to B_prime.

4.4. Conclusion

One of the most difficult problems in drug design is the prediction of a protein-ligand binding pose. Although molecular docking can effectively generate hypothetical binding poses, the correct pose among many candidates is still difficult to predict (K. Liu and Kokubo, 2017). Molecular dynamics and associated methods are close to becoming routine drug discovery computing tools in determining the conformational evolution of protein structures when bound to small molecules (De Vivo *et al.,* 2016). In this chapter, 13 compounds were simulated with their respective protein structures. Also, the ligand-free (Apo) structure was simulated and this was used as a reference structure in determining the conformational changes when the protein structures are bound to the various ligands. A total, 15 MD simulations were be performed, 8 simulations for H_prime and 7 simulations for B_prime. Post MD analysis such as RMSD which is a measure of the conformational stability of a structure during simulation revealed conformational changes in the protein structures when bound to their respective ligands. RMSF identified residues with the highest fluctuations. H_prime exhibited the largest flexibility around the NBD subdomains IIA and IIB in all ligand-bound state. B_prime exhibited the greatest flexibility around subdomains SBDα also known as the alpha-helical lid in all ligand-bound state except when bound to SANC00430 (Open_430) which showed conformational fluctuation in the hydrophobic linker and SBDβ also known as the Beta substrate binding domain. Rg analysis revealed that all protein structures remained compact during the entire simulation. MM-PBSA was done to calculate the binding free energy of the protein-ligand complex. Van der Waals energy has been identified as the principal energetic term contributing to the ligand binding. Per-residue contribution identified residues contributing favourably to the binding free energy. Of notable importance are some of the allosteric residues identified by Penkler *et al.,* 2017. This includes

G196, G197, F216, E217, L219, A220, T221, N222, L320, K321, V322, D326, A583, A586, K587 and I601 for H_prime. Allosteric residues contributing significantly to B_prime includes P143, A144, Y145, F146, A153, V165, L392, D393, G405, V407, T409, V436, I438, V440, K446, D450, N451, V474, T475, F476, I478, D479, A480, D481, K502, A503, S504, S505. Of the 7 protein-ligand complexes of H_prime subjected to MM-PBSA, 5 complexes had negative binding energy ($\Delta G_{binding}$). Of the 6 protein-ligand complexes of B_prime subjected to MM-PBSA, 5 complexes had negative binding energy ($\Delta G_{binding}$). These complexes with negative binding energy ($\Delta G_{binding}$) were isolated for further analysis in the next chapter. The next chapter will be focused on examining the major conformational differences between ligand-free and ligand-bound structures.

CHAPTER FIVE: PRINCIPAL COMPONENT ANALYSIS

5.1. Chapter overview

Protein structures when bound to ligand(s) undergo relatively small conformational changes that are essential for biological activity. These changes range from a small movement of loops or side chains in the nucleotide-binding site to large-scale domain bending or even partial folding and unfolding of the protein structure (Brylinski and Skolnick, 2008). The characterization of these changes and their mechanism of interconversion are essential for biological processes. In this chapter, we aim to understand/analyse the severity of these changes resulting from the binding of ligands using principal component analysis (PCA). PCA reveals the most important motions in protein caused by the interactions with small molecules. H_prime and B_prime in both ligand-free and ligand-bound forms will be subjected to principal component analysis using both gromacs and MODE-TASK tools.

5.1.2. Principal component analysis (PCA)

The dynamics of proteins are established as changes in molecular structures or conformation depending on time. MD simulations of protein structures provide the positional movement of each atom in relation to a fixed reference frame of time (Ross *et al.,* 2018). This method although useful in characterizing the fluctuations in individual atom cannot be used in describing the inherent dynamic motion of protein structures due to its shortness of achievable times, hence the need to incorporate principal component analysis (PCA) that study long time dynamics by reducing the degree of freedom (DOF) to a few collective DOF which accounts for the essential dynamics of the system (Balsera *et al.,* 1996). PCA is a standard statistical tool used to analyse large multivariate data. This tool identifies patterns in data and expresses these patterns to emphasize their degrees of similarities and differences (Gkeka *et al.,* 2015). By definition, PCA is used to reduce the size of a data set consisting of a large number of interrelated attributes, while retaining as much of the variation as possible in the original data set. In respect to this study, PCA which is a covariance-matrix-based mathematical technique was used as a technique in the identification of global, correlated motions in atomic simulations of protein (Maisuradze *et al.,* 2009). PCA is a method that takes the trajectory of a simulation of molecular dynamics in terms of a small number of variables, sometimes referred to as essential degrees of freedom and extracts the most

important nodes in the movement of the molecule and this is performed on the Cartesian coordinates of the molecule (Haider *et al.,* 2008). This process is carried out by transforming the original set of attributes into a smaller set of attributes known as the principal components (PCs). The principal components are not correlated and ordered so that the first few components retain most of the variation in all the original characteristics. PCA acts on the covariance matrix, C, of the protein structure acquired from any length simulation of MD. The covariance matrix describes the protein's accessible DOF, such as the Cartesian coordinates that define atomic displacements in each conformation that comprises a trajectory. In order to obtain the covariance matrix, C, the protein coordinates are first superimposed on a reference structure which is usually the initial or average coordinates. To obtain C, first the protein coordinates are superimposed on a reference structure, usually the initial coordinates, or the average coordinates. The displacement vector described as the Cartesian coordinates for each residue at a time point t is obtained as $\Delta R_i(t)$. For a set of M coordinates, these are organized in the trajectory fluctuation matrix of order 3N x 3M.

$$\Delta R = \begin{bmatrix} \Delta R_1(t_1) & \Delta R_1(t_2) & \Delta R_1(t_M) \\ \Delta R_2(t_1) & \Delta R_2(t_2) & \Delta R_2(t_M) \\ \Delta R_3(t_1) & \Delta R_3(t_2) & \Delta R_3(t_M) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \Delta R_n(t_1) & \Delta R_n(t_2) & \Delta R_n(t_M) \end{bmatrix}$$

5.2. Methodology

PCA was performed on the MD trajectory of protein. The first step in PCA is the preparation of the trajectory obtained from MD simulation. Preparation of trajectory includes removal of periodicity (periodic boundary conditions) and removal of water molecules. GROMACS tool was used for used for the preparation of trajectory. Once the trajectory files are prepared, the next step is to run PCA. This tool requires two input files; a trajectory file (*xtc*) and a topology file (*tpr*). This was performed using MODE-TASK which is a range of tools for analysing and comparing protein dynamics from MD simulations. Running PCA includes standardizing the data and calculating the correlation matrix. The eigenvalues and eigenvectors of the correlation matrix are calculated, and these values are examined and interpreted. Finally, the transformed values are plot using a 2D plot using a python script.
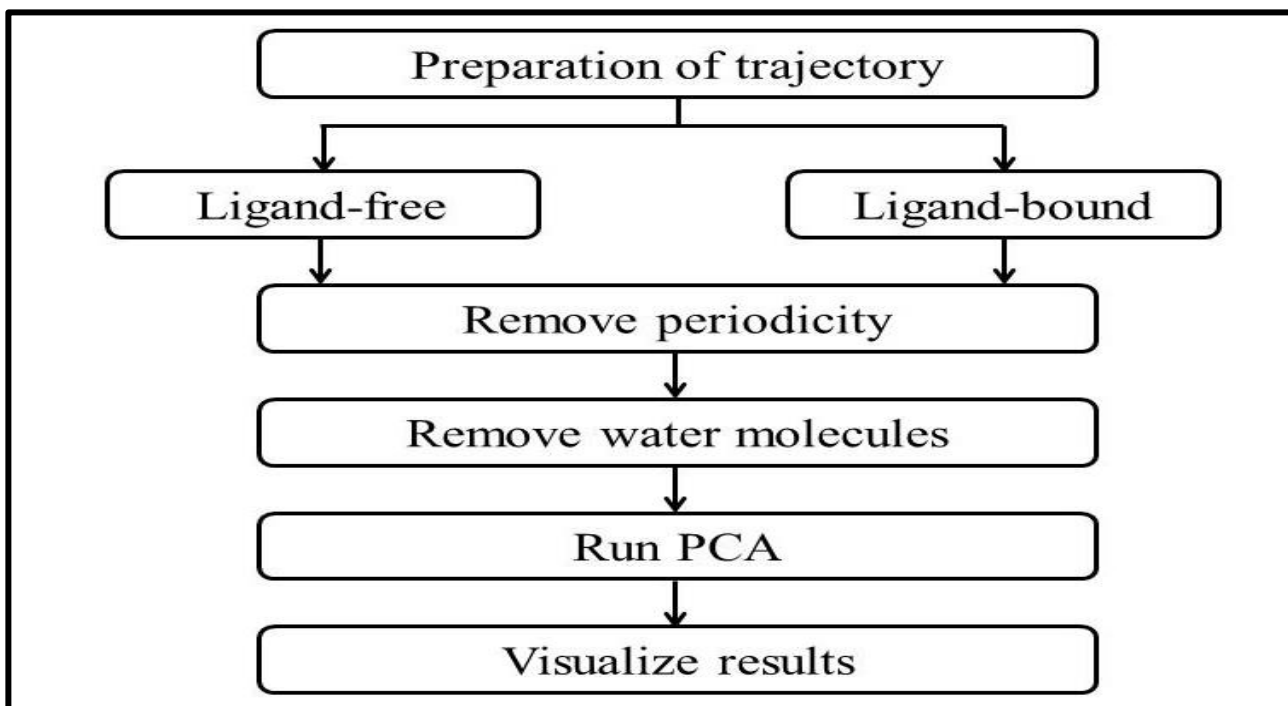
Figure 5.1: Summary of PCA workflow using MODE-TASK.

5.3 Results and Discussion

5.3.1. H_prime

The overall motion of H_prime was obtained from MD simulation; however, this motion was divided into principal components where each component represents a functional protein movement. Each component has its own set of coordinates which explains the motion of each atom in the system. The first two principal components (PC1 and PC2) retain the majority of the variance in the original distribution, providing a useful description of the system's conformational space. We first executed PCA for the ligand-free protein structure and compared the plots with that of the ligand-bound protein structures using the Cartesian coordinates of all backbone atoms. Figure 5.2 depicts the motion of the protein structures in both states during MD simulation and are displayed along PC1 and PC2, each colour is coded for each trajectory time. **A** represents the projection of Apo structure dynamics along PC1 and PC2; PC1 explained 38% of the variance while PC2 explained 29%. **B** represents the projection of protein in complex with SANC00499 along PC1 and PC2; PC1 explained 48% of the variance while PC2 explained 25%. **C** represents the projection of protein in complex with SANC00685 along PC1 and PC2; PC1 explained 45% of the variance while PC2 explained 26%. **D** represents the projection of protein in complex with SANC00492 along PC1 and PC2; PC1 explained 54% of the variance while PC2 explained 24%. **E** represents

the projection of protein in complex with SANC00340 along PC1 and PC2; PC1 explained 53% of the variance while PC2 explained 20%. **F** represents the projection of protein in complex with SANC00357 along PC1 and PC2; PC1 explained 42% of the variance while PC2 explained 29%.

Figure 5.2: Principal component analysis showing the structural relationship in terms of the conformational differences described by PC1 and PC2.

PC1 and PC2 accounts for the most variance in the protein's motion and these motions are responsible for the change in protein conformation observed during MD simulation. Figure 5.2 shows how the coordinates of the ligand-free and ligand-bound change significantly over time for each component 1 and 2. Some motions were observed when the protein structure was in complex with some of the ligands. Protein-ligand complexes that displayed significant movement during the simulation include protein in complex with SANC00499, SANC00357and SANC00492, and SANC00685. SANC00340 varied less over the MD simulation.

5.3.2. B_prime

Figure 5.3 shows the 2D plots of the first two PCs (PC1 and PC2). This plot also shows the essential dynamics of the protein structures during MD simulation. **A** represents the projection of Apo structure dynamics along PC1 and PC2; PC1 explained 77% of the variance while PC2 explained 13%. **B** represents the projection of protein in complex with SANC00284 along PC1 and PC2; PC1 explained 59% of the variance while PC2 explained 19%. **C** represents the projection of protein in complex with SANC00477 along PC1 and PC2; PC1 explained 71% of the variance while PC2 explained 10%. **C** represents the projection of protein in complex with SANC00477 along PC1 and PC2; PC1 explained 65% of the variance while PC2 explained 17%. **D** represents the projection of protein in complex with SANC00556 along PC1 and PC2; PC1 explained 52% of the variance while PC2 explained 21%. **E** represents the projection of protein in complex with SANC00676 along PC1 and PC2; PC1 explained 50% of the variance while PC2 explained 23%. **F** represents the projection of protein in complex with SANC00378 along PC1 and PC2; PC1 explained 59% of the variance while PC2 explained 19%.
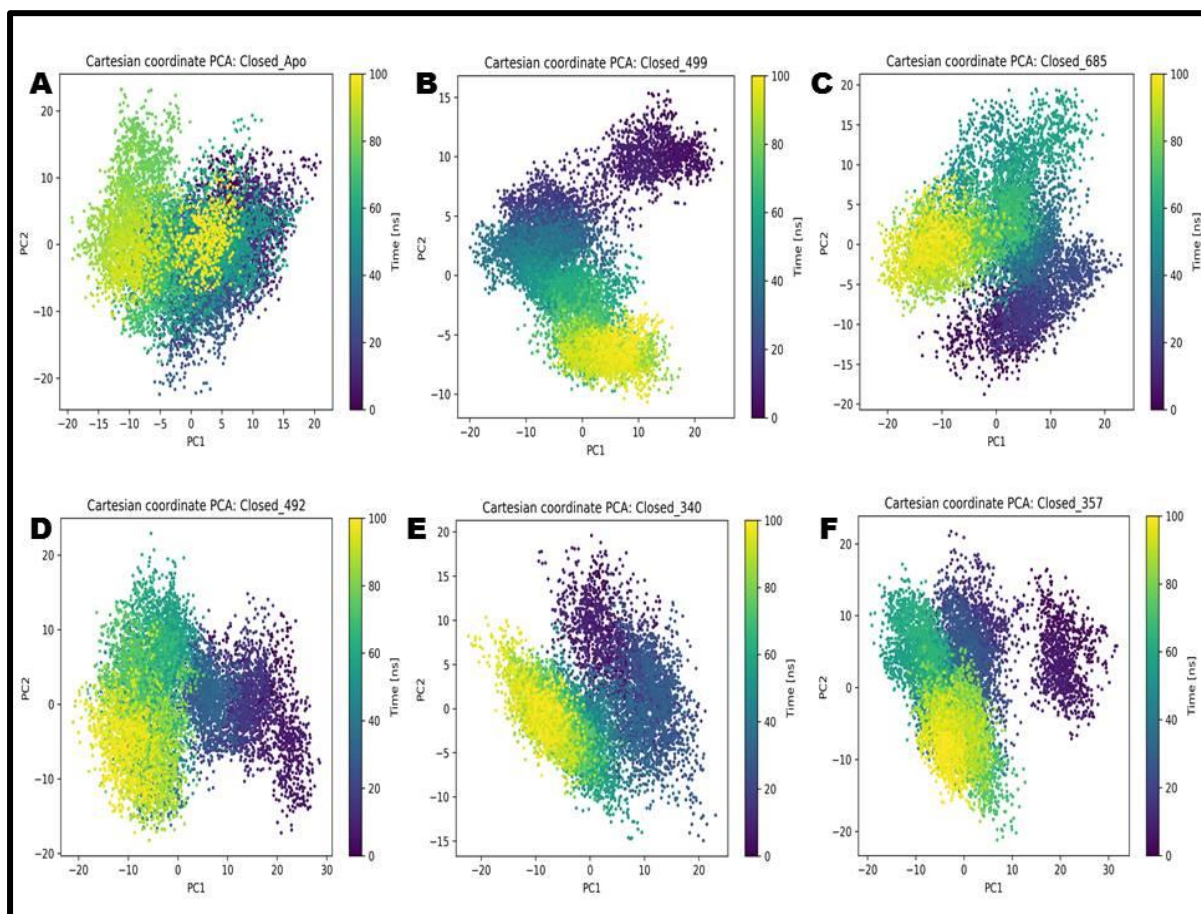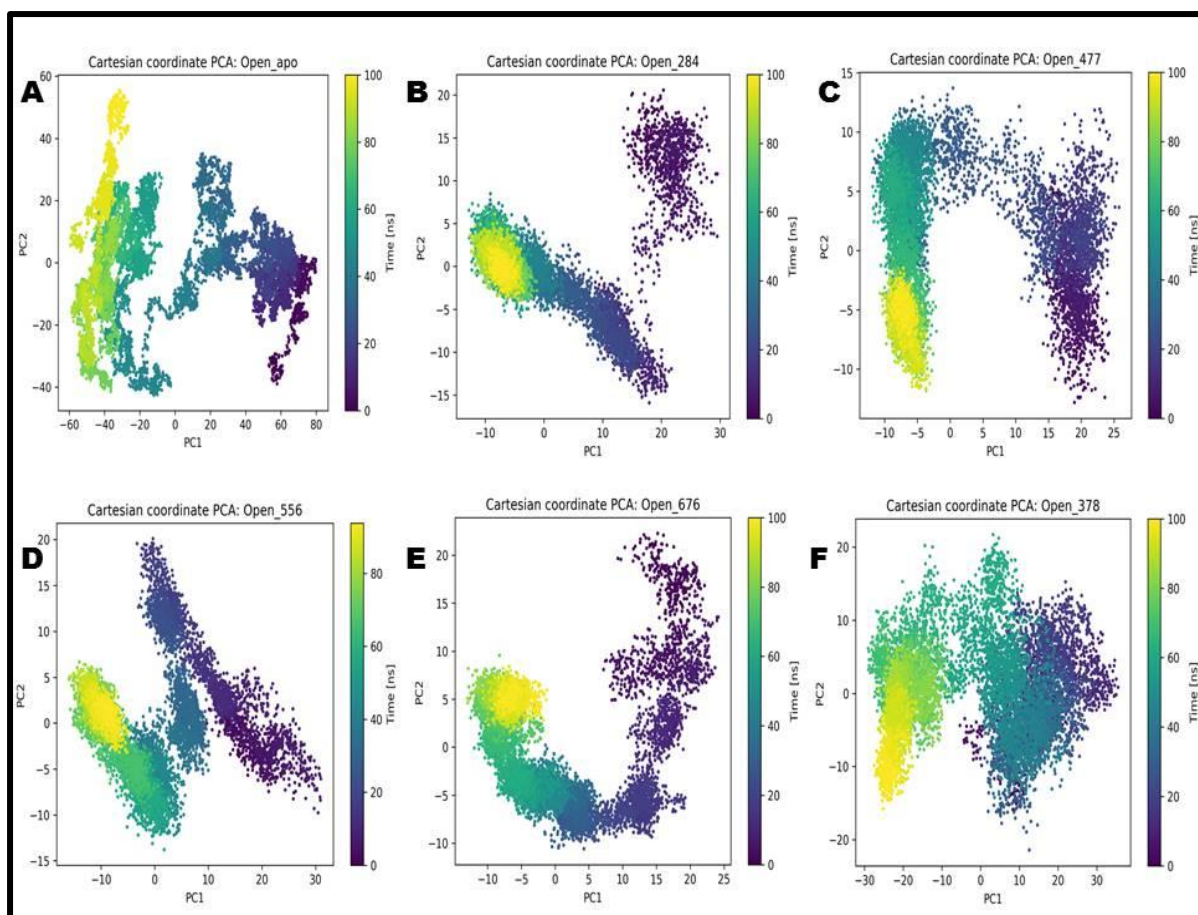
Figure 5.3: Principal component analysis showing the structural relationship in terms of the conformational differences described by PC1 and PC2.

In order to understand the conformational changes induced by the bound ligands, trajectories from MD simulation were projected into the two-dimensional subspace spanned by PC1 and PC2 as seen in Figure 5.3. From the above figure, we observed differences in motion between the ligand-free and ligand-bound structures. SANC00284, SANC00477, SANC00556, and SANC00676 showed very strong alteration of the motion during MD simulation. SANC00378 varied less during the simulation as it clustered together in the above figure.

5.4. Conclusion

Several biological functions such as allosteric regulation which is the basis of this study, are strongly linked to protein conformational transitions (Harada *et al.,* 2015). One of the most attractive features of PCA is the ability to identify the overall patterns of motions and to present data findings in some informative plots. PCA identifies configurational space with only a few degrees of freedom in which harmonic motion occurs by reducing the dimensionality of the data obtained from molecular dynamics simulations (Haider *et al.,* 2008). From this study, the dominant motions observed during MD simulation were identified. When bound to its respective ligands the motions were not similar indicating some effects the ligand has on the protein structures.

CHAPTER SIX: CONCLUSION AND FUTURE PROSPECTS

5.1 Concluding remarks

Hsp70s are the most extensively studied members of the heat shock family of proteins and are abundant in the bacterial and eukaryotic systems. These powerful molecular chaperones play a significant role in client (substrate protein) homeostasis and as such are important for cell development, and survival. Their wide distribution among different species, besides its connection to various diseases of global importance, such as malaria and cancer predisposes them as a promising therapeutic target. In this study, *E. coli* Hsp70 was used as a model organism. Allosteric sites of this protein in its multifunctional states prior to this study have not yet been explored for inhibitory design. Allosteric sites contain a lot of unexplored potentials in drug discovery, therefore; targeting allosteric sites of the protein will provide a new avenue towards the discovery of selective therapeutics agents for the treatment of a wide range of diseases. Targeting allosteric sites can offer limited advantages, therefore, in order to utilize allostery, it is fundamental to predict allosteric sites, allosteric modulators and residues implicated in the propagation of allosteric signals (Greener and Sternberg, 2018). FTMAP, SiteMap, and Allosite were employed in the identification of potential allosteric ligand binding sites. These tools revealed that Hsp70s in its open and closed conformation contains an immense number of unexplored binding sites. In its closed conformation (H_prime); FTMAP identified 10 allosteric binding sites, SiteMap identified 5 allosteric binding sites, while Allosite identified 9 allosteric binding sites. In its open conformation (B_prime); FTMAP identified 16 allosteric binding sites, SiteMap is identified 5 allosteric binding sites, while Allosite identified 9 allosteric binding sites. Consensus sites were then identified from these 3 prediction tools and the sites whose surrounding residues overlap with allosteric residues identified by Penkler et al., 2017 were isolated. Ligands binding to the isolated sites were selected for further analysis.

The protein structures were subjected to high throughput virtual screening against 623 SANCDB compounds and ligands were selected based on their binding poses, binding energies, and interactions to identified allosteric residues. Ligands fulfilling these criteria were isolated and taken to molecular dynamics (MD) simulations. In total 13 ligands were identified; 7 for H_prime and 6 for B_prime. Ligands binding to H_prime include SANC00499, SANC00685, SANC00492, SANC00340, SANC00357, SANC00378, and

SANC00430. Ligands binding to B_prime include SANC00284, SANC00378, SANC00556, SANC00676, SANC00477, and SANC00430.

Post-MD simulations analysis revealed the structural dynamics of the protein structures when bound to the selective ligands. RMSF identified residues with higher fluctuation and Rg revealed that when bound to the various ligands, the complexes remained compact. MM-PBSA calculations also revealed residues contributing significantly to binding free energy (BFE) of the complex and some of the residues contributing to BFE are residues identified as allosteric residues by Penkler *et al.,* 2017. Results from these analyses revealed potential allosteric modulators which are SANC00499, SANC00685, SANC00492, SANC00340 and SANC00357 for H_prime and SANC00284, SANC00477, SANC00576, SANC00676, and SANC00378. The impacts of the allosteric modulators were determined using principal component analysis which revealed the dominant motions observed during MD simulation.

5.2 Future prospects

Future work of this experiment includes an extension of the MD simulations to 200ns in order to provide a comprehensive evaluation of the compounds (SANC00499, SANC00685, and SANC00477). Other databases containing chemical compounds can be explored for compounds with inhibitory effects. Other future prospects include the analysis of ligands binding to other promising allosteric sites that were identified by the servers used in this research.

REFERENCES

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindah, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. https://doi.org/10.1016/j.softx.2015.06.001

Ahmed, M. (2018). Acute cholangitis - an update. *World Journal of Gastrointestinal Pathophysiology*. https://doi.org/10.4291/wjgp.v9.i1.1

Alanazi, M. Q., Alqahtani, F. Y., & Aleanizy, F. S. (2018). An evaluation of E. coli in urinary tract infection in emergency department at KAMC in Riyadh, Saudi Arabia: Retrospective study. *Annals of Clinical Microbiology and Antimicrobials*. https://doi.org/10.1186/s12941-018-0255-z

Arcon, J. P., Defelipe, L. A., Modenutti, C. P., López, E. D., Alvarez-Garcia, D., Barril, X., … Martí, M. A. (2017). Molecular Dynamics in Mixed Solvents Reveals Protein-Ligand Interactions, Improves Docking, and Allows Accurate Binding Free Energy Predictions. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/acs.jcim.6b00678

Asea, A., Rehli, M., Kabingu, E., Boch, J. A., Baré, O., Auron, P. E., … Calderwood, S. K. (2002). Novel signal transduction pathway utilized by extracellular HSP70. Role of toll-like receptor (TLR) 2 and TLR4. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M200497200

Assimon, V. A., Gillies, A. T., Rauch, J. N., & Gestwicki, J. E. (2013). Hsp70 protein complexes as drug targets. *Current Pharmaceutical Design*. https://doi.org/CPD-EPUB-20120816-6 [pii]

Balaburski, G. M., Leu, J. I.-J., Beeharry, N., Hayik, S., Andrake, M. D., Zhang, G., … Murphy, M. E. (2013). A Modified HSP70 Inhibitor Shows Broad Activity as an Anticancer Agent. *Molecular Cancer Research*. https://doi.org/10.1158/1541-7786.MCR-12-0547-T

Ballester, P. J., Mangold, M., Howard, N. I., Marchese Robinson, R. L., Abell, C., Blumberger, J., & Mitchell, O. B. O. (2012). Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification. *Journal of the*

Royal Society Interface. https://doi.org/10.1098/rsif.2012.0569

Balsera, M. A., Wriggers, W., Oono, Y., & Schulten, K. (1996). Principal component analysis and long time protein dynamics. *Journal of Physical Chemistry*. https://doi.org/10.1021/jp9536920

Baylis, C., Uyttendaele, M., Joosten, H., Davies, A., & Heinz, H. J. (2011). *The Enterobacteriaceae and their significance to the food industry*. *ILSI Europe Report Series*. https://doi.org/2011/10.996/30

Beglov, D., Hall, D. R., Wakefield, A. E., Luo, L., Allen, K. N., Kozakov, D., … Vajda, S. (2018). Exploring the structural origins of cryptic sites on proteins. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1711490115

Bell, S. L., Chiang, A. N., & Brodsky, J. L. (2011). Expression of a malarial Hsp70 improves defects in chaperone-dependent activities in ssa1 mutant yeast. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0020047

Bellelli, A., & Brunori, M. (2011). Hemoglobin allostery: Variations on the theme. In *Biochimica et Biophysica Acta - Bioenergetics*. https://doi.org/10.1016/j.bbabio.2011.04.004

Benjamin, I. J., & Mcmillan, D. R. (1998). Stress (Heat Shock) Proteins Molecular Chaperones in Cardiovascular Biology and Disease. *Circ Res*. https://doi.org/10.1161/01.RES.83.2.117

Berendsen, H. J. C., van der Spoel, D., & van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*. https://doi.org/10.1016/0010-4655(95)00042-E

Bertelsen, E. B., Chang, L., Gestwicki, J. E., & Zuiderweg, E. R. P. (2009). Solution conformation of wild-type E. coli Hsp70 (DnaK) chaperone complexed with ADP and substrate. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0903503106

Bessong, P. O., Obi, C. L., Andréola, M. L., Rojas, L. B., Pouységu, L., Igumbor, E., … Litvak, S. (2005). Evaluation of selected South African medicinal plants for inhibitory properties against human immunodeficiency virus type 1 reverse transcriptase and

integrase. *Journal of Ethnopharmacology*. https://doi.org/10.1016/j.jep.2005.01.056

Blount, Z. D. (2015). The unexhausted potential of E. coli. *ELife*. https://doi.org/10.7554/eLife.05826

Blundell, T. L., & Patel, S. (2004). High-throughput X-ray crystallography for drug discovery. *Current Opinion in Pharmacology*. https://doi.org/10.1016/j.coph.2004.04.007

Bowers, K. J., Sacerdoti, F. D., Salmon, J. K., Shan, Y., Shaw, D. E., Chow, E., … Moraes, M. A. (2006). Molecular dynamics---Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC '06*. https://doi.org/10.1145/1188455.1188544

Brenke, R., Kozakov, D., Chuang, G. Y., Beglov, D., Hall, D., Landon, M. R., … Vajda, S. (2009). Fragment-based identification of druggable "hot spots" of proteins using Fourier domain correlation techniques. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp036

Brooijmans, N., & Kuntz, I. D. (2003). Molecular Recognition and Docking Algorithms. *Annual Review of Biophysics and Biomolecular Structure*. https://doi.org/10.1146/annurev.biophys.32.110601.142532

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*. https://doi.org/10.1002/jcc.540040211

Brylinski, M., & Skolnick, J. (2008). What is the relationship between the global structures of apo and holo proteins? *Proteins: Structure, Function and Genetics*. https://doi.org/10.1002/prot.21510

Buchner, J. (1996). Supervising the fold: functional principles of molecular chaperones. *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*. https://doi.org/10.1096/fj.1530-6860

Bukau, B., & Horwich, A. L. (1998). The Hsp70 and Hsp60 chaperone machines. *Cell*. https://doi.org/10.1016/S0092-8674(00)80928-9

Cakmak, Y. O. (2009). A review of the potential effect of electroacupuncture and moxibustion on cell repair and survival: The role of heat shock proteins. *Acupuncture in Medicine*. https://doi.org/10.1136/aim.2009.001420

Calloni, G., Chen, T., Schermann, S. M., Chang, H. C., Genevaux, P., Agostini, F., … Hartl, F. U. (2012). DnaK Functions as a Central Hub in the E. coli Chaperone Network. *Cell Reports*. https://doi.org/10.1016/j.celrep.2011.12.007

Canny, G. O., & McCormick, B. A. (2008). Bacteria in the intestine, helpful residents or enemies from within? *Infection and Immunity*. https://doi.org/10.1128/IAI.00187-08

Chalova, V. I., Sirsat, S. A., O'Bryan, C. A., Crandall, P. G., & Ricke, S. C. (2009). Escherichia coli, an intestinal microorganism, as a biosensor for quantification of amino acid bioavailability. *Sensors*. https://doi.org/10.3390/s90907038

Chauret, C. (2011). Survival and control of Escherichia coli O157:H7 in foods, beverages, soil and water. *Virulence*. https://doi.org/10.4161/viru.2.6.18423

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today*. https://doi.org/10.1016/j.drudis.2018.01.039

Chiappori, F., Fumian, M., Milanesi, L., & Merelli, I. (2015). DnaK as antibiotic target: Hot spot residues analysis for differential inhibition of the bacterial protein in comparison with the human Hsp70. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0124563

Chow, D. A. (2005). Physiological activities of the natural immune system. *NeuroImmune Biology*. https://doi.org/10.1016/S1567-7443(05)80020-1

Christopoulos, A. (2002). G Protein-Coupled Receptor Allosterism and Complexing. *Pharmacological Reviews*. https://doi.org/10.1124/pr.54.2.323

Christopoulos, A., May, L. T., Avlani, V. A., & Sexton, P. M. (2004). G-protein-coupled receptor allosterism: the promise and the problem(s). *Biochemical Society Transactions*. https://doi.org/10.1042/BST0320873

Chung, H. J., Bang, W., & Drake, M. A. (2006). Stress response of Escherichia coli. *Comprehensive Reviews in Food Science and Food Safety*. https://doi.org/10.1111/j.1541-4337.2006.00002.x

Clarkson, C., Stærk, D., Hansen, S. H., Smith, P. J., & Jaroszewski, J. W. (2006). Discovering new natural products directly from crude extracts by HPLC-SPE-NMR: Chinane diterpenes in Harpagophytum procumbens. *Journal of Natural Products*. https://doi.org/10.1021/np050504g

Cui, Q., & Karplus, M. (2008). Allostery and cooperativity revisited. *Protein Science*. https://doi.org/10.1110/ps.03259908

D.A. Case and R.M. Betz and W. Botello-Smith and D.S. Cerutti and T.E. Cheatham and III and T.A. Darden and R.E. Duke and T.J. Giese and H. Gohlke and A.W. Goetz and N. Homeyer and S. Izadi and P. Janowski and J. Kaus and A. Kovalenko and T.S. Lee and S. (2018). AMBER 2018. *University of California, San Francisco*.

Davies, J., Johnson, A. P., & Hope, R. (2017). Identifying hospital-onset Escherichia coli bacteraemia cases from English mandatory surveillance: the case for applying a two-day post-admission rule. *Journal of Hospital Infection*. https://doi.org/10.1016/j.jhin.2017.06.031

Davis, J. E., Voisine, C., & Craig, E. a. (1999). Intragenic suppressors of Hsp70 mutants: interplay between the ATPase- and peptide-binding domains. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.96.16.9269

De Vivo, M., Masetti, M., Bottegoni, G., & Cavalli, A. (2016). Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry*. https://doi.org/10.1021/acs.jmedchem.5b01684

De Wet, H., Van Heerden, F. R., & Van Wyk, B. E. (2011). Alkaloidal variation in Cissampelos capensis (Menispermaceae). *Molecules*. https://doi.org/10.3390/molecules16043001

Donnenberg, M. S., & Whittam, T. S. (2001). Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic Escherichia coli. *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI12404

Drwal, M. N., & Griffith, R. (2013). Combination of ligand- and structure-based methods in virtual screening. *Drug Discovery Today: Technologies*. https://doi.org/10.1016/j.ddtec.2013.02.002

Du Toit, K., Elgorashi, E. E., Malan, S. F., Drewes, S. E., Van Staden, J., Crouch, N. R., & Mulholland, D. A. (2005). Anti-inflammatory activity and QSAR studies of compounds isolated from Hyacinthaceae species and Tachiadenus longiflorus Griseb. (Gentianaceae). *Bioorganic and Medicinal Chemistry*. https://doi.org/10.1016/j.bmc.2005.01.036

Einav, T., Mazutis, L., & Phillips, R. (2016). Statistical Mechanics of Allosteric Enzymes. *Journal of Physical Chemistry B*. https://doi.org/10.1021/acs.jpcb.6b01911

Ekins, S., Mestres, J., & Testa, B. (2007). In silico pharmacology for drug discovery: Methods for virtual ligand screening and profiling. *British Journal of Pharmacology*. https://doi.org/10.1038/sj.bjp.0707305

Elsas, J. D. Van, Semenov, A. V, Costa, R., & Trevors, J. T. (2010). Survival of Escherichia coli in the environment : fundamental and public health aspects. *The ISME Journal*. https://doi.org/10.1038/ismej.2010.80

English, C. A., Sherman, W., Meng, W., & Gierasch, L. M. (2017). The Hsp70 interdomain linker is a dynamic switch that enables allosteric communication between two structured domains. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M117.789313

Fontaine, S. N., Rauch, J. N., Nordhues, B. A., Assimon, V. A., Stothert, A. R., Jinwal, U. K., … Dickey, C. A. (2015). Isoform-selective genetic inhibition of constitutive cytosolic Hsp70 activity promotes client Tau degradation using an altered co-chaperone complement. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M115.637595

Fouche, G., Cragg, G. M., Pillay, P., Kolesnikova, N., Maharaj, V. J., & Senabe, J. (2008). In vitro anticancer screening of South African plants. *Journal of Ethnopharmacology*. https://doi.org/10.1016/j.jep.2008.07.005

Freiburger, L. A., Baettig, O. M., Sprules, T., Berghuis, A. M., Auclair, K., & Mittermaier, A. K. (2011). Competing allosteric mechanisms modulate substrate binding in a dimeric enzyme. *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.1978

Genevaux, P., Georgopoulos, C., & Kelley, W. L. (2007). The Hsp70 chaperone machines of Escherichia coli: A paradigm for the repartition of chaperone functions. *Molecular Microbiology*. https://doi.org/10.1111/j.1365-2958.2007.05961.x

Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*. https://doi.org/10.1517/17460441.2015.1032936

Gerek, Z. N., & Ozkan, S. B. (2011). Change in allosteric network affects binding affinities of PDZ domains: Analysis through perturbation response scanning. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002154

Ghersi, D., & Sanchez, R. (2009). Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. *Proteins: Structure, Function and Bioinformatics*. https://doi.org/10.1002/prot.22154

Giridhar, R. (2012). Drug discovery: Past and present. *Journal of Advanced Pharmaceutical Technology & Research*, *3*(1), 2. https://doi.org/10.4103/2231-4040.93554

Gkeka, P., Papafotika, A., Christoforidis, S., & Cournia, Z. (2015). Exploring a Non-ATP pocket for potential allosteric modulation of pi3kα. *Journal of Physical Chemistry B*. https://doi.org/10.1021/jp506423e

Greener, J. G., & Sternberg, M. J. (2018). Structure-based prediction of protein allostery. *Current Opinion in Structural Biology*. https://doi.org/10.1016/j.sbi.2017.10.002

Gribaldo, S., Lumia, V., Creti, R., Conway De Macario, E., Sanangelantoni, A., & Cammarano, P. (1999). Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. *Journal of Bacteriology*. https://doi.org/10.1016/j.ejc.2006.02.002

Grover, A. K. (2013). Use of allosteric targets in the discovery of safer drugs. *Medical Principles and Practice*. https://doi.org/10.1159/000350417

Gudipati, S., M, R., Mankad, A., Pandya, H., & Jasrai, Y. . (2018). Molecular docking based screening of Noggin inhibitors. *Bioinformation*, *14*(01), 015–020. https://doi.org/10.6026/97320630014015

Haider, S., Parkinson, G. N., & Neidle, S. (2008). Molecular dynamics and principal components analysis of human telomeric quadruplex multimers. *Biophysical Journal*. https://doi.org/10.1529/biophysj.107.120501

Halgren, T. (2007). New method for fast and accurate binding-site identification and analysis. *Chemical Biology and Drug Design*. https://doi.org/10.1111/j.1747-0285.2007.00483.x

Halgren, T. A. (2009). Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/ci800324m

Hamza, A., Wei, N. N., & Zhan, C. G. (2012). Ligand-based virtual screening approach using a new scoring function. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/ci200617d

Hao, W.-L., & Lee, Y.-K. (2004). Microflora of the gastrointestinal tract: a review. *Methods in Molecular Biology (Clifton, N.J.)*. https://doi.org/10.1385/1-59259-766-1:491

Harada, R., Takano, Y., Baba, T., & Shigeta, Y. (2015). Simple, yet powerful methodologies for conformational sampling of proteins. *Physical Chemistry Chemical Physics*. https://doi.org/10.1039/c4cp05262e

Hatherley, R., Blatch, G. L., & Bishop, Ö. T. (2014). Plasmodium falciparum Hsp70-x: A heat shock protein at the host-parasite interface. *Journal of Biomolecular Structure and Dynamics*. https://doi.org/10.1080/07391102.2013.834849

Hatherley, R., Brown, D. K., Musyoka, T. M., Penkler, D. L., Faya, N., Lobb, K. A., & Tastan Bishop, Ö. (2015). SANCDB: A South African natural compound database. *Journal of Cheminformatics*, *7*(1), 1–9. https://doi.org/10.1186/s13321-015-0080-8

Health Protection Agency. (2013). UK Standards for Microbiology Investigations Identification of Campylobacter species. *Standards Unit, Microbiology Services*, *ID23*, 1–22. https://doi.org/ID 7

Helmstaedt, K., Krappmann, S., & Braus, G. H. (2001). Allosteric Regulation of Catalytic Activity: Escherichia coli Aspartate Transcarbamoylase versus Yeast Chorismate Mutase. *Microbiology and Molecular Biology Reviews*. https://doi.org/10.1128/MMBR.65.3.404-421.2001

Hilser, V. J., Wrabl, J. O., & Motlagh, H. N. (2012). Structural and Energetic Basis of Allostery. *Annual Review of Biophysics*. https://doi.org/10.1146/annurev-biophys-050511-102319

Hoffman, P. S., & Garduno, R. A. (1999). Surface-associated heat shock proteins of Legionella pneumophila and Helicobacter pylori: roles in pathogenesis and immunity. *Infectious Diseases in Obstetrics and Gynecology*. https://doi.org/10.1155/S1064744999000125

Hospital, A., Goñi, J. R., Orozco, M., & Gelpí, J. L. (2015). Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*. https://doi.org/10.2147/AABC.S70333

Hou, T., Wang, J., Li, Y., & Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/ci100275a

Huang, W., Lu, S., Huang, Z., Liu, X., Mou, L., Luo, Y., … Zhang, J. (2013). Allosite: A method for predicting allosteric sites. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt399

Huang, W., Nussinov, R., & Zhang, J. (2017). Computational tools for allosteric drug discovery: Site identification and focus library design. In *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-4939-6637-0_23

Ishii, S., & Sadowsky, M. J. (2008). Escherichia coli in the Environment: Implications for Water Quality and Human Health. *Microbes and Environments*, *23*(2), 101–108. https://doi.org/10.1264/jsme2.23.101

Jakupovic, J., Zdero, C., Grenz, M., Tsichritzis, F., Lehmann, L., Hashemi-Nejad, S. M., & Bohlmann, F. (1989). Twenty-one acylphloroglucinol derivatives and further constituents from south african Helichrysum species. *Phytochemistry*. https://doi.org/10.1016/0031-9422(89)80195-5

Jiang, J., Maes, E. G., Taylor, A. B., Wang, L., Hinck, A. P., Lafer, E. M., & Sousa, R. (2007). Structural Basis of J Cochaperone Binding and Regulation of Hsp70. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2007.08.022

Kampinga, H. H., & Craig, E. A. (2010). The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm2941

Kim, L. S., & Kim, J. H. (2011). Heat shock protein as molecular targets for breast cancer therapeutics. *J Breast Cancer*. https://doi.org/10.4048/jbc.2011.14.3.167

Kityk, R., Kopp, J., Sinning, I., & Mayer, M. P. (2012). Structure and Dynamics of the ATP-Bound Open Conformation of Hsp70 Chaperones. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2012.09.023

Koorbanally, C., Crouch, N. R., & Mulholland, D. A. (2006). The phytochemistry and ethnobotany of the southern African genus Eucomis (Hyacinthaceae; Hyacinthoideae). *Phytochemistry: Advances in Research*. https://doi.org/10.1021/nn901374b

Koorbanally, C., Mulholland, D. A., & Crouch, N. R. (2005). A novel 3-hydroxy-3-benzyl-4-chromanone-type homoisoflavonoid from Albuca fastigiata (Ornithogaloideae: Hyacinthaceae). *Biochemical Systematics and Ecology*. https://doi.org/10.1016/j.bse.2004.08.009

Koorbanally, C., Mulholland, D. A., & Crouch, N. R. (2006). Norlignans and homoisoflavanones from two South African Drimiopsis species (Hyacinthaceae: Hyacinthoideae). *Biochemical Systematics and Ecology*. https://doi.org/10.1016/j.bse.2005.12.011

Koorbanally, N. A., Crouch, N. R., Harilal, A., Pillay, B., & Mulholland, D. A. (2006). Coincident isolation of a novel homoisoflavonoid from Resnova humifusa and Eucomis montana (Hyacinthoideae: Hyacinthaceae). *Biochemical Systematics and Ecology*. https://doi.org/10.1016/j.bse.2005.08.003

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., … Vajda, S. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols*. https://doi.org/10.1038/nprot.2015.043

Kragol, G., Hoffmann, R., Chattergoon, M. A., Lovas, S., Cudic, M., Bulet, P., … Otvos, L. (2002). Identification of crucial residues for the antibacterial activity of the proline-rich peptide, pyrrhocoricin. *European Journal of Biochemistry*. https://doi.org/10.1046/j.1432-1033.2002.03119.x

Kumar, S. J., Stokes, J., Singh, U. P., Scissum Gunn, K., Acharya, A., Manne, U., & Mishra, M. (2016). Targeting Hsp70: A possible therapy for cancer. *Cancer Letters*. https://doi.org/10.1016/j.canlet.2016.01.056

Kumari, R., Kumar, R., & Lynn, A. (2014). G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/ci500020m

Kuroda, M., Mimaki, Y., Yokosuka, A., Hasegawa, F., & Sashida, Y. (2002). Cholestane glycosides from the bulbs of Ornithogalum thyrsoides and their cytotoxic activity against HL-60 leukemia cells. *Journal of Natural Products*. https://doi.org/10.1021/np020114j

Kuroda, M., Mimaki, Y., Yokosuka, A., & Sashida, Y. (2001). Cholestane glycosides from the bulbs of Galtonia candicans and their cytotoxicity. *Chemical & Pharmaceutical Bulletin*. https://doi.org/10.1248/cpb.49.1042

Kuroda, M., Ori, K., & Mimaki, Y. (2006). Ornithosaponins A-D, four new polyoxygenated steroidal glycosides from the bulbs of Ornithogalum thyrsoides. *Steroids*. https://doi.org/10.1016/j.steroids.2005.10.001

Lääveri, T., Vilkman, K., Pakkanen, S. H., Kirveskari, J., & Kantele, A. (2018). A prospective study of travellers' diarrhoea: analysis of pathogen findings by destination in various (sub)tropical regions. *Clinical Microbiology and Infection*. https://doi.org/10.1016/j.cmi.2017.10.034

Lambers, J. (2009). Three-Dimensional Coordinate Systems, 1–5.

Lamprecht, C., Romanis, M., Huisamen, N., Carinus, A., Schoeman, N., Sigge, G. O., & Britz, T. J. (2014). Escherichia coli with virulence factors and multidrug resistance in the Plankenburg River. *South African Journal of Science*, *110*(9–10), 1–6. https://doi.org/10.1590/sajs.2014/20130347

Landon, M. R., Lieberman, R. L., Hoang, Q. Q., Ju, S., Caaveiro, J. M. M., Orwig, S. D., … Ringe, D. (2009). Detection of ligand binding hot spots on protein surfaces via fragment-based methods: Application to DJ-1 and glucocerebrosidase. *Journal of Computer-Aided Molecular Design*. https://doi.org/10.1007/s10822-009-9283-2

Laufen, T., Mayer, M. P., Beisel, C., Klostermeier, D., Mogk, A., Reinstein, J., & Bukau, B. (1999). Mechanism of regulation of Hsp70 chaperones by DnaJ cochaperones. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.96.10.5452

Lavecchia, A., & Giovanni, C. (2013). Virtual Screening Strategies in Drug Discovery: A Critical Review. *Current Medicinal Chemistry*. https://doi.org/10.2174/09298673113209990001

Leão-Vasconcelos, L. S. N. de O., Lima, A. B. M., Costa, D. de M., Rocha-Vilefort, L. O., Oliveira, A. C. A. de, Gonçalves, N. F., … Prado-Palos, M. A. (2015). Enterobacteriaceae ISOLATES FROM THE ORAL CAVITY OF WORKERS IN A BRAZILIAN ONCOLOGY HOSPITAL. *Rev. Inst. Med. Trop. Sao Paulo*. https://doi.org/10.1590/S0036-46652015000200004

Lee, P. S., & Lee, K. H. (2003). Escherichia coli - A Model System That Benefits from and Contributes to the Evolution of Proteomics. *Biotechnology and Bioengineering*. https://doi.org/10.1002/bit.10848

Leu, J. I. J., Pimkina, J., Frank, A., Murphy, M. E., & George, D. L. (2009). A Small Molecule Inhibitor of Inducible Heat Shock Protein 70. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2009.09.023

Leu, J. I. J., Zhang, P., Murphy, M. E., Marmorstein, R., & George, D. L. (2014). Structural basis for the inhibition of HSP70 and DnaK chaperones by small-molecule targeting of a C-terminal allosteric pocket. *ACS Chemical Biology*. https://doi.org/10.1021/cb500236y

Li, X., Chen, Y., Lu, S., Huang, Z., Liu, X., Wang, Q., … Zhang, J. (2013). Toward an understanding of the sequence and structural basis of allosteric proteins. *Journal of Molecular Graphics and Modelling*. https://doi.org/10.1016/j.jmgm.2012.12.011

Li, X., Shao, H., Taylor, I. R., & Gestwicki, J. E. (2016). Targeting Allosteric Control Mechanisms in Heat Shock Protein 70 (Hsp70). *Current Topics in Medicinal Chemistry*. https://doi.org/http://dx.doi.org/10.2174/1568026616666160413140911

Li, Z., & Srivastava, P. (2004). Heat-shock proteins. *Current Protocols in Immunology / Edited by John E. Coligan ... [et Al.]*. https://doi.org/10.1002/0471142735.ima01ts58

Lionta, E., Spyrou, G., Vassilatis, D., & Cournia, Z. (2014). Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Current Topics in Medicinal Chemistry*. https://doi.org/10.2174/1568026614666140929124445

Liu, J., & Nussinov, R. (2016). Allostery: An Overview of Its History, Concepts, Methods,

and Applications. *PLoS Computational Biology*.
https://doi.org/10.1371/journal.pcbi.1004966

Liu, J., Yan, Q., Luo, F., Shang, D., Wu, D., Zhang, H., … Xin, Y. (2015). Acute
cholecystitis associated with infection of Enterobacteriaceae from gut microbiota.
Clinical Microbiology and Infection. https://doi.org/10.1016/j.cmi.2015.05.017

Liu, K., & Kokubo, H. (2017). Exploring the Stability of Ligand Binding Modes to Proteins
by Molecular Dynamics Simulations: A Cross-docking Study. *Journal of Chemical
Information and Modeling*. https://doi.org/10.1021/acs.jcim.7b00412

Llorach-Pares, L., Nonell-Canals, A., Sanchez-Martinez, M., & Avila, C. (2017). Computer-
aided drug design applied to marine drug discovery: Meridianins as alzheimer's disease
therapeutic agents. *Marine Drugs*. https://doi.org/10.3390/md15120366

Lu, S., Li, S., & Zhang, J. (2014). Harnessing Allostery: A Novel Approach to Drug
Discovery. *Medicinal Research Reviews*. https://doi.org/10.1002/med.21317

Maisuradze, G. G., Liwo, A., & Scheraga, H. A. (2009). Principal Component Analysis for
Protein Folding Dynamics. *Journal of Molecular Biology*.
https://doi.org/10.1016/j.jmb.2008.10.018

Martínez, L. (2015). Automatic identification of mobile and rigid substructures in molecular
dynamics simulations and fractional structural fluctuation analysis. *PLoS ONE*.
https://doi.org/10.1371/journal.pone.0119264

Mayer, M. P., Brehmer, D., Gässler, C. S., & Bukau, B. (2001). Hsp70 chaperone machines.
Advances in Protein Chemistry. https://doi.org/10.1016/S0065-3233(01)59001-4

Mayer, M. P., & Bukau, B. (2005). Hsp70 chaperones: Cellular functions and molecular
mechanism. *Cellular and Molecular Life Sciences*. https://doi.org/10.1007/s00018-004-
4464-6

Meng, X.-Y., Zhang, H.-X., Mezei, M., & Cui, M. (2011). Molecular Docking: A Powerful
Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design*.
https://doi.org/10.2174/157340911795677602

Meyer, J. J. M., Rakuambo, N. C., & Hussein, A. A. (2008). Novel xanthones from
Securidaca longepedunculata with activity against erectile dysfunction. *Journal of*

Ethnopharmacology. https://doi.org/10.1016/j.jep.2008.06.018

Michael, S., Auld, D., Klumpp, C., Jadhav, A., Zheng, W., Thorne, N., … Simeonov, A. (2008). A Robotic Platform for Quantitative High-Throughput Screening. *ASSAY and Drug Development Technologies*. https://doi.org/10.1089/adt.2008.150

Mirzaei, H., Zarbafian, S., Villar, E., Mottarella, S., Beglov, D., Vajda, S., … Kozakov, D. (2015). Energy minimization on manifolds for docking flexible molecules. *Journal of Chemical Theory and Computation*. https://doi.org/10.1021/ct500155t

Miyata, Y., Chang, L., Bainor, A., McQuade, T. J., Walczak, C. P., Zhang, Y., … Gestwicki, J. E. (2010). High-throughput screen for escherichia coli heat shock protein 70 (Hsp70/DnaK): ATPase assay in low volume by exploiting energy transfer. *Journal of Biomolecular Screening*, *15*(10), 1211–1219. https://doi.org/10.1177/1087057110380571

Morimoto, R. I., & Cuervo, A. M. (2009). Protein homeostasis and aging: Taking care of proteins from the cradle to the grave. In *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*. https://doi.org/10.1093/gerona/gln071

Morris, G. M., & Lim-Wilby, M. (2008). Molecular docking. *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-59745-177-2_19

Motlagh, H. N., Wrabl, J. O., Li, J., & Hilser, V. J. (2014). The ensemble nature of allostery. *Nature*. https://doi.org/10.1038/nature13001

Nachin, L., Nannmark, U., & Nyström, T. (2005). Differential roles of the universal stress proteins of Escherichia coli in oxidative stress resistance, adhesion, and motility. *Journal of Bacteriology*. https://doi.org/10.1128/JB.187.18.6265-6272.2005

Nataro, J. P., & Kaper, J. B. (1998). Diarrheagenic Escherichia coli. *Clinical Microbiology Reviews*. https://doi.org/file://Z:\References\Text Files\00000004472.txt

Ngan, C. H., Bohnuud, T., Mottarella, S. E., Beglov, D., Villar, E. A., Hall, D. R., … Vajda, S. (2012). FTMAP: Extended protein mapping with user-selected probe molecules. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks441

Nollen, E. a a, & Morimoto, R. I. (2002). Chaperoning signaling pathways: molecular chaperones as stress-sensing "heat shock" proteins. *Journal of Cell Science*.

Nussinov, R., & Tsai, C.-J. (2015). The Design of Covalent Allosteric Drugs. *Annual Review of Pharmacology and Toxicology*. https://doi.org/10.1146/annurev-pharmtox-010814-124401

Nylandsted, J., Gyrd-Hansen, M., Danielewicz, A., Fehrenbacher, N., Lademann, U., Høyer-Hansen, M., … Jäättelä, M. (2004). Heat Shock Protein 70 Promotes Cell Survival by Inhibiting Lysosomal Membrane Permeabilization. *The Journal of Experimental Medicine*. https://doi.org/10.1084/jem.20040531

Nys, M., Wijckmans, E., Farinha, A., Yoluk, Ö., Andersson, M., Brams, M., … Ulens, C. (2016). Allosteric binding site in a Cys-loop receptor ligand-binding domain unveiled in the crystal structure of ELIC in complex with chlorpromazine. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1603101113

Olsson, M. H. M., SØndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent treatment of internal and surface residues in empirical p Kapredictions. *Journal of Chemical Theory and Computation*. https://doi.org/10.1021/ct100578z

Pagadala, N. S., Syed, K., & Tuszynski, J. (2017). Software for molecular docking: a review. *Biophysical Reviews*. https://doi.org/10.1007/s12551-016-0247-1

Park, C.-J., & Seo, Y.-S. (2015). Heat Shock Proteins: A Review of the Molecular Chaperones for Plant Immunity. *The Plant Pathology Journal*. https://doi.org/10.5423/PPJ.RW.08.2015.0150

Patschull, A. O. M., Gooptu, B., Ashford, P., Daviter, T., & Nobeli, I. (2012). In silico assessment of potential druggable pockets on the surface of α1-antitrypsin conformers. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0036612

Patury, S., Miyata, Y., & Gestwicki, J. E. (2009). Pharmacological Targeting of the Hsp70 Chaperone. *Current Topics in Medicinal Chemistry*. https://doi.org/CTMC-Abs-024-9-15 [pii]

Penkler, D., Sensoy, Ö., Atilgan, C., & Tastan Bishop, Ö. (2017). Perturbation-Response Scanning Reveals Key Residues for Allosteric Control in Hsp70. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/acs.jcim.6b00775

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., … Schulten, K.

(2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*. https://doi.org/10.1002/jcc.20289

Pirkkala, L., Nykänen, P., & Sistonen, L. (2001). Roles of the heat shock transcription factors in regulation of the heat shock response and beyond. *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*. https://doi.org/10.1096/fj00-0294rev

Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*. https://doi.org/10.1006/jcph.1995.1039

Powers, M. V., Jones, K., Barillari, C., Westwood, I., Van Montfort, R. L. M., & Workman, P. (2010). Targeting HSP70: The second potentially druggable heat shock protein and molecular chaperone? *Cell Cycle*. https://doi.org/10.4161/cc.9.8.11204

Proudfoot, J. R. (2008). High-Throughput Screening and Drug Discovery. *The Practice of Medicinal Chemistry*, *384*(November), 144–158. https://doi.org/10.1016/B978-0-12-374194-3.00007-X

Przyborski, J. M., Diehl, M., & Blatch, G. L. (2015). Plasmodial HSP70s are functionally adapted to the malaria parasite life cycle. *Frontiers in Molecular Biosciences*. https://doi.org/10.3389/fmolb.2015.00034

Rapp, O., & Yifrach, O. (2017). Using the MWC model to describe heterotropic interactions in hemoglobin. *PloS One*. https://doi.org/10.1371/journal.pone.0182871

Roca, C., Requena, C., Sebastián-Pérez, V., Malhotra, S., Radoux, C., Pérez, C., … Campillo, N. E. (2018). Identification of new allosteric sites and modulators of AChE through computational and experimental tools. *Journal of Enzyme Inhibition and Medicinal Chemistry*. https://doi.org/10.1080/14756366.2018.1476502

Rodina, A., Patel, P. D., Kang, Y., Patel, Y., Baaklini, I., Wong, M. J. H., … Chiosis, G. (2013). Identification of an allosteric pocket on human Hsp70 reveals a mode of inhibition of this therapeutically important protein. *Chemistry and Biology*. https://doi.org/10.1016/j.chembiol.2013.10.008

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

Bioinformatics (Oxford, England), *34*(21), 3759–3763.
https://doi.org/10.1093/bioinformatics/bty427

Rudiger, S., Buchberger, A., & Bukau, B. (1997). Interaction of Hsp70 chaperones with substrates. *Nat Struct.Biol*. https://doi.org/10.1038/nsb0597-342

Samreen, Z., & Ling, J. X. (2014). *Escherichia coli as a model organism.*

Schiavone, S., & Trabace, L. (2018). Small molecules: Therapeutic application in neuropsychiatric and neurodegenerative disorders. *Molecules*, *23*(2).
https://doi.org/10.3390/molecules23020411

Schlecht, R., Scholz, S. R., Dahmen, H., Wegener, A., Sirrenberg, C., Musil, D., … Bukau, B. (2013). Functional analysis of Hsp70 inhibitors. *PLoS ONE*.
https://doi.org/10.1371/journal.pone.0078443

Schlesinger, M. J. (1990). Heat Shock Proteins. *The Journal of Biological Chemistry*.
https://doi.org/10.1016/j.virol.2010.11.005

Schmidtke, P., & Barril, X. (2010). Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*.
https://doi.org/10.1021/jm100574m

Schrödinger. (2016). Release 2016-1: Schrödinger Suite 2016-1 Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2016; Impact, Schrödinger, LLC, New York, NY, 2016; Prime.

Schulze, J. O., Saladino, G., Busschots, K., Neimanis, S., Süß, E., Odadzic, D., … Biondi, R. M. (2016). Bidirectional Allosteric Communication between the ATP-Binding Site and the Regulatory PIF Pocket in PDK1 Protein Kinase. *Cell Chemical Biology*.
https://doi.org/10.1016/j.chembiol.2016.06.017

Schüttelkopf, A. W., & Van Aalten, D. M. F. (2004). PRODRG: A tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica Section D: Biological Crystallography*. https://doi.org/10.1107/S0907444904011679

Seeliger, D., & De Groot, B. L. (2010). Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *Journal of Computer-Aided Molecular Design*.
https://doi.org/10.1007/s10822-010-9352-6

Shang, J., Lu, S., Jiang, Y., & Zhang, J. (2016). Allosteric modulators of MEK1: drug design and discovery. *Chemical Biology and Drug Design*. https://doi.org/10.1111/cbdd.12780

Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature*. https://doi.org/10.1038/nature03197

Singh, S. B., & Pettit, G. R. (1987). Isolation , Structure , and Synthesis of combretastatin A-2, A-3, and B-2. *Can J Chem*, *65*, 2390–2396. https://doi.org/10.1021/jo00278a023

Sinha, S., Medhi, B., & Sehgal, R. (2014). Challenges of drug-resistant malaria. *Parasite*. https://doi.org/10.1051/parasite/2014059

Song, K., Liu, X., Huang, W., Lu, S., Shen, Q., Zhang, L., & Zhang, J. (2017). Improved Method for the Identification and Validation of Allosteric Sites. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/acs.jcim.7b00014

Srivastava, A., Nagai, T., Srivastava, A., Miyashita, O., & Tama, F. (2018). Role of computational methods in going beyond x-ray crystallography to explore protein structure and dynamics. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms19113401

Stetz, G., & Verkhivker, G. M. (2016). Probing Allosteric Inhibition Mechanisms of the Hsp70 Chaperone Proteins Using Molecular Dynamics Simulations and Analysis of the Residue Interaction Networks. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/acs.jcim.5b00755

Sugiki, T., Furuita, K., Fujiwara, T., & Kojima, C. (2018). Current NMR techniques for structure-based drug discovery. *Molecules*. https://doi.org/10.3390/molecules23010148

Suplatov, D., & Švedas, V. (2015). Study of functional and allosteric sites in protein superfamilies. *Acta Naturae*.

Taj, M. K., Samreen, Z., Ling, J. X., Taj, I., & Yunlin, W. (2014). Escherichia coli as a Model Organism. *International Journal of Engineering Research and Science & Technology*, *3*(April), 1–10.

Teague, S. J. (2003). Implications of protein flexibility for drug discovery. *Nature Reviews Drug Discovery*. https://doi.org/10.1038/nrd1129

Tee, W. V., Guarnera, E., & Berezovsky, I. N. (2018). Reversing allosteric communication: From detecting allosteric sites to inducing and tuning targeted allosteric response. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1006228

Thayer, K. M., Lakhani, B., & Beveridge, D. L. (2017). Molecular Dynamics-Markov State Model of Protein Ligand Binding and Allostery in CRIB-PDZ: Conformational Selection and Induced Fit. *Journal of Physical Chemistry B*. https://doi.org/10.1021/acs.jpcb.7b02083

Thermodynamics, L. O. F., & Reactions, A. (n.d.). Energy, enzymes, metabolism, 1–15.

Truglio, J., Graziano, M., Vedanthan, R., Hahn, S., Rios, C., Hendel-Paterson, B., & Ripp, J. (2012). Global health and primary care: Increasing burden of chronic diseases and need for integrated training. *Mount Sinai Journal of Medicine*. https://doi.org/10.1002/msj.21327

Wagner, J. R., Lee, C. T., Durrant, J. D., Malmstrom, R. D., Feher, V. A., & Amaro, R. E. (2016). Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chemical Reviews*. https://doi.org/10.1021/acs.chemrev.5b00631

Wang, C., Greene, D., Xiao, L., Qi, R., & Luo, R. (2018). Recent Developments and Applications of the MMPBSA Method. *Frontiers in Molecular Biosciences*. https://doi.org/10.3389/fmolb.2017.00087

Win, D. T., & Budhraja, P. (2007). The IBM BlueGene (Supercomputer) Project Capability in Science Application (Molecular Dynamics and Protein Folding), *10*(4), 237–247.

Young, J. C. (2010). Mechanisms of the Hsp70 chaperone system. *Biochemistry and Cell Biology*. https://doi.org/10.1139/o09-175

Zhuravleva, A., Clerico, E. M., & Gierasch, L. M. (2012). An interdomain energetic tug-of-war creates the allosterically active state in Hsp70 molecular chaperones. *Cell*. https://doi.org/10.1016/j.cell.2012.11.002

SUPPLEMENTARY DATA

Video S1: Movement of the ligand-free structure in H_prime after 100ns MD simulation. The protein is represented as a cartoon while ATP (blue) is represented in licorice, and Mg^{2+} (blue) is represented as a ball.

Video S2: Movement of the protein-SANC00499 structure after 100ns MD simulation. The protein is represented as a cartoon, ATP (blue) is represented in licorice, Mg^{2+} (blue) is represented as a ball, and SANC00499 (lilac) is represented in licorice.

Video S3: Movement of the protein-SANC00685 structure after 100ns MD simulation. The protein is represented as a cartoon, ATP (blue) is represented in licorice, Mg^{2+} (blue) is represented as a ball, and SANC00685 (lilac) is represented in licorice.

Video S4: Movement of the protein-SANC00492 structure after 100ns MD simulation. The protein is represented as a cartoon, ATP (blue) is represented in licorice, Mg^{2+} (blue) is represented as a ball, and SANC00492 (lilac) is represented in licorice.

Video S5: Movement of the protein-SANC00357 structure after 100ns MD simulation. The protein is represented as a cartoon, ATP (blue) is represented in licorice, Mg^{2+} (blue) is represented as a ball, and SANC00357 (lilac) is represented in licorice.

Video S6: Movement of the ligand-free structure in B_prime after 100ns MD simulation. The protein is represented as a cartoon, ATP (blue) is represented in licorice, and Mg^{2+} (blue) is represented as a ball.

Video S7: Movement of the protein-SANC00284 structure after 100ns MD simulation. The protein is represented as a cartoon, ATP (blue) is represented in licorice, Mg^{2+} (blue) is represented as a ball, and SANC00284 (lilac) is represented in licorice.

Video S8: Movement of the protein-SANC00477 structure after 100ns MD simulation. The protein is represented as a cartoon, ATP (blue) is represented in licorice, Mg^{2+} (blue) is represented as a ball, and SANC00477 (lilac) is represented in licorice.

Video S9: Movement of the protein-SANC00556 structure after 100ns MD simulation. The protein is represented as a cartoon, ATP (blue) is represented in licorice, Mg^{2+} (blue) is represented as a ball, and SANC00556 (lilac) is represented in licorice.

Video S10: Movement of the protein-SANC00676 structure after 100ns MD simulation. The protein is represented as a cartoon, ATP (blue) is represented in licorice, Mg^{2+} (blue) is represented as a ball, and SANC00676 (lilac) is represented in licorice.

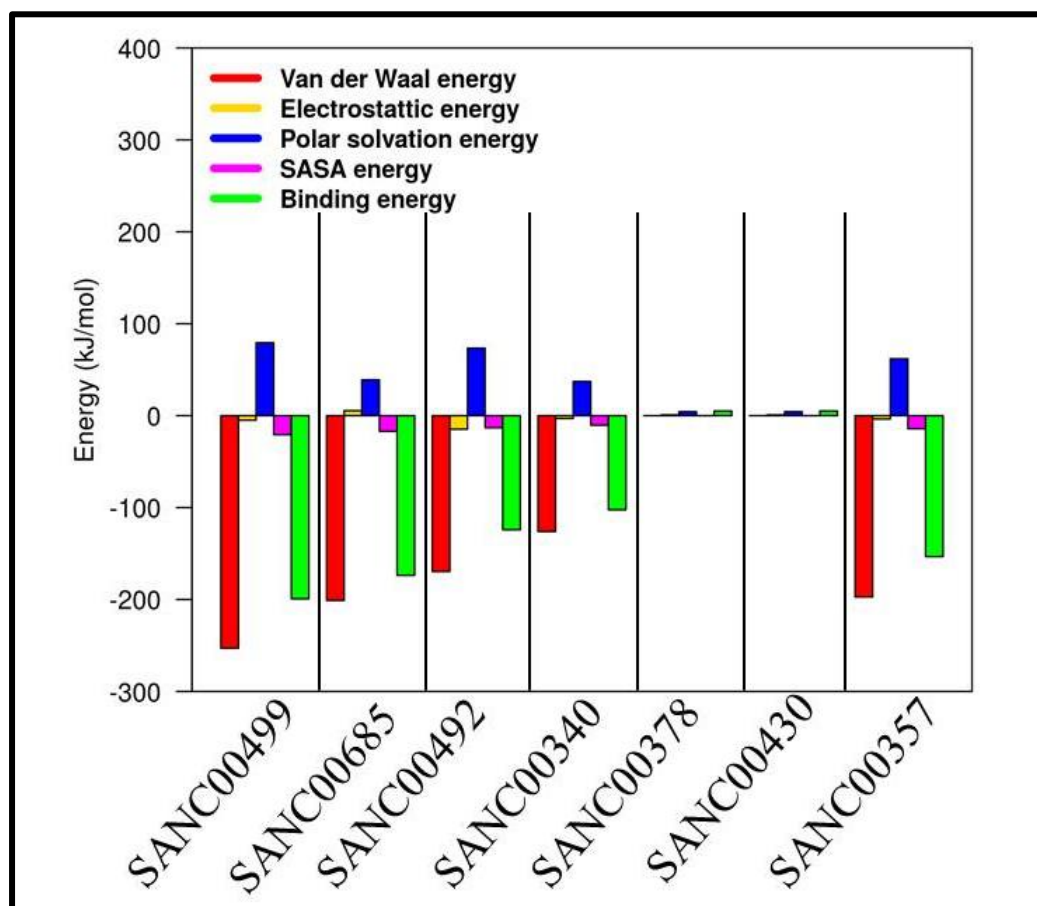**Figure S1:** Binding free energy components on H_prime in complex with respective ligands

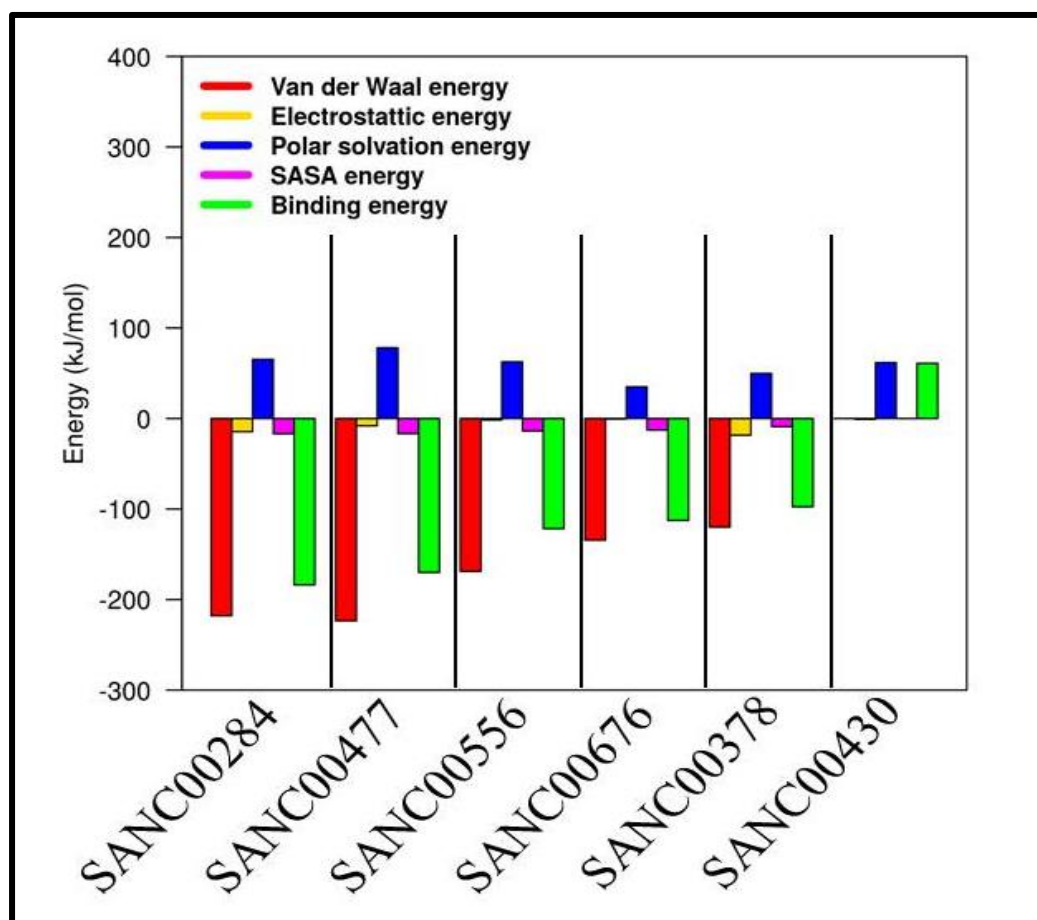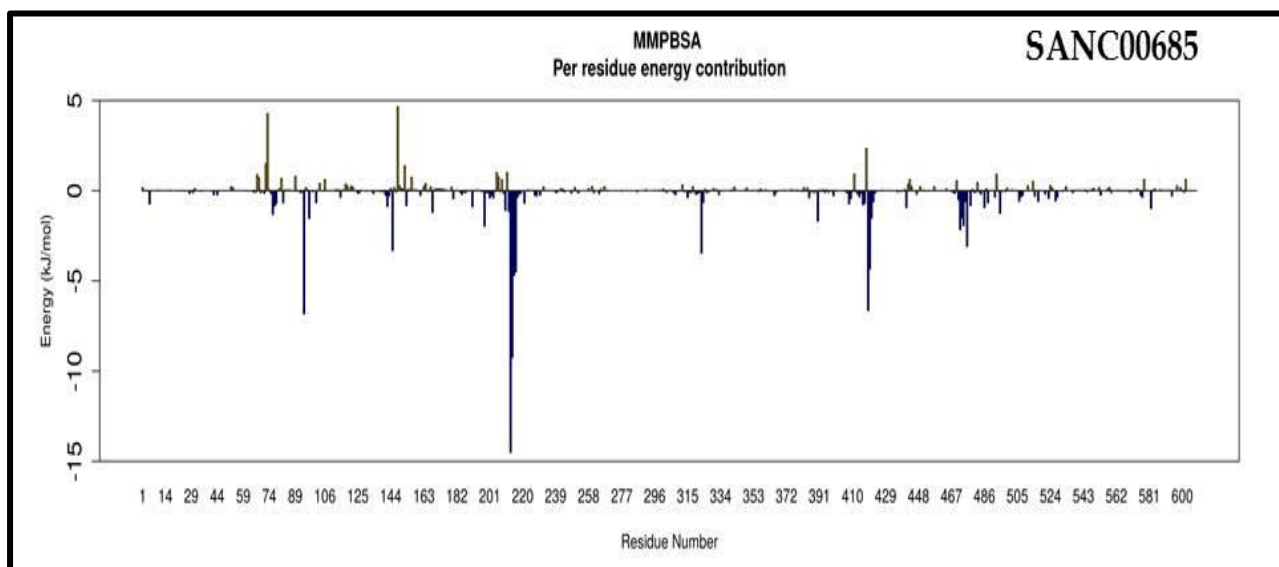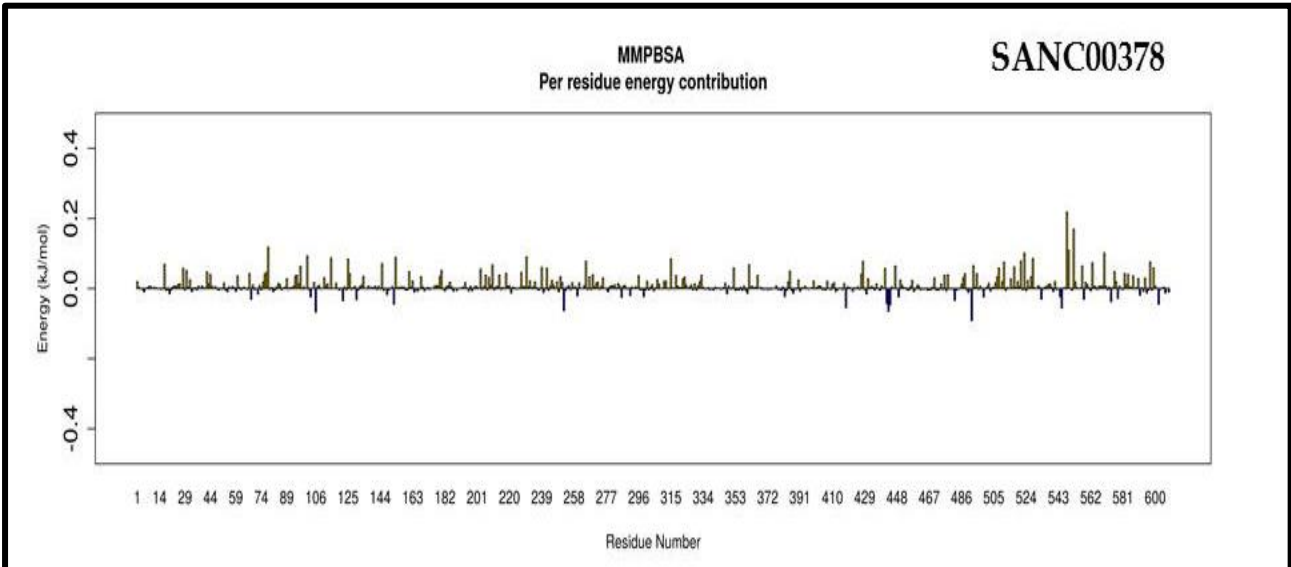**Figure S1:** Binding free energy components on B_prime in complex with respective ligands
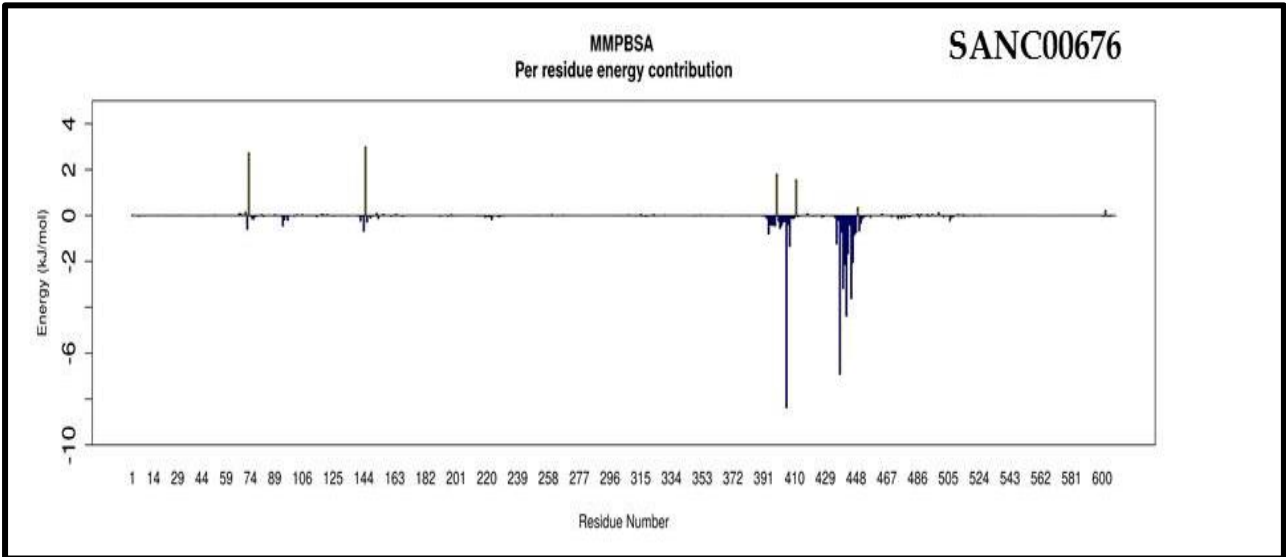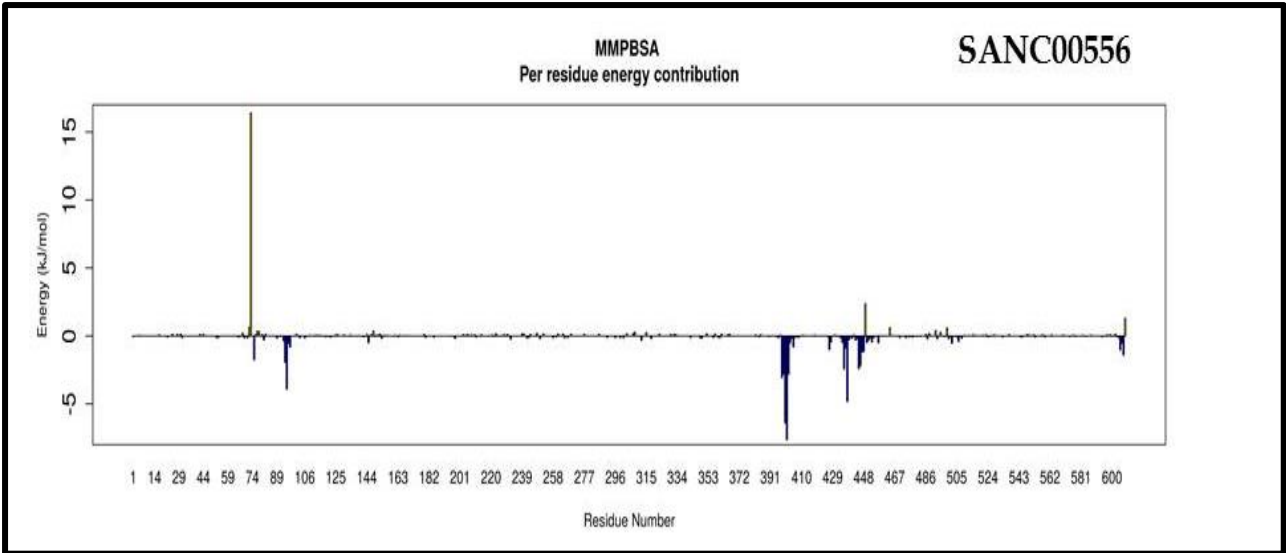
Figure S3: Bar plot showing per residue contribution to the total binding free energy of protein-ligand bound complexes in H_prime

104

MMPBSA
Per residue energy contribution

SANC00430

MMPBSA
Per residue energy contribution

SANC00357
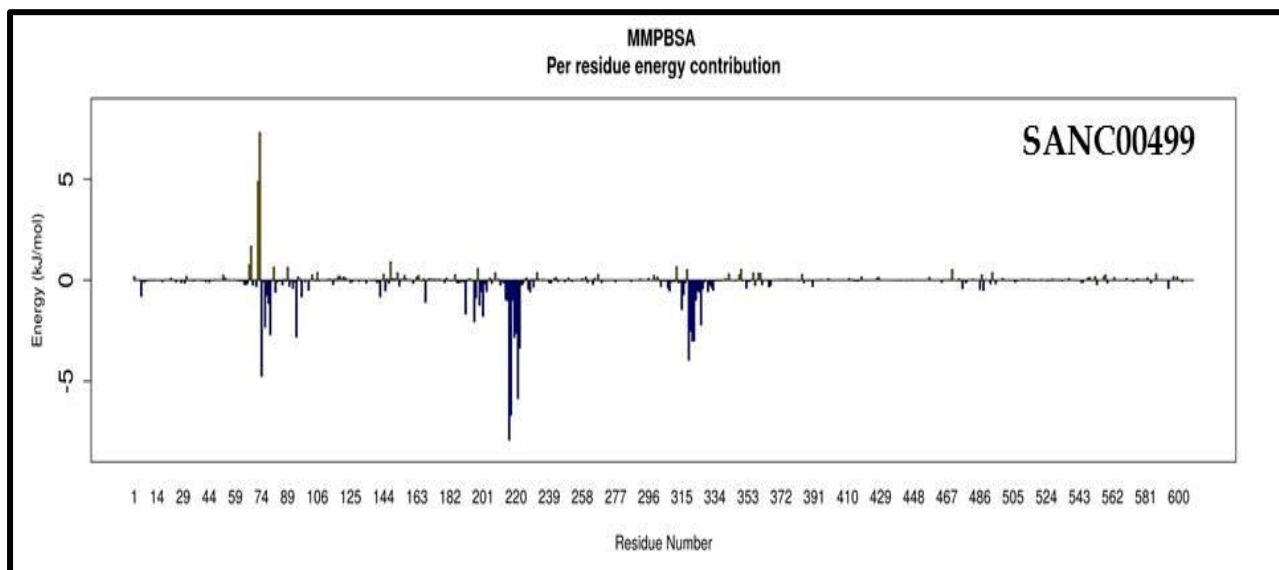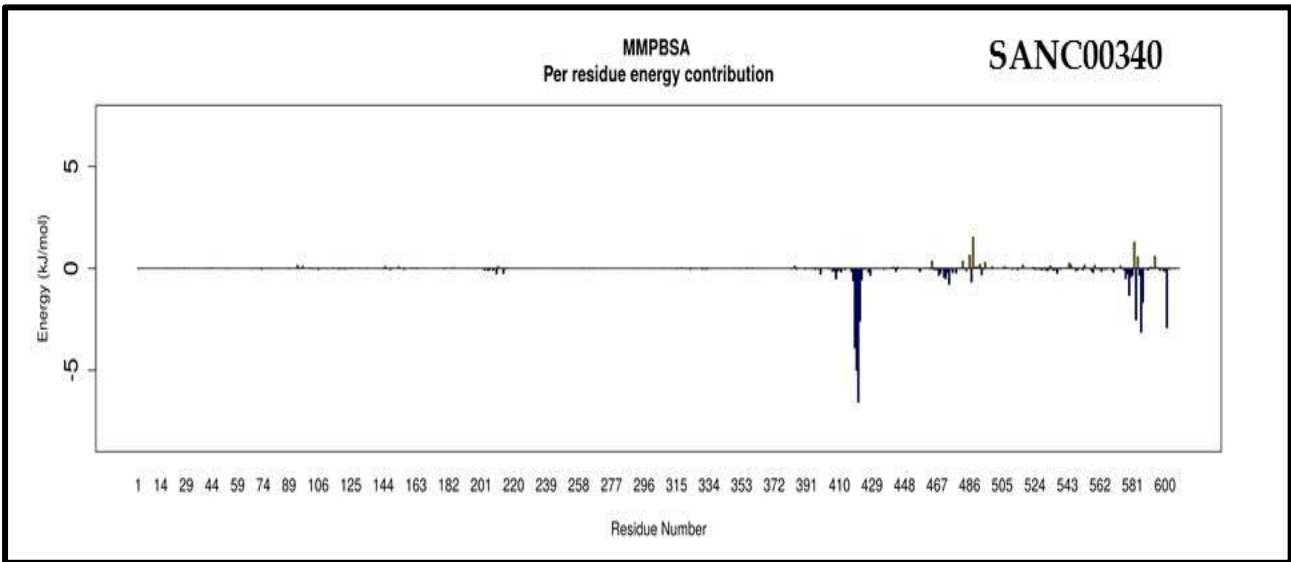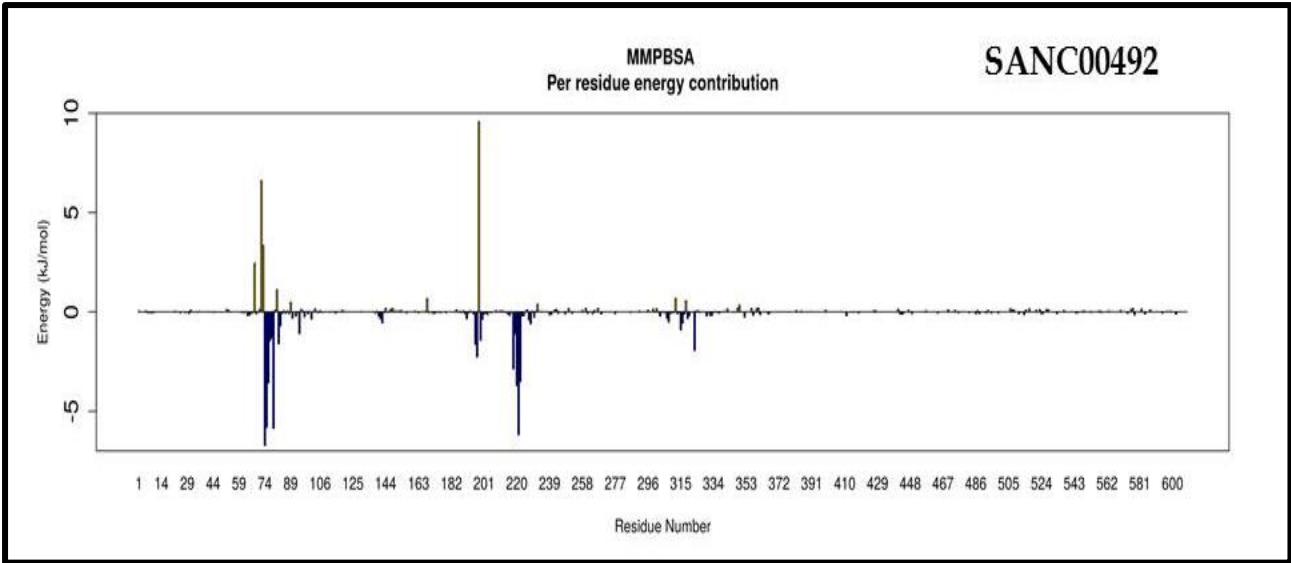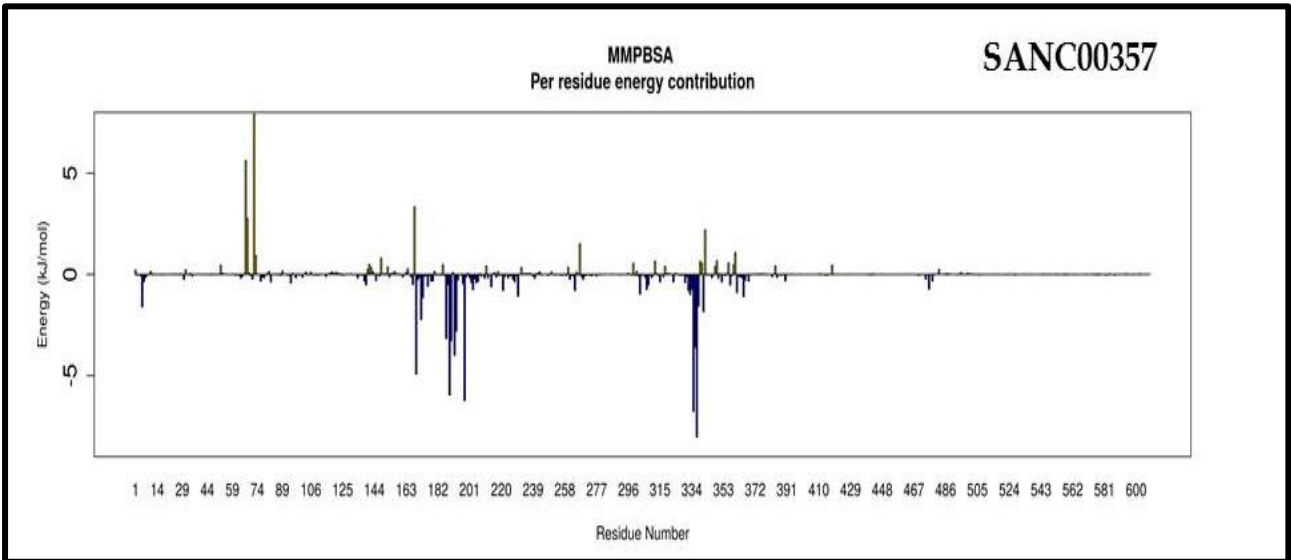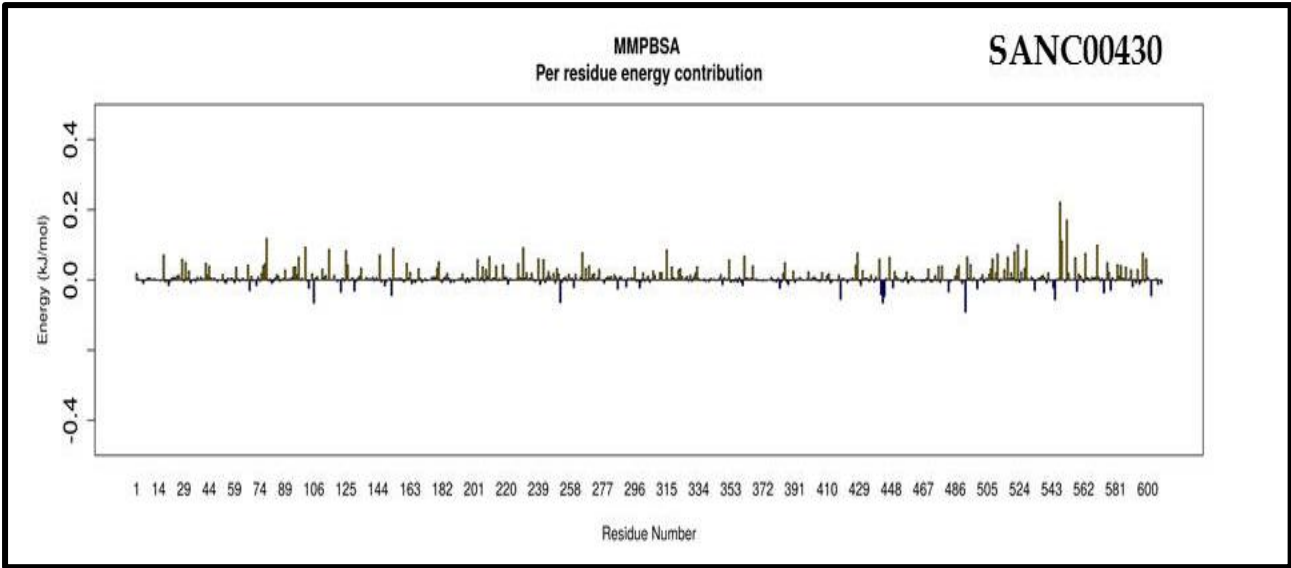
Figure S4: Bar plot showing per residue contribution to the total binding free energy of protein-ligand bound complexes in B_prime

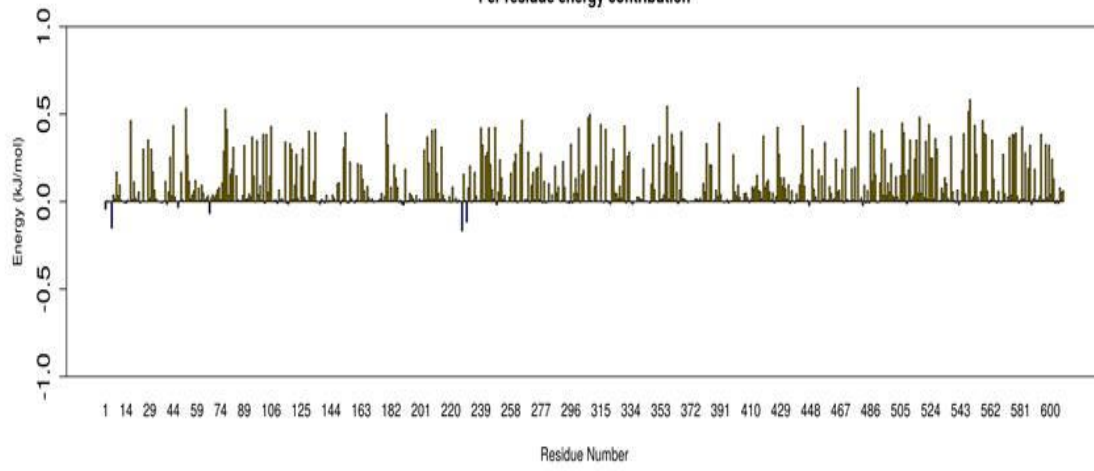Per residue energy contribution chart titled "MMPBSA" with "SANC00430" label. X-axis: Residue Number, Y-axis: Energy (kJ/mol).
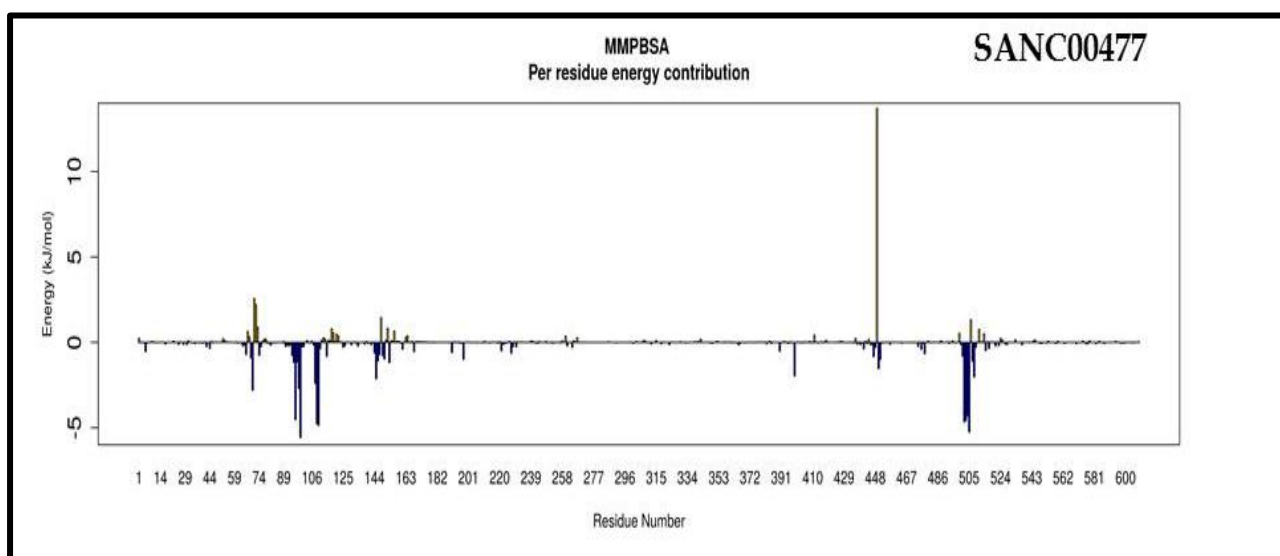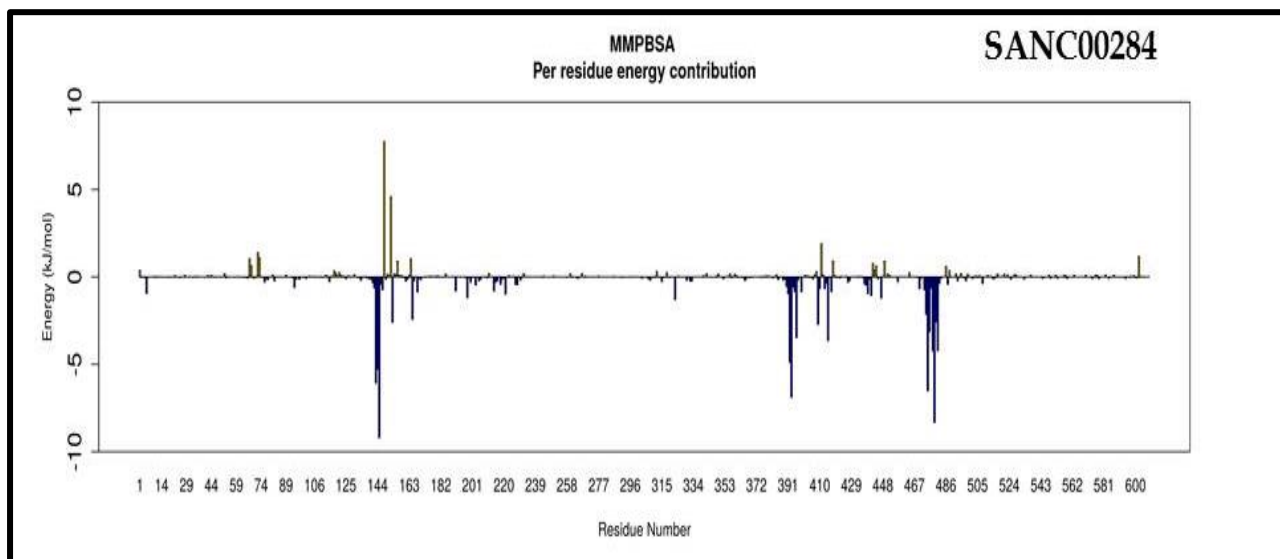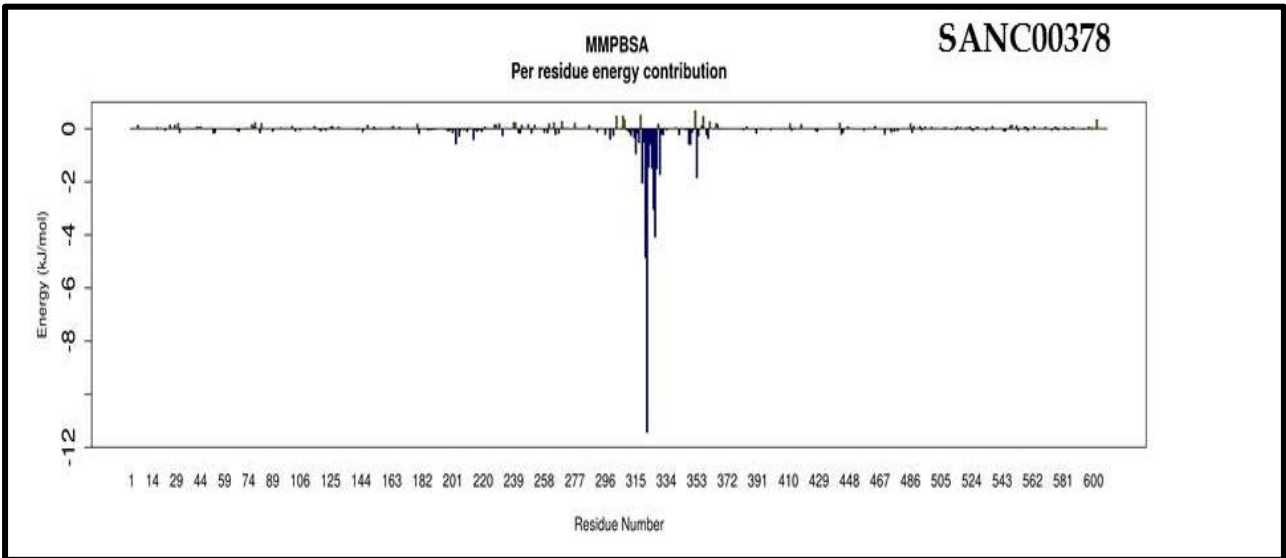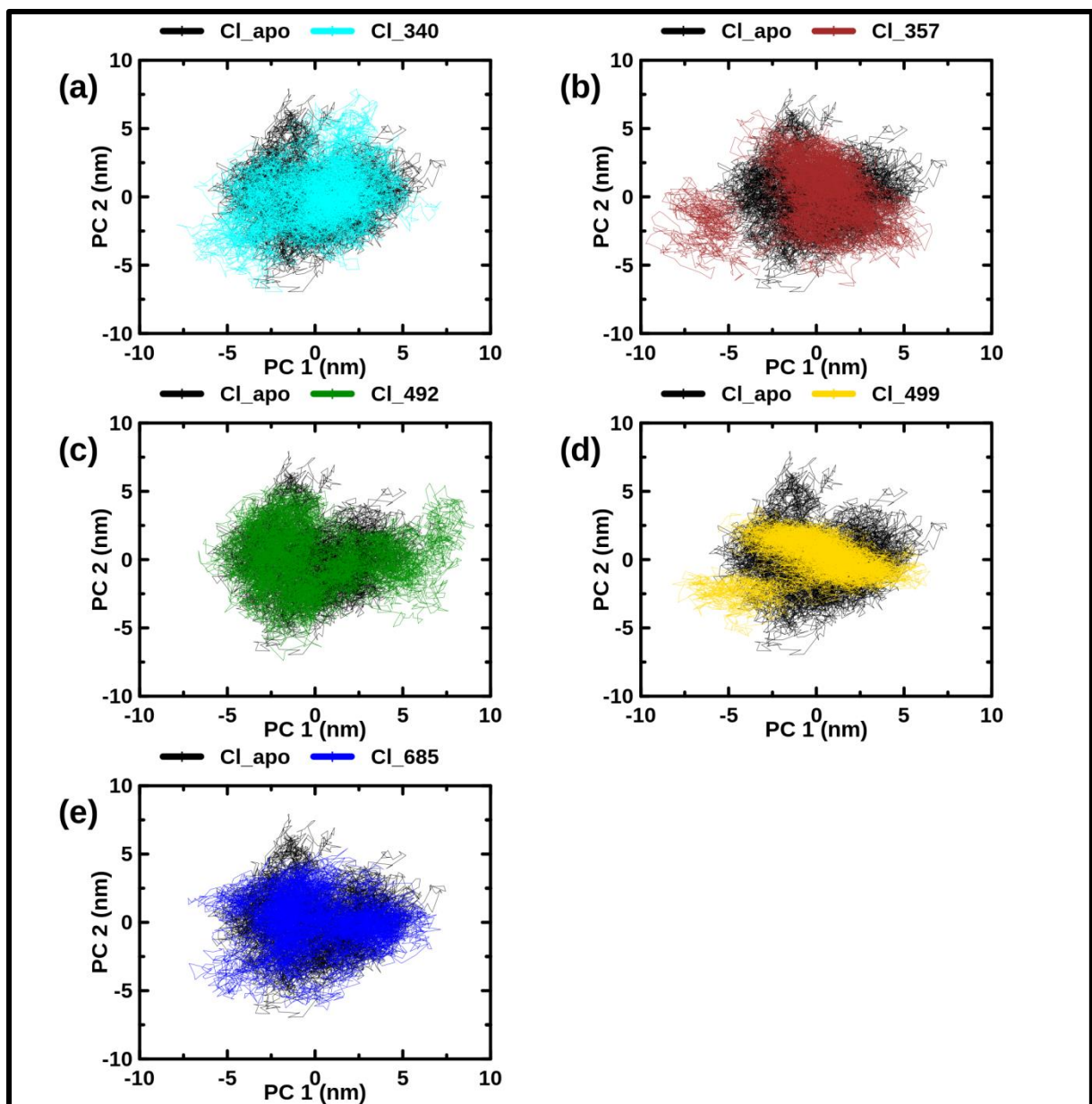
Figure S5: PCA of H_prime plotted on R

Figure S6: PCA of B_prime plotted on R