In silico analysis of human Hsp90 for the identification of novel anti-cancer drug target sites and natural compound inhibitors

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

of

MASTER OF SCIENCE OF RHODES UNIVERSITY

by

Coursework / Thesis

in

Bioinformatics and Computational Molecular Biology

in the Department of Biochemistry and Microbiology

Faculty of Science

by

David Lawrence Penkler

July 2015

ABSTRACT

The 90-KDa heat shock protein (Hsp90) is part of the molecular chaperone family, and as such it is involved in the regulation of protein homeostasis within cells. Specifically, Hsp90 aids in the folding of nascent proteins and re-folding of denatured proteins. It also plays an important role in the prevention of protein aggregation. Hsp90's functionality is attributed to its several staged, multi-conformational ATPase cycle, in which associated client proteins are bound and released. Hsp90 is known to be associated with a wide array of client proteins, some of which are thought to be involved in multiple oncogenic processes. Indeed Hsp90 is known to be directly involved in perpetuating the stability and function of multiple mutated, chimeric and over-expressed signalling proteins that are known to promote the growth and survival of cancer cells. Hsp90 inhibitors are thus thought to be promising therapeutic agents for cancer treatment. A lack of a 3D structure of human Hsp90 however has restricted Hsp90 inhibitor development in large to *in vivo* investigations. This study, aims to investigate and calculate hypothetical homology models of the full human Hsp90 protein, and to probe these structural models for novel drug target sites using several *in silico* techniques. A multi-template homology modelling methodology was developed and in conjunction with protein-protein docking techniques, two functionally important human Hsp90 structural models were calculated; the nucleotide free "v-like" open and nucleotide bound closed conformations. Based on the conservation of ligand binding, virtual screening experiments conducted on both models using 316 natural compounds indigenous to South Africa, revealed three novel putative target sites. Two binding pockets in close association with important Hsp90-Hop interaction residues and a single binding pocket on the dimerization interface in the C-terminal domain. Targeted molecular docking experiments at these sites revealed two compounds (721395-11-5 and 264624-39-7) as putative inhibitors, both showing strong binding affinities for at least one of the three investigated target sites. Furthermore both compounds were found to only violate one Lipinski's rules, suggesting their potential as candidates for further drug development. The combined work described here provides a putative platform for the development of next generation inhibitors of human Hsp90.

DECLARATION

The research described in this thesis was carried out as part of the one-year MSc coursework and research thesis programme in Bioinformatics and Computational Molecular Biology, from 15 July 2014 to 31 January 2015 under the supervision of Prof Özlem Taştan Bishop.

I, David Lawrence Penkler, declare that this thesis submitted to Rhodes University is wholly my own work and has not previously been submitted for a degree at this or any other institution.

Signature ……..

Date …….

RESEARCH OUTPUTS

Parts of the research described in this thesis have been published or submitted for peer review in the following:

David L. Penkler and Özlem Tastan Bishop. ***In silico* analysis of human Hsp90 for the identification of novel drug target sites.** *South African Society for Bioinformatics and South African Genetics Society Joint Congress* (2014) Poster presentation P1-6, pg. 44

Rowan Hatherley, David K. Brown, Thommas M. Musyoka, David L. Penkler, Ngonidzashe Faya, KevinA. Lobb and Özlem Tastan Bishop. 2015. **SANCDB a South African Natural Compound Database.** Journal of Cheminformatics 1(7) pp29 doi: 10.1186/s13321-015-0080-8

David K. Brown, Thommas M. Musyoka, David L. Penkler and Özlem Tastan Bishop. 2015. **JMS: A workflow management system and web-based cluster front-end for the Torgue resource manager**. arXiv preprint arXiv:1501.06907

DEDICATION

This thesis is dedicated to my Omi Rose Penkler

The source of much of my inspiration and motivation.

~ Despite the odds no challenge is ever too big ~

ACKNOWLEDGEMENTS

Prof. Özlem Tastan Bishop, for recognizing my potential and doing everything within her power to help me achieve it. No student could ask for a better supervisor. Thank you.

Rowan Hatherley, for all his assistance and advice with the homology modelling. Thank you for the many interesting discussions and for allowing me the opportunity to bounce ideas and general musings off you.

David Brown, for all his help in familiarizing me with the server and cluster. I have learnt much from you over the past year and your patience in helping me understand the concept of parallelization was an added bonus.

Thommas Musyoka, thank you big man for all the ideas, encouragement and many laughs.

My dearest mother Carol for her endless support, love and guidance. You are an absolute blessing.

My loving girlfriend and partner Jo-Anne Laurence. Your energetic approach to life and its many challenges has been my most constant motivator. Without your daily love and support, this degree may just have beaten me. I'm lucky to have you in my life.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. The opinions expressed and derived conclusions are those of the author and are not necessarily to be attributed to the NRF.

I also acknowledge Rhodes University for providing additional funding through the Prestigious Henderson Scholarship.

TABLE OF CONTENTS

Abstract i

Declaration ii

Research outputs iii

Dedication iv

Acknowledgements v

Table of Contents vi

List of figures xi

List of tables xiii

List of web servers and applications xiv

List of Abbreviations xv

Chapter 1: Introduction 1

1.1 Overview of cancer and current treatments 1

1.2 The molecular chaperone family 2

1.3 Characterization of the 90-KDa molecular chaperone 3

1.4 Structure and ATPase cycle of Hsp90 4

1.5 Hsp90's ATPase cycle 6

1.6 The intracellular function of Hsp90 7

1.7 The extracellular function of Hsp90 8

1.8 Hsp90 and its link to cancer 9

1.9 Current Hsp90 anti-cancer therapies 10

1.10 *In silico* drug discovery and computational studies 11

1.11 Project motivation 12
1.11.1 Problem statement and knowledge gap 12
1.11.2 Aims and objectives 13
Chapter 2: Homology Modelling 14
2.1 Introduction 14
2.2 Homology modelling process 15
2.2.1 Template identification 16
2.2.2 Target – Template alignment 17
2.2.3 Model building 18
2.2.4 Model evaluation 18
2.3 Protein – protein docking 20
2.4 Energy Minimization - GROMACS 20
2.5 chapter objectives 21
2.6 Methodology 22
2.6.1 Template identification and retrieval 22
2.6.2 Structural sequence alignment 22
2.6.3 Model building 22
2.6.4 Model validation 23
2.6.5 Protein-protein docking 23
2.6.6 Energy minimization 23
2.7 Results and discussion 24

2.7.1 Template identification and retrieval 24

2.7.2 Alignment file construction and model building 30

2.7.2.1 Construction of closed conformation alignment file 30

2.7.2.2 Model building and evaluation of Hsp90 in closed conformation 32

2.7.2.3 Construction of open conformation alignment file 34

2.7.2.4 Model building and evaluation of single open conformation monomer 36

2.7.2.5 Protein – Protein docking of single Hsp90 monomers 37

2.7.3 Energy minimization 37

2.8 Chapter conclusions 40

Chapter 3: Molecular Docking 44

3.1 Introduction 44

3.2 The thermodynamics of protein-ligand binding 44

3.3 Types of protein-ligand interactions and other contributing factors 46

3.3.1 Electrostatic interactions 46

3.3.2 Hydrophobic interactions 47

3.3.3 Protein binding site water molecules 48

3.3.4 Solvation and desolvation 48

3.3.5 Protein flexibility 49

3.4 Prediction of protein-ligand interactions 49

3.4.1 Free energy calculations 49

3.4.2 MM-PBSA and MM-GBSA methods 50

3.4.3 Docking and scoring 50

3.4.3.1 Empirical scoring....51
3.4.3.2 Knowledge-based scoring....51
3.4.3.3 Force-field scoring....52
3.5 Chapter Objectives....52
3.6 Methodology....53
3.6.1 Construction of mini South African compound database....53
3.6.2 Molecular docking procedures....54
3.6.2.1 Blind docking parameters....55
3.6.2.2 Targeted docking parameters....55
3.6.3 Control docking procedure....55
3.6.4 Docking analysis and scoring....56
3.6.4.1 Blind docking analysis....56
3.6.4.2 Targeted docking analysis and scoring....56
3.7 Results and discussion....57
3.7.1 South African natural compounds database....57
3.7.2 Control study for the validation of blind docking procedure....57
3.7.3 Whole protein virtual screening....59
3.7.3.1 Closed conformation screening and target site identification....61
3.7.3.2 Open conformation screening and target site identification....62
3.7.4 Targeted docking studies....62
3.7.4.1 Targeting Hop-Hsp90 interacting residues....63
3.7.4.1.1 Quantitative analysis of target site 1....65
3.7.4.1.2 Quantitative analysis of target site 2....69

3.7.4.2 Targeting the CTD dimerization site 71
3.7.5 Evaluation of putative lead compounds 73
3.8 Chapter conclusions 75
Chapter 4: Concluding remarks 76
4.1 Homology modelling of human Hsp90 76
4.2 Structure-based targeted drug discovery 77
4.3 Future work and prospects 78
References 80
Appendix A 92
A-1: Example PIR file format used in homology model building 92
A-2: Example of MODELLER script used for executing model building 97
Appendix B 98
B-1: Example of docking parameter file used for AutoDock4 runs 98
B-2: X-score parameter file 99
Appendix c 101
C-1: LigPlot+ 2D interaction maps for all ligands bound at Target site 1 101
C-2: LigPlot+ 2D interaction maps for all ligands bound at Target site 2 105
C-3: LigPlot+ 2D interaction maps for all ligands bound at Target site 3 110

LIST OF FIGURES

Figure 1-1: Schematic representation of Hsp90's three distinct structural domains 4

Figure 1-2: 3D rendering of N-terminal of Hsp90.....5

Figure 1-3: Graphical representation of the ATP dependent binding mechanism 6

Figure 1-4: Schematic demonstrating Hsp90's functional structural life cycle. 7

Figure 2-1: Schematic depicting the homology modelling process.....16

Figure 2-2: The structure of yeast Hsp90 (2CG9) 25

Figure 2-3: Target-Template structural multiple sequence alignment 28

Figure 2-4: Graphical summary of templates used 29

Figure 2-5: Graphical representation of homodimer PIR alignment file.....31

Figure 2-6: Knotting observed during model building process 32

Figure 2-7: Graphical representation of the normalized DOPE-Z scores 33

Figure 2-8: Conformational differences in N-terminal domain.....35

Figure 2-9: Graphical representation of monomer PIR alignment file.....35

Figure 2-10: MetaMQAPII analysis of model 44.....36

Figure 2-11: Energy potentials of conformation models.....38

Figure 2-12: MetaMQAPII representation of closed conformation models 39

Figure 2-13: Homology model of human Hsp90 in closed conformation 40

Figure 2-14: Homology model of human Hsp90 in open conformation 41

Figure 3-1: Schematic representation main non-bonded interaction types.....47

Figure 3-2: Schematic overview of protein-ligand interactions 48

Figure 3-3: Overview of stage requirements for compound database addition.....53

Figure 3-4: Overview of docking preparation with AutoDock4.....54

Figure 3-5: Blind docking of inhibitor ACP against the NTD ATP binding pocket.....58

Figure 3-6: Comparison of LigPlot+ 2D interaction maps.....59

Figure 3-7: Schematic representation of the simulation space used in blind docking 60

Figure 3-8: Whole protein screening Hsp90 homology models.....61

Figure 3-9: Schematic overview of target sites 1 and 2 64

Figure 3-10: Flow diagram summarizing the best ligand filtering procedure 65

Figure 3-11: Graphical representation of LigPlot+ analysis for target site 1.....66

Figure 3-12: Graphical representation of empirical binding free energy scores 67

Figure 3-13: Schematic representations of target site 1 bound by ligand 721395-11-5 68

Figure 3-14: Graphical representation of LigPlot+ analysis for target site 2.....69

Figure 3-15: Binding groove of target site 2 with bound ligand 264624-39-7 70

Figure 3-16: CTD dimerization target site 3.....71

Figure 3-17: Graphical representation of LigPlot+ analysis for target site 3.....72

Figure 3-18: The CTD dimerization site bound by ligand 264624-39-7.....73

LIST OF TABLES

Table 1-1: Mammalian chaperone families and their known functions 3

Table 1-2: The possible functions of different extracellular Hsp90 8

Table 2-1: Detailed summary of the protein structure templates 26

Table 2-2: DOPE-Z and MetaMQAPII evaluation scores for closed conformation model 34

Table 2-3: DOPE-Z and MetaMQAPII evaluation scores for open conformation model 36

Table 2-4: Summary of ClusPro protein-protein docking results 37

Table 3-1: Summary of all gridbox parameters used 63

Table 3-2: Summary of analysis of best scoring ligands according to Lipinski's rule of 5 74

LIST OF WEB SERVERS AND APPLICATIONS

1. Genesilico online server

   https://genesilico.pl/toolkit/

2. HHpred online server

   http://toolkit.tuebingen.mpg.de/hhpred

3. NCBI BLAST

   http://blast.ncbi.nlm.nih.gov/

4. PROMALS 3D

   http://prodata.swmed.edu/promals3d/promals3d.php

5. RCSB

   http://www.rcsb.org/pdb/home/home.do

6. SciFinder

   https://scifinder.cas.org/scifinder/view/scifinder/scifinderExplore.jsf

LIST OF ABBREVIATIONS

ADT	Autodock tools
ATP	Adenosine tri-phosphate
ADP	Adenosine di-phosphate
CTD	C-terminal domain
DOPE	Discrete Optimized Protein Energy
GDT	Global distance test
GDT_TS	Global distance test total scores
GROMACS	GROningen MAchine for Chemical Simulations
Hsp	Heat shock protein
KDa	Kilo-Daltons
LRT	Linear Response Theory
M-domain	Middle domain
MD	Molecular dynamics
MM-GBSA	Molecular Mechanics with Generalized Born and Surface Area
MM-PBSA	Molecular Mechanics with Poisson-Boltzmann and Surface Area
MSA	Multiple Sequence Alignment
MW	Molecular weight
NTD	N-terminal domain
PDB	Protein Data Bank
PRS	Perturbation-Response Scanning
RMSD	Root-mean-square deviation

T1	Target site 1
T2	Target site 2
T3	Target site 3
3D	Three-dimensional

CHAPTER 1: INTRODUCTION

1.1 Overview of cancer and current treatments

Simply defined, cancer is a disease that is characterised by the division of tissue cells outside the limitations of regulated cell division (Lundgren et al. 2007). This uncontrolled cell division is thought to arise in normal cells when programmed cell death (apoptosis) is lost. Typically these rapidly dividing cells originate from the same location and come together to form tissue masses, commonly referred to as tumours. A tumour can be further classified as being either benign or malignant. The former describes a tumour with limited growth and a localisation restricted to one area and one tissue type. A malignant tumour on the other hand is when cancerous tissue cells move throughout the body via the blood or lymph systems and invade other healthy tissue cells, a process called metastasis. This transformation, invasion and metastasis of tumours is heavily dependent on extracellular and intracellular cell signalling transduction pathways (Kumar & Weaver 2009).

Cancer has been reported to be the second most common cause of death in the United States, exceeded only by heart disease, and can be held accountable for 1 in every 4 deaths (American Cancer Society 2015). In 2015, approximately 589 430 Americans are expected to die of cancer alone, taking the daily death rate due to cancer to approximately 1620 people per day (American Cancer Society 2015). There are over 100 different types of cancer recorded today, each being classified according to the type of tissue cell it affects (Langton et al. 2014). A report by the American Cancer Society (2015), has estimated that approximately 1 658 370 new cancer cases are expected to be diagnosed in 2015.

The treatment of cancer is very much dependent on the type of cancer. If metastasis has not occurred, it is possible to surgically remove the cancerous cells from the body however in the case of malignant cancers, treatment becomes much more complex and includes, radiotherapy, chemotherapy, immunotherapy and hormone therapy (Langton et al. 2014). Unfortunately there is no one treatment for cancer and patients often require a combination of therapies, however all treatments ultimately look to prevent rapid cell division, by disruption or destruction of cell signalling pathways (Langton et al. 2014). Although current

statistics indicate that the current cancer survival rate is 68%, up from 49% recorded in 1977, treatment of cancer remains complex (American Cancer Society 2015). While current therapies have shown promising results, the challenge of only targeting cancer cells and not healthy cells, remains a current and serious drawback. This being said, a perfect cancer treatment that targets cancer cells alone is a much sought after ideal and a hot topic in medical and clinical research areas. Currently, one of these research areas includes the targeting of biologically important proteins known as molecular chaperones. This highly conserved set of proteins are known to be associated with a multitude of oncogenic proteins, an association that ultimately aids in the progression and metastasis of several related cancers. The following sections in this chapter, briefly describe the molecular chaperones and in particular the 90 KDa heat shock protein, the subject of this thesis.

1.2 The molecular chaperone family

The term molecular chaperone was first identified by Laskey in 1978 and were described as a set of highly conserved proteins that were only expressed when induced by various types of cellular stress. Chaperones have since been defined as proteins that have strong cyto-protective properties and as such their main function is the stabilising of non-native protein conformers by binding (Hartl 1996). Chaperones are highly regarded as essential "house-keeping" proteins and work together to facilitate in cellular processes such as protein folding, oligomeric assembly, degradation and subcellular transportation of unstable protein conformers (Koga *et al.,* 2009). The biological importance of chaperones is highlighted by their vast abundance (1-2% of cytosolic proteins) in almost every cell type (Buchner 1996), and are found in plants, prokaryotes, eukaryotes and fungi (Gupta 1995). Being ubiquitous and highly conserved, Csermely *et al.,* 1997 suggested that chaperones may have played a major role in the evolution of modern day enzymes.

Although chaperones are required for the duration of a cell's life time, it is during times of cellular stress that the role of chaperones becomes most essential. Cellular stress whether it be heat-shock, poisoning, acidosis or any abrupt change to a cells immediate environment, leads to protein denaturation, protein miss-folding and/or an array of cell signalling errors (Csermely *et al.,* 1998). A cell's recovery mechanism after such an event is the induction of chaperone synthesis, which coordinate the restoration of cellular homeostasis. It is not

surprising therefore that these essential “cellular paramedics” are referred to as heat-shock or stress proteins (Whitesell & Lindquist 2005).

The understanding of cellular function and characterisation of molecular chaperones is still somewhat limited, however two decades ago the knowledge base was so little that chaperone classification was based on the respective molecular weights of identified chaperones (Csermely *et al.,* 1998) , this classification still holds true today. There are five classified mammalian heat-shock proteins (Hsps) currently known today; 100-kDa, 90-kDa, 70-kDa, 60-kDa (Table 1) and the small Hsp families. Each family has members that are expressed either constitutively or are regulated inductively and each is targeted to different subcellular compartments (Schmitt *et al.,* 2007).

Table 1-1: Mammalian chaperone families and their known functions (Schmitt *et al.,* 2007)

Common Names	Characteristic function
Hsp27, crystallins, small heat-shock proteins	· Prevent protein aggregation · Release proteins from aggregates
Hsp60, chaperonins	· Prevent protein aggregation · Help with protein folding
Hsp70, grp78, Bip	· Prevent protein aggregation · Help with protein folding
Hsp90, grp94	· Prevent protein aggregation
Hsp110	· Release proteins from aggregates

As reviewed by Csermely *et al.,* (1998), the 90-kDa family of chaperones appear to be the most passive of all the mammalian chaperones, as they are considered to only prevent protein aggregation, and aid in guiding protein folding (Welch & Brown 1996). Indeed Young *et al.,* (1997) showed that Hsp90 can only aid in protein refolding in conjunction with other co-chaperones (see Section 1.4). The prokaryote Hsp90 otherwise known as HtpG has been reported to be dispensable (Bardwell & Craig 1988) but mammalian Hsp90 is considered to be essential to cell survival in eukaryotes (Buchner 1996).

1.3 Characterization of the 90-KDa molecular chaperone

The 90-kDa chaperone family is comprised of the 90-kDa heat-shock protein (Hsp90) and the 94-kDa glucose regulated protein (grp94) as well as the mitochondrial TNF receptor-associated protein (Trap1). While grp94 is restricted to the endoplasmic reticulum, Hsp90 is

mostly cytosolic. Hsp90 and grp94 share a 50% sequence identity (Gupta 1995), and their parallel existence is thought to be a result of a gene duplication event that would have occurred at a very early stage in the evolution of the eukaryotic cell (Csermely *et al.,* 1998).

Hsp90 differs from its grp94 counterpart in that it has two distinct isoforms Hsp90α and Hsp90β. These two isoforms share 76% sequence identity and are also thought to be a result of gene duplication (Moore *et al.,* 1989). Biochemical studies by Moore *et al.,* (1989) indicated two distinct differences between the two; firstly, Hsp90β is slightly larger than Hsp90α being 86-kDa and 84-kDa respectively; secondly, Hsp90β is considerably less inducible than its Hsp90α counterpart, and is often referred to as hsc90, as it is the constitutively expressed cognate of the 90-kDa family.

1.4 Structure and ATPase cycle of Hsp90

The structure of Hsp90 was first described as being asymmetrical and oblong in shape, and it was only after further analysis in the late 80's and early 90's that it was found to be a phosphorylated dimer (Rose *et al.,* 1987; Minami *et al.,* 1991), with 2-3 covalently bound phosphate molecules per monomer(Iannotti et al. 1988; Goetz et al. 2003). A few year later Minami *et al.,* (1994) revealed that the dimerization of the protein was essential for its functionality as a chaperone and that although considered to be a homodimer, occurrences of monomeric forms were observed albeit rarely. Hsp90 is considered to be largely hydrophobic and its hydrophobicity increases after heat-shock (Yamamoto *et al.,* 1991). Detailed biochemical and electron microscopy studies revealed that Hsp90 could be divided up into three distinct structural regions (Figure 1-1), namely the nucleotide binding N-terminal domain (NTD), a flexible middle segment coupled to a highly charged linker region, and a C-terminal domain (CTD) which provides the site for dimerization (Binart *et al.,* 1989).

Figure 1-1: Schematic representation of Hsp90's three distinct structural domains of a single monomer. The N-terminal domain (green), the middle domain and linker region (blue) and the C-terminal domain (yellow)

Early studies showed that Hsp90 not only contains an ATP-binding domain, but also has the ability to phosphorylate itself. This nucleotide binding domain was later confirmed after the 3D-structure of Hsp90's N-terminus was solved by Prodromou *et al.,* (1997) and Stebbins *et al.,* (1997), the former analysing Hsp90 from yeast and the later that of humans (Figure 1-2). The two structures were found to be almost identical both being comprised of an 8 stranded β-sheet, of which one face is covered by several α-helices, which form the deep nucleotide binding pocket near the center. The NTD was shown to be a highly conserved region of the protein over several species (Gupta 1995).

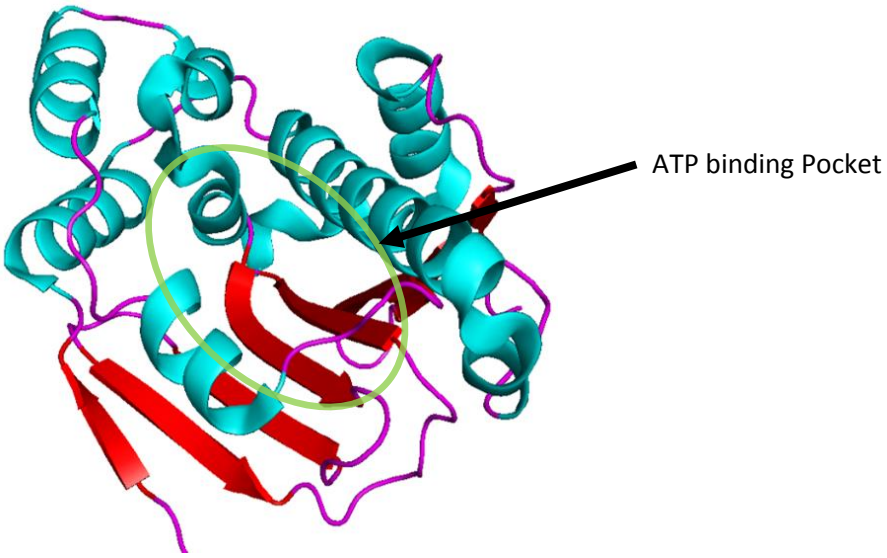


Figure 1-2: 3D rendering of N-terminal ATP binding domain of Hsp90 (PDB 3T10) showing its secondary structure constituents, beta sheet (blue), and alpha helices (red). The ATP binding pocket is shown and labelled.

The C-terminal domain (CTD), as previously mentioned provides the site for dimerization (Harris *et al.,* 2004), and also contains an -MEEVD motif which is thought to be involved in protein-protein interactions, specifically between Hsp90 and the co-chaperone Hop (Chen *et al.,* 1998). The NTD and CTD are joined by a middle 55-60 residue loop segment which is referred to as the hinge region or M domain, a name attributed by its high degree of flexibility (Toft 1998). Gupta (1995) showed the region to be specific to eukaryotes. This domain highly charged with a 64% composition of acidic amino acids and as such is thought to be involved in protein interactions particularly hormone receptors (Tbarka *et al.,* 1993). The highly packed region also includes two serine phosphorylation sites in humans (Lees-Miller & Anderson 1989) as well as a bipartite nuclear localization sequence (Nardai *et al.,* 1996). Interestingly, studies by Louvion *et al.,* (1996), indicated that the middle domain does not appear to be

needed for any life-sustaining functions but rather may be involved in some regulatory function. Csermely *et al.,* (1998) suggests however that one should be mindful of the high flexibility of Hsp90 and as such the many conformational changes it can take. The may be as yet unknown monomeric forms or even higher oligomeric forms which are functionally equivalent.

1.5 Hsp90's ATPase cycle

In *Saccharomyces cerevisiae*, Hsp90 has a measurable albeit slow ATPase activity, with an affinity for ATP (K_D > 100 µM) that is approximately 10 times poorer than that of other chaperonins. Its rate of catalysis however is comparable to that of Hsp70 without co-chaperone or peptide stimulation (approx. 0.5 min^{-1}) (Young et al. 2001). Hsp90 has been mutated to abolish ATPase activity *in vivo*, and interestingly it cannot substitute for the essential wild-type protein in yeast (Obermann et al. 1998).

In the absence of bound nucleotide, Hsp90 dimers are extended and arranged in an open v-like state (Southworth & Agard 2011), a state in which the chaperone is poised, ready for client loading and binding. It has been suggested that Hsp90 shares a structural mechanism with other homodimeric ATPases such as; DNA gyrase, topoisomerase II, MutL, and the histidine kinases CheA and EncZ (Dutta & Inouye 2000). The mechanism is trigged by ATP binding, which induces inter-subunit contacts between the nucleotide binding domains in the homodimer (Dutta & Inouye 2000; Shiau et al. 2006). This dimerization of nucleotide binding domains is thought to be critical for triggering ATP hydrolysis.

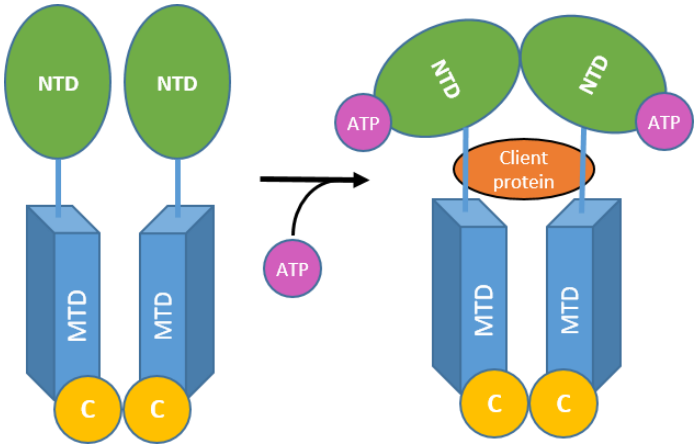


Figure 1-3: Schematic representation of the ATP dependent binding mechanism. (**A**) When ATP is not bound, the N-terminals are extended in an open fashion ready for the capture of a client protein. (**B**) After client protein capture, ATP binds at the N-terminal inducing a conformational change causing the N-terminals to "clamp down", holding the client protein in place (adapted from Brown et al. 2007)

1.6 The intracellular function of Hsp90

As previously mentioned, like most chaperone proteins, Hsp90 helps suppress the aggregation of unstable proteins, and in doing so increases the yield of refolded cellular protein (Wiech *et al.,* 1992; Jakob *et al.,* 1995). Its role in the folding of nascent proteins or the refolding of non-native proteins is only accomplice as a part of a complex chaperone machine which forms the foldosome. This folding process as depicted in (Figure 1-4) is a several step cycle.

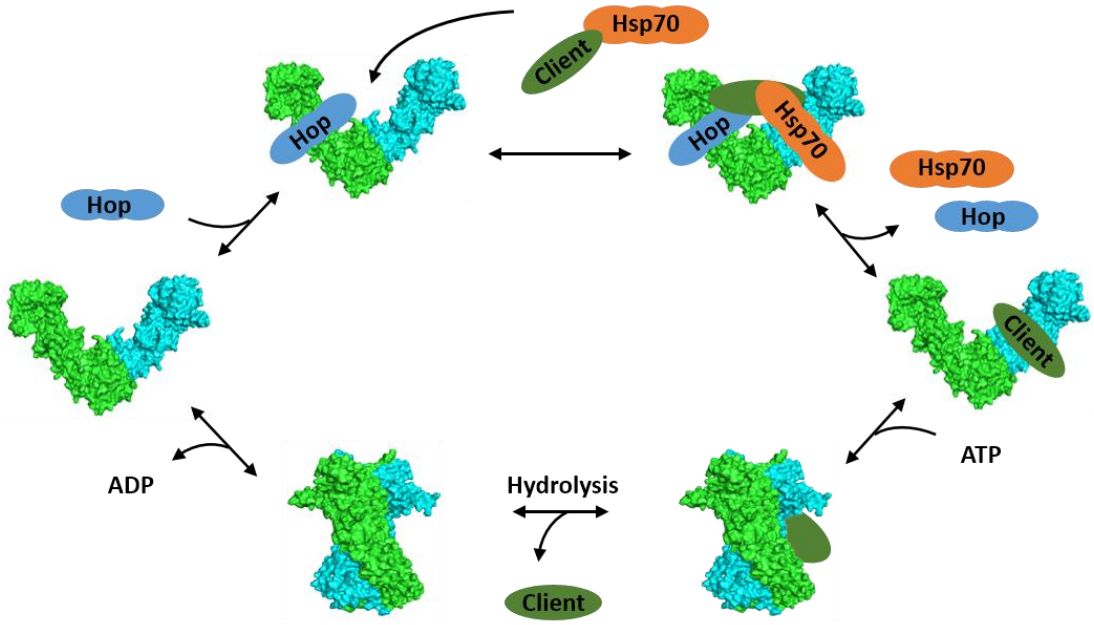


Figure 1-4: Schematic demonstrating Hsp90's functional cycle. The dimerized protein is arranged in an open conformation, where upon Hop and Hsp70-client protein complex bind. ATP binding at the NTD triggers the clamping action of the NTD arms and protein folding takes place. The subsequent hydrolysis of ATP causes the release of the newly folded client, triggering a final conformational shift back to the open v-like state. Adapted from Southworth and Agard (2004)

The first step is initiated by the binding of the non-native target protein by Hsp70 in complex with Hsp40. This complex is then associated via Hsp70, with the "open" conformation of Hsp90 via the organization protein Hop (Pearl *et al.,* 2008). Hop binds strongly with Hsp90, where its interactions with the C-terminal -MEEVD region as well as a binding site in the N-terminal prevent Hsp90 from "closing". By maintaining an "open" conformational state, the client protein can be transferred from the Hsp70/40 complex to Hsp90 seamlessly (Terasawa *et al.,* 2008). When ATP binds Hsp90, the protein p23 displaces Hop, dissociating it and the bound Hsp70/40 from the Hsp90client protein complex, which then moves to a mature from

(Neckers 2003). Hsp90 ATPase is then activated by the binding of Aha1 to the hinge region of Hsp90 and assists in a conformational change that allows ATP to once again bind (Meyer *et al.,* 2004), and its subsequent hydrolysis stimulates the release of the refolded target protein (Terasawa *et al.,* 2008), leading to its activation (Csermely et al. 1998; Jackson 2013)

1.7 The extracellular function of Hsp90

As previously mentioned mammalian Hsp90 has two main isoforms, Hsp90α and Hsp90β. While Hsp90β is constitutively expressed, Hsp90α is induced and expressed in large quantities when the cellular environment is exposed to stress (Sreedhar *et al.,* 2004). Therefore, in the event of an infection, one would expect Hsp90α to be expressed and carry out its cellular maintenance role in the cytosol, this is not necessarily the case however. A number of studies have suggested that there is a pool of Hsp90 located on the cell surface and facing the extracellular space (Tsutsumi & Neckers 2007). Indeed Hsp90 has been identified on the surface of several cell types (Table 1-2).

Table 1-2: The possible functions of different extracellular Hsp90 (Tsutsumi & Neckers 2007)

Cell types	Function
Lung cancer	Immune response
Hepatoma	NA
Fibrosarcoma	Immune response and metastasis
Lymphoma	Immune response
Macrophage, dendrites	Immune response
Monocyte	Immune response
Melanoma	Immune response, invasion/migration and metastasis
Neural cell	Migration
Dermal fibroblast	Migration

As reviewed by Csermely *et al.* (1998), Hsp90α has been found to also play a key role in the stimulation of the immune system when it is located in the extracellular space or plasma membrane (Schmitt *et al.,* 2007). Given that Hsp90 is able to bind a large array of different client proteins, as an extracellular protein, Hsp90α (eHsp90) is thought to play a role in antigen presentation to MHC class I molecules and as such has been identified as a key regulator of the host's immune response (Arnold-Schild *et al.,* 1999).

eHsp90's main function however has been reported to be its involvement in cell motility and wound healing (Stellas *et al.,* 2010; Li *et al.,* 2007; Li *et al.,* 2012; Li *et al.,* 2013). Eustace *et al.* (2004) reported that Hsp90α and not Hsp90β promotes cell motility by binding and activating the metalloproteinase 2 matrix (MMP 2). This activation was shown to be specifically induced by the M-domain (aa 272 – 617) of Hsp90α (Song *et al.,* 2010).

1.8 Hsp90 and its link to cancer

Hsp90 has been found it be quantitatively overexpressed or qualitatively overactive in tumour cells (Kamal *et al.,* 2003). Sahu *et al.,* (2012) reported Hsp90α to account for 2-3% of the total cellular proteins in normal cells and up to 7% in certain tumour cell lines. As reviewed by Li *et al.* (2012) normal cells do not secrete Hsp90 unless they are triggered by environmental insults. Indeed it has been reported that heat-shock at 44°C induced the secretion of Hsp90α in B-cells via nano-vesicles called exosomes (Clayton *et al.,* 2005). Hypoxia, another form of cellular stress, especially in tumour cells, was also found to induce Hsp90α, and that its secretion was regulated by the hypoxia-inducible factor-1α (HIF-1α) (Li *et al.,* 2007). HIF-1α is now considered to be a key regulator in the secretion of Hsp90α.

Hypoxia is a known micro-environmental stress connected to the growth, invasion and metastasis of many solid tumour cancers (Simon & Keith 2008). Cancer cells are forced to adapt to a constant hypoxic environment in order for their survival and continued growth. HIF-1α has been reported to be overexpressed in 40% of tumours in humans (Semenza 2007). Li *et al.,* (2012) report that breast cancer cells over express HIF-1α, and that this over expression leads to the constitutive expression and secretion of Hsp901α. The mechanism for the release of eHsp90α due to environmental cues in normal cells is still not fully understood and requires further investigation.

It is believed that some of the client proteins associated with cytosolic Hsp90 include oncogene products (Trepel et al. 2010). It is sensible therefore to presume that as a cytosolic housekeeping protein, Hsp90 protects the stability and therefore oncogenecity of these products (Lundgren *et al.,* 2007). Examples of known oncoproteins associated with Hsp90 include; HER-2 in breast cancer (Garrett & Arteaga 2011), BCR-ABL in chronic myeloid

leukaemia (Vaidya *et al.,* 2011) and EGFR in non-small cell lung cancer (Hammerman *et al.,* 2009)

1.9 Current Hsp90 anti-cancer therapies

It stands to reason that the theme of anti-cancer drugs, is the targeting of a cellular molecule that is essential to the survival of a cancer cell (Li *et al.,* 2013). Given their regulatory role in cell survival, Hsp90s present themselves as such a molecular target for anti-cancer therapies. Their inhibition in tumour cells would lead to the simultaneous disruption of multiple signalling pathways critical to their growth and survival (Whitesell & Lin 2012). This mass inhibitory approach is thought to address the intrinsic heterogeneity and complexity of the numerous genetic defects characteristic of most known clinical cancers (Trepel *et al.,* 2010).

Over the past two decades research in drug design has targeted the inhibition of Hsp90-oncoprotein complexes in anti-cancer therapies (Jhaveri et al. 2012; Jhaveri et al. 2014; Jego et al. 2013). Clinical results however have been far less promising than previously hoped for, with serious clinical issues being raised over the unselective targeting of Hsp90 in both normal cells and those of tumour cells. Indeed one of the most promising natural compound candidates, a benzoquinone ansamycin Geldanamycin displayed interesting anti-tumour properties *in* vivo, but failed Phase I clinical trials due to severe cytotoxicity (Supko et al. 1995). The analogue tanespimycin however, when given in combination with trastzumab to patients with HER2+ metastatic breast cancer, showed the greatest clinical activity, as evidenced by significant tumour regression (Jhaveri et al. 2014). Other than the poor selectivity of the current Hsp90 inhibitors, another complication in their development is the intrinsic complexity of molecular oncogenesis as well as the frequent emergence of resistance to current treatments (Whitesell & Lin 2012). Together these hindrances have deemed current Hsp90 inhibitors to be unsatisfactory anti-cancer therapies. Jhaveri et al. 2014, however suggest that investing into the understanding the fundamentals of drug delivery and patient selection for the current unique inhibitors, will allow for the next generation of Hsp90 targeted therapies to reach their full potential as anti-cancer therapeutics.

1.10 *In silico* drug discovery and computational studies

As previously discussed, the molecular mechanisms of Hsp90 are generally regulated through ligand-based modulation of protein dynamics, as well as inter-domain communication, and these events lead to the activation of functionally specific conformational states. Despite not having a fully solved structure of Hsp90, recent years have seen a remarkable progress in the understanding of the structural biology of Hsp90. The upshot being the development of numerous computational analyses, which have led to a much improved understanding of the dynamics and mechanisms of the chaperone at an atomic resolution (Verkhivker et al. 2009). This improved understanding has facilitated the development of *in silico* Hsp90 drug discovery studies.

Computer-based virtual screening approaches have uncovered novel potent chemotypes of Hsp90 inhibitors, while highlighting the importance of receptor flexibility and the many conformational changes in Hsp90's chaperone cycle (Huth et al. 2007; Park et al. 2007; Barril et al. 2005). Indeed Mahanta and colleges developed a novel Geldanamycin analogue Hsp90 alpha-inhibitor, using *in silico* techniques (Mahanta et al. 2013). This analogue inhibitor is thought to have increased binding affinity, efficacy and less toxic for the therapy of breast cancer. Overall, computer-based ligand screening has led to a number of novel potent chemotypes of Hsp90 inhibitors, suggesting that the integration of structural and computational approaches may be a viable route in the development of therapeutic strategies.

Despite the continuous optimism for the success of Hsp90 inhibitors that target the NTD ATP binding site (Plescia et al. 2005; Marcu et al. 2000), the emerging strategies in the development of potent and selective anti-cancer therapies are beginning to focus on targeting allosteric binding sites and disrupting the interactions between Hsp90, its co-chaperones and client proteins. Interrupting protein client binding with Hsp90 (Hieronymus et al. 2006; Zhang et al. 2008) may provide a different treatment approach, in which the activity of Hsp90 is not completely lost and rather specific client proteins are prevented from binding to the chaperone complex (Verkhivker et al. 2009).

The CTD of Hsp90 is considered to be a very important domain being the site of chaperone dimerization, studies however have demonstrated an additional binding site for inhibitors within this region of the protein (Garnier et al. 2002; Söti et al. 2002). This binding site favours GTP over ATP, which potentially enables the selective design of novel inhibitors against this site. One such selective inhibitor called cisplatin has been discovered having a different Hsp90-dependent kinase inhibitory pattern to Geldanamycin (Söti et al. 2003). Another putative inhibitor CTD binding site was identified using a combination of several computational approaches (Sgobba et al. 2010), providing a starting point for further experimental studies aimed at validating the CTD as a viable drug target.

1.11 Project motivation

Although much work to date has been completed on Hsp90 both *in vitro* and *in silico*, complete understanding of Hsp90's multitude conformational mechanisms remains conjecture. Hsp90's role in the progression and maturation of several cancers has been well established. Despite numerous targeted drug discovery studies, a relatively non-toxic, efficient anti-Hsp90 therapy remains elusive. Given the recent advances in the field of structural and computational biology, the rapid growth of crystallographic structures of Hsp90-inhibitor complexes provides an ideal experimental platform from which to launch *in silico* structure-based drug discovery development.

1.11.1 Problem statement and knowledge gap

Despite the numerous crystallographic structures available to research scientists, the lack of a complete structure for human Hsp90 remains a limiting knowledge gap. Given that Hsp90 is a highly flexible and dynamic protein, it is not unreasonable to surmise that any structure-based study of Hsp90 will in some way be affected by interactions, whether they by ligand based, of inter-subunit, in other parts of the proteins. As such, without complete structural model of human Hsp90 any conclusions made with respect to structural studies will be severely limited by this apparent lack of structural information.

1.11.2 Aims and objectives

This project aims to directly address the structural limitations applicable to research directed toward human Hsp90, by using a complex homology modelling approach to calculate a full sequence model of human Hsp90 from which to build a platform and pave the way for rational drug design and identification of next-generation Hsp90 inhibitors. The objectives of this project are thus two fold; firstly and most importantly to accurately calculate human Hsp90 homology models in two of its most drastic conformations, namely its open v-like state and closed highly compact state. The second objective is entirely reliant on the accuracy of the homology modelling and will involve structure-based, *in silico* drug discovery techniques in an attempt to identify novel binding pockets and potential target sites. This objective will include the construction of a library of small molecule compounds of South African origin. This library will be used in a series of *in silico* docking experiments, to elucidate novel binding pockets on both open and closed state Hsp90. Well bound compounds will be further analysed and investigated for potential use as lead compound candidates.

CHAPTER 2: HOMOLOGY MODELLING

This chapter focuses on the implementation of homology modelling techniques in conjunction with protein-protein docking and energy minimization, to calculate accurate homology models of full length human Hsp90. The chapter is preceded by a brief overview and explanation of the aforementioned techniques, and is followed a detailed description of the methodology followed to calculate full length homology models of human Hsp90 in open 'v-like' and closed nucleotide bound conformations.

2.1 Introduction

Homology modelling is a computational technique used for the determination and prediction of protein tertiary structure. The rapid evaluation and development of fast and relatively cheap sequencing techniques has led to a global explosion of genetic data, with the number of new entries in the National Center for Biotechnology Information (NCBI) database doubling on average every 18 months since 1985. The annual NCBI report for 2014 indicated that the number of stored sequences has grown from 10 million to 100 million since January 2000, and that of the 100 million sequences, there are 46 968 574 non-redundant, fully annotated protein sequences (Anon 2014). Despite this veritable mountain of available protein sequence data, there is comparably little or no known relative 3D-structural data to match.

In spite of limited understanding as to the mechanism by which native proteins fold into their functional forms, improved computing power and efficiency at a hardware level has opened the door for advancements in *ab initio* prediction of protein tertiary structure directly from amino acid sequence (Venclovas et al. 2003). *Ab initio* methods must take into consideration the physicochemical properties of all amino acid combinations involved in the folding process and the strain on computation thus increases with an increase in sequence length. Despite the establishment of several different algorithmic approaches, *ab initio* methods still remain a very time consuming process and the computational simulation of protein folding has been limited to very small peptides (Xu & Zhang 2012; Jayaram et al. 2014).

In the past, the techniques and methods used for garnering protein structural information have relied solely on the experimental techniques of X-ray crystallography, Nuclear Magnetic Resonance (NMR) and high resolution electron microscopy. These techniques have been extensively developed and currently contribute to the majority of all known 3D protein structures. Despite this, there is still a large gap present between the number of available 3D structures and protein sequences. When no experimental structure exists for a given protein, homology modelling has proved to be a viable alternative, given a close homolog with known structure.

It is well documented that protein function is well conserved over evolutionary stages (Illergård et al. 2009). A phenomenon explained by natural selection and the consistency observed in protein folding and chemistry. It thus stands to reason that protein structure, a physical characteristic that defines protein function, is as highly conserved. Indeed Illergård and colleagues showed that a protein's structure is conserved up to ten times more readily than its sequence. Given a protein sequence with no known 3D structure, it is not unreasonable to consider the feasibility of computationally extrapolating structural information from close homologues with known 3D structures to build an accurate 3D protein model. Homology modelling is one such computational technique and is often referred to as comparative modelling, owing to the fact that it draws on structural comparisons between the protein of interest and any close homologs with known 3D structure.

2.2 Homology modelling process

Homology modelling is a multi-step process that requires both the computation of complex mathematical algorithms as well as the accurate identification of functional patterns in a protein sequence, such as protein folds and turns (Eiben et al. 2012). The process is logically ordered and provided that suitable homologous templates are identified and accurately aligned, a large part of it can be automated and can be broken down into four main steps (Figure 2-1), namely template identification, target-template alignment, model building, and validation.

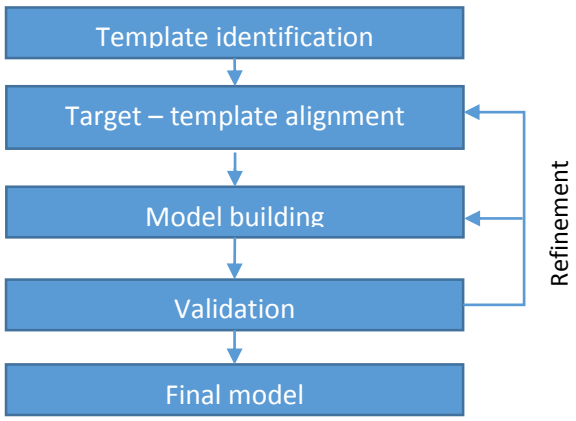


Figure 2-1: Schematic depicting the logical work flow of the homology modelling process. Should validation prove the generation of an unsatisfactory model, refinement either to the current model or the sequence alignment maybe required.

2.2.1 Template identification

This is the first step in the homology modelling pipeline and in conjunction with the sequence alignment is absolutely crucial to the accuracy of the final homology model. Because homology modelling is built on the premise that tertiary structure is conserved between homologous proteins, it is of utmost importance that the homology between the protein of interest and template be accurately determined. Structural databases are thus searched for suitable templates based on the criteria of homology and quality. Should a close homolog of good quality be identified, its structural data is extrapolated to the target sequence, a process that is strictly guided by sequence alignment alone, highlighting the importance of this step.

The Protein Data Bank (PDB) is a freely accessible database of experimental structural macromolecule data, and currently holds approximately 105732 structures (Anon 2015). Structural data is presented to the user in a specific file format (PDB), which contains structural information for each atom in the protein's structure. Searching this already vast database for the structural information of a particular protein is fairly simple and can be refined using the advanced search functionality. However obtaining the structural information of close homologs to a particular protein is a much more challenging task, and better achieved using 3rd party tools such as the HHPred web-server (Söding et al. 2005).

HHPred is a remote interactive server that was developed to provide biologists with a method for database sequence searching that allows for fast protein homology detection and

structure prediction (Söding et al. 2005). The algorithm is the first to implement pairwise comparison of hidden Markov models (HMMs), giving it a significant efficiency speed advantage over similar software such as BLAST and PSI-BLAST. HHPred can be used to search a several different databases other that the PDB including, SCOP (Hubbard et al. 1999), Pfam (Finn et al. 2014), SMART (Ponting et al. 1999), COGs (Koonin 2002) and CDD (Marchler-Bauer et al. 2002). While the aforementioned software is fully capable of detecting homology, a recent study (Söding 2005) found that HHPred had the highest sensitivity and alignment accuracy for remote homology detection.

2.2.2 Target – Template alignment

In this alignment, the target sequence, which is the full amino acid sequence of the protein to be modelled, is aligned to the known "structural" sequence of a suitable template, and is widely regarded as being the most important step in comparative homology modelling, where poorly aligned sequence segments result in low quality homology models. When the spatial data for residues cannot be resolved experimentally, they are listed in their respective PDB file as "missing residues". When aligning the target sequence to a template sequence, one must account for these missing residues, by aligning the target sequence to only those residues present in the template PDB file, thus ensuring that only those residues with known structural data are. As such this step in the homology modelling process is best accomplished using software that is capable of extracting the amino acid sequence of solved residues directly from the PDB file. Although there are several structural alignment programs available, PROMALS3D (Pei et al. 2008) is an example of such software.

PROMALS3D is an alignment tool that derives structural constraints through structure-based alignments and combines them with sequence constraints to build consistency-based multiple sequence alignments (Pei et al. 2008). The algorithm used by PROMALS3D improves the alignment quality of distantly related sequenced by combining several advanced techniques such as database searching for additional homologs, prediction of secondary structure as well as probalistic consistency of profile-profile comparisons. Being capable of secondary structure prediction, and given that structure-based alignments are regarded as being the gold standard, PROMALSD essentially bridges structure-based alignments with

sequence alignments to generate high quality multiple alignments that are consistent with both sequence and structural information (Pei et al. 2008).

2.2.3 Model building

There are many different programs and web servers that are capable of automating not only the homology modelling process, but also the evaluation of built models. While these programs and webservers make protein modelling possible for both professionals as well as non-specialists, manual intervention is often a necessity in order to maximise accuracy. Discussed here are the techniques used for comparative homology modelling using the program MODELLER (Sali & Blundell 1993).

MODELLER was originally written and developed by Sali and Blundell as a program that predicted the tertiary structure of proteins. The software was written to be implemented within the Python programming environment and by extrapolating intra-protein interaction information for existing natively folded protein structures, uses comparative techniques. The modelling algorithm calculates 3D protein models using the satisfaction of spatial restraints, a technique well known in NMR. Spatial restraints are expressed as a probability density function for each protein sub-structure. Models are thus built to satisfy these spatial restraints based on the alignment of target to template sequences (Sali & Blundell 1993). The required input for MODELLER, includes the alignment file containing the target-template alignment, as well as the original PDB file of the template. To ensure that model building returns an optimum protein model, MODELLER utilizes an iterative process whereby several models are calculated, such that models that are not statistically significant can be discarded and the resulting models sampled (Eswar et al. 2001). Given that loop regions are highly variable and thus very difficult to predict and model, MODELLER includes some elementary *ab initio* structure prediction, which greatly improves the modelling of these regions.

2.2.4 Model evaluation

Regardless of the number of models generated during model building, individual assessment of each model is required such that they can be ranked according to a statistical measure and the optimal model identified. The model building program MODELLER caters for this with a Discrete Optimised Protein Energy (DOPE) score. This score is an atomic distance-dependent

statistical potential which can be used to evaluate the fold of a model (Shen & Sali 2006). The score is based on the physical reference state of non-interacting atoms within a uniform sphere, which corresponds with the finite size and spherical shape of proteins and is dependent on a sample native structure. The normalised DOPE score (Eramian et al. 2008) is a standard Z-score, and is derived directly from raw DOPE scores, where positive scores are regarded as being poor models, and negative scores less than -0.5, deemed accurate and closest to the native structure (Eswar et al. 2001). Being a Z-score this qualitative value is often referred to as the DOPE-Z score.

Besides the DOPE-Z scoring from MODELLER, there is a wide array of online webservers dedicated to protein quality assessment. One such server is the MetaMQAPII Meta-server. This software is designed to accurately assess both the local and overall tertiary structural quality of a protein model (Pawlowski et al. 2008). The strength of assessment on the Meta-server is vastly increased with the incorporation of results from eight other model quality assessment servers; ProSA, VERIFY3D (Wiederstein & Sippl 2007),TUNE (Lin et al. 2002), BALA-SNAPP, ANOLEA (Krishnamoorthy & Tropsha 2003), PROQRES (Wallner & Elofsson 2006) and REFINER (Boniecki et al. 2004). The MetaMQAPII algorithm assesses each residue in a given model by grouping it in one of 315 electrostatic environment groups. A linear regression model is developed for each of these groups, allowing for the extrapolation and determination of the specific RMSD of a particular residue within a given group from its position in a native structure (Pawlowski et al. 2008). MetaMQAPII outputs the results of an assessment in a PDB coordinate file, where the B-factor values for each residue in the model, is modified to represent an overall assessment score for that particular residue. When visualising this PDB file in most molecular visualisation viewers and colouring the secondary structures by the B-factor spectrum, the model is accordingly coloured by using an RMSD score based on the deviation from the native structure, where blue represents accurate low regions where the RMSD value is very small and red inaccurate regions where RMSD value is high (Pawlowski et al. 2008). This novel technique allows the user to quickly identify problem areas within a protein model. MetaMQAPII also returns a log file which contains a predicted Global Distance Test Total Score (GDT_TS), which is an overall RMSD value that describes the predicted deviation of the model from the true protein structure in angstroms.

2.3 Protein – protein docking

In silico protein-protein docking is the prediction of protein interactions between two proteins that are known to interact (Comeau et al. 2003). The major challenge of this technique is to take the coordinates of two unbound molecules, and obtain a statistically accurate model for a bound complex. To date docking methods have improved substantially with the development of several different techniques and approaches, such as the powerful fast Fourier transform (FFT) correlation methods (Gabb et al. 1997), the Monte Carlo methods as used by RosettaDock (Gray et al. 2003) or the high ambiguity driven biomolecular docking (HADDOCK) (Dominguez et al. 2003). One of the most recent debutants, ClusPro (Comeau et al. 2003), is an automated rigid-body docking webserver that uses a discrimination algorithm capable of filtering docked conformations, and ranking them based on their clustering properties. This filtering algorithm uses an empirical free energy evaluation method that selects conformations with the low desolvation and electrostatic energies (Comeau et al. 2003). The server performs three essential computational steps: (1) rigid-body docking using the FFT correlation functionality of PIPER; (2) root mean square deviation (RMSD)-based clustering of all generated structures to find the largest clusters that represent, statistically, the most likely models of the bound complex; and (3) the refinement of these clustered structures (Kozakov et al. 2013).

2.4 Energy Minimization - GROMACS

As the name implies, energy minimization is a procedure that attempts to minimize the potential energy of a system to its lowest possible point. It is used in molecular modelling to prevent steric clashes or any inappropriate geometry in the modelled protein. The model building process does not take intra-protein forces within the homology model into account and thus a homology model cannot truly be said to be in a native conformation. Energy minimization is a technique in which to circumvent this and the GROMACS package is highly regarded as being one of the most suitable platforms.

The GROningen MAchine for Chemical Simulations (GROMACS) is a molecular dynamics package that was designed primarily for protein, lipid and nucleic acid simulations (Berendsen et al. 1995). GROMACS accomplishes minimization by constructing a topology file which

contains all the information necessary for defining the molecule of interest in simulation. This information includes non-bonded parameters (charges and atom types) as well as bonded parameters (bonds, bond angle and dihedrals) and is based on a chosen force filed. Once the topology for the molecule has been defined, a simulation box or unit cell must be generated. This box essentially defines the simulation space around the molecule of interest. The size of this box must be defined and a suitable solvent such as water added to the simulation space. At this point the system is solvated and contains the molecule of interest, which holds a net charge. GROMACS solves this charge discrepancy by adding counter ions to the solution thus neutralizing the molecule. Allowing for brief simulations, the molecule's structure is then allowed to relax into a more native state.

2.5 Chapter objectives

The over-riding goal of this chapter was to build accurate homology models of human Hsp90. This objective is made fairly complex, given Hsp90 naturally exists as a homodimer and that when dimerized it undergoes several conformational changes during its functional cycle. As previously discussed, prior to protein client binding, Hsp90 is dimerized at the C-terminal alone and is arranged in an "open" conformation at the N-terminal. After client binding however, dimerization is observed at both terminals in a tightly bound "closed" conformation. Given these two functional conformations, the primary objective was divided into two parts: 1) modelling of human Hsp90 in closed conformation and 2) modelling human Hsp90 in its open conformation.

Being a hypothetical and statistically based approach, the accuracy of homology modelling is very much questionable. This is not surprising though, given that it is often necessary to base the modelling on more than one structural template, where each and every template introduces new structural restraints and inaccuracies. The importance of validation and evaluation of homology models thus cannot be underestimated. The final objectives of this chapter were thus, 1) to improve the quality of built models as much as possible using energy minimization techniques, and 2) to fully vet and critically evaluate both sets of models, such that they are fully prepared for future experimentation.

2.6 Methodology

2.6.1 Template identification and retrieval

The protein sequence for human Hsp90 was retrieved from the Genbank database (NCBI, AAI21063.1) and submitted to the HHPred webserver and protein-BLAST, for the identification of homologous protein sequences with and without known structure respectively. Results returned by HHPred were analysed and sorted according to the percentage amino acid identity shared between the target sequence and homolog sequence, as well as the expected value (E-value), and only sequences that held experimental structural data and not predicted secondary structure were retained. Lastly these respective PDB files were closely examined and critically assessed according to their resolution, R-free and R-values.

2.6.2 Structural sequence alignment

All structural template sequences retrieved (PDB IDs: 2CG9, 2CGE, 3T10, and 2O1U) were aligned to the target sequence along with several non-structural homologs (NCBI accession numbers: gi_67542540, gi_47522774, gi_34395877, gi_60592792, gi_25565303, and gi_55730837) using the structural alignment program PROMALS3D. Where template PDB files contained more than one chain, only the chains relevant to the target sequence were used. All other alignment parameters were left to the default values. The resulting alignment was visualized and edited using Jalview, removing any large target inserts and unnecessary gapped regions. Once edited, the homologous sequences that held no structural data were removed from the alignment and the remaining template and target sequences were converted into PIR file format, in preparation for model building. The alignment file for the closed homodimer included the following PDB templates 2CG9, 2CGE, 2O1U and 3T10, while the open homodimer 2CG9, 2CGE and 3T10. The rationale for using these specific templates is discussed in the Results and Discussion Section, and template details are shown in Table 2-1.

2.6.3 Model building

Homology models were calculated using the program MODELLER 9v13. All modelling was carried out in parallel over three nodes and 192 cores on a local server, to generate 100

unique homology models. Other parameters included were a very slow refinement and the generation of fitted models using the M3ALIGN method.

2.6.4 Model validation

Initial model evaluation was completed by calculating the normalized DOPE-Z score for each of the 100 models and ranking them accordingly. The top three scoring models were retained and visually analyse using PyMOL visualizer (Schrodinger LLC 2010). Models that contained no steric clashes or obvious geometric discrepancies were submitted to the MetaMQAPII webserver for further qualitative statistical evaluation, and the best scoring model retained for further analysis.

2.6.5 Protein-protein docking

Modelled monomers of human Hsp90 were docked to one another using the ClusPro 2.0 webserver via specified C-terminal residue interactions, to calculate an open dimerized conformation homology model. As required by the ClusPro webserver, the chain A monomer was set as the receptor protein and the chain B monomer the binding ligand. The acceptor residues were set as follows a-610, a-622, a-626, a-627, a-630; b-610, b-622, b-626, b-627, b-630 based on work done by Rastelli (2014). All other available settings were left to the default option.

2.6.6 Energy minimization

Energy minimization was accomplished using GROMACS version 4.5, and the CHARM27 (MacKerell et al. 2001) force field. The protein was centred in a cubic box with a maximum edge space of 1.0 nm. The net negative charge in all cases was neutralized by adding an equivalent number of sodium ions to the solution. The group 13 SOL water solution was chosen as the solvent to ensure that the added counter ions were embedded within the solution and not added to the protein. The energy potential after simulation was graphically analysed using the Xmgrace plotting tool.

2.7 Results and discussion

2.7.1 Template identification and retrieval

Human Hsp90 protein sequence AAI21063.1 is 732 amino acid residues in length and represents the full length Hsp90 sequence, and being sequenced as the cytosolic alpha isoform, the most active form of the protein (Freeman & Morimoto 1996), AAI21063.1 was selected as the sequence of interest in this study, and from this point will be referred to as the target.

HHPred template identification showed that the closest full sequence structural homolog of human Hsp90 belonged to *Saccharomyces cerevisiae* or baker's yeast, PDB ID 2CG9. Template 2CG9 had the highest score out of all the templates returned by HHPred. A sequence identity of 63% (Table 2-1), suggesting it to be a good candidate for use as a full sequence template. HHPred however is highly sensitive and capable of detecting homology well below the twilight zone (< 20% sequence identity) (Söding et al. 2005), giving rise to the strong possibility that template hits with high sequence identity alone may represent false positives. It is therefore necessary to verify the validity of putative hits using means other than sequence identity alone. Thus conservation of secondary structure prediction (SS) as well as the probability and expected value (E-value) were taken into consideration. The probability in this instance represents an estimate of how likely it is for a template hit to be homologous to the query sequence while taking secondary structure into account; where anything over 95% is regarded as being fairly certain and anything greater than 50%, worth further investigation (Söding et al. 2005). The E-value on the other hand, is not affected by secondary structure, and represents a probability measure of the chance of finding alternative hits with a better score, if the database were to only contain unrelated hits to the query sequence (Söding et al. 2005). Template 2CG9 had a 100% probability and a significantly low E-value of $1x10^{-161}$ (Table 2-1), thus providing further support that in this case, the structure of 2CG9 to be a suitable full sequence structural template for the homology modelling of human Hsp90.

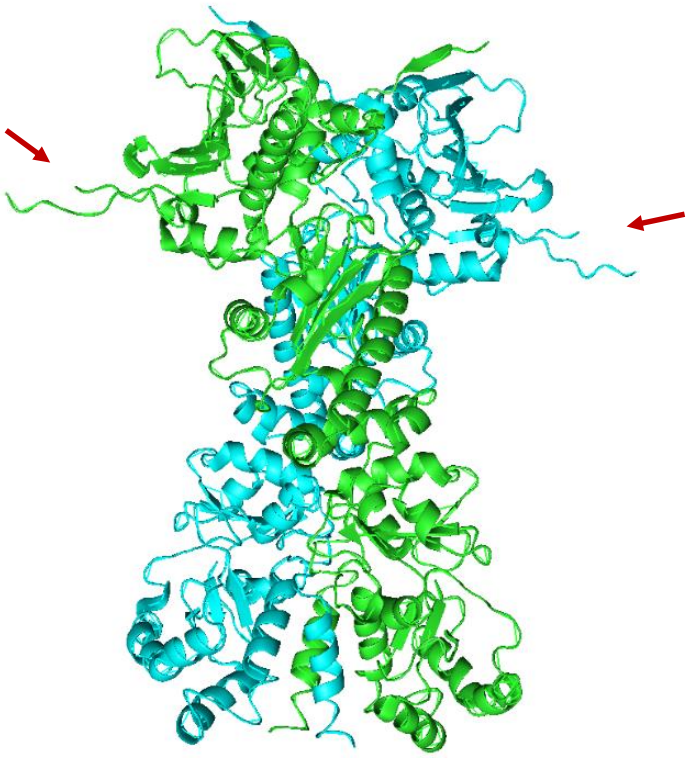


Figure 2-2: The structure of yeast Hsp90 (2CG9) in closed conformation visualised in PyMOL. Examples of missing residues in both chain-A (green) and chain-B (cyan) are indicated (red arrows).

Visual analysis of 2CG9 (Figure 2-2) showed a full homodimer with identical monomers arranged in the closed conformation (Shiau et al. 2006). Closer inspection however revealed missing residues in both chains. Reviewing the PDB coordinate indicated that 2CG9 was solved by X-ray crystallography and confirmed that residues 1, 217-261, 330-338 and 598-610 in chain A were missing as well as residues 1, 217-261 and 598-610 in chain B. This loss of structural data is most likely attributed to resolution limitations, indeed the overall quality of 2CG9 was found to be substantially low with a lofty resolution of 3.10 Å and high scoring R-value and R-free values of 0.314 and 0.353 respectively (Table 2-1).

Although 2CG9 scored remarkably well with HHPred, the quality of its structural data is clearly suboptimal for use as a single homology modelling template, and due to the incomplete nature of this structure, modelling of the regions with missing residues would be impossible. In an attempt to circumvent these limitations and ensure the best accuracy possible, a multi template homology modelling approach was taken.

Table 2-1: Detailed HHPred summary of the structural protein templates used for homology modelling process

TEMPLATE	ORGANISM	SCORE	PROBABILITY	E-VALUE	SS	% ID	RANGE	CHAINS	SOLVED	RESOLUTION	R-VALUE	R-FREE	REFERENCE
2CG9	*S. Cerevisiae*	1396.7	100.0	$1x10^{-161}$	41.1	63%	17-732	A & B	XRD	3.10 Å	0.314	0.353	(Ali et al. 2006)
2CGE	*S. Cerevisiae*	982.6	100.0	$3x10^{-118}$	37.1	63%	293-732	A & B	XRD	3.00 Å	0.258	0.288	(Ali et al. 2006)
2O1U	*C. Lupus familiaris*	1316.9	100.0	$4x10^{-152}$	46.8	50%	12-732	A & B	XRD	2.40 Å	0.245	0.289	(Dollins et al. 2007)
3T10	*Homo sapiens*	452.7	100.0	$1x10^{-55}$	20.7	100%	16-228	A	XRD	1.47 Å	0.209	0.234	(J. Li et al. 2012)

As previously mentioned, HHPred analysis returned several structural homologs of Hsp90, each isolated from a different eukaryotic organism. The rationale for selecting 2CGE, 201U and 3T10 (Table 2-1) as additional templates to 2CG9, is best explained by analysing the structural alignment between the target sequence and chains A and B of 2CG9 (Figure 2-3). This MSA draws immediate attention to several inserted regions in the target sequence (red boxes). Indeed this alignment confirms that there are missing residues in the PDB file of 2CG9, as it is not surprising to note that the inserted regions in the alignment correspond accordingly with the afore mentioned missing residues. The loss of structural data at the very beginning and very end of a crystalized protein is not uncommon, due to difficulty in capturing these regions during crystallography. The rather large 51 residue insert in the middle domain and a 13 residue insert in the CTD however, clearly highlight the shortcomings of 2CG9 as a sole template, and it would have been prudent to continue the homology modelling process using this template alone. Thus, the templates listed in Table 2-1 were selected, as additional templates as each contained structural information for one or more of these inserted regions (see Figure 2-4).

Looking at the qualitative data in Table 2-1, the templates chosen were all solved at higher resolutions than that of 2CG9, and had better R-free and R-value scores. Figure 2-4 shows how each template covered the gapped regions shown in Figure 2-3. Interestingly, no template structures could be identified for the large 51 residue insert in the middle domain, and the C-terminal. Modelling these region without structural information would result in the generation of an undefined loop segment with no recognizable folding pattern. To prevent this from occurring, both regions were deleted from the alignment and thus not included in the model building process. While this is clearly not an optimal solution, as deleting these regions drastically reduces the relative true accuracy of the final model in comparison to the native protein, by using multiple templates which are of a higher quality than the primary template, the remainder of the protein can be modelled fairly accurately.

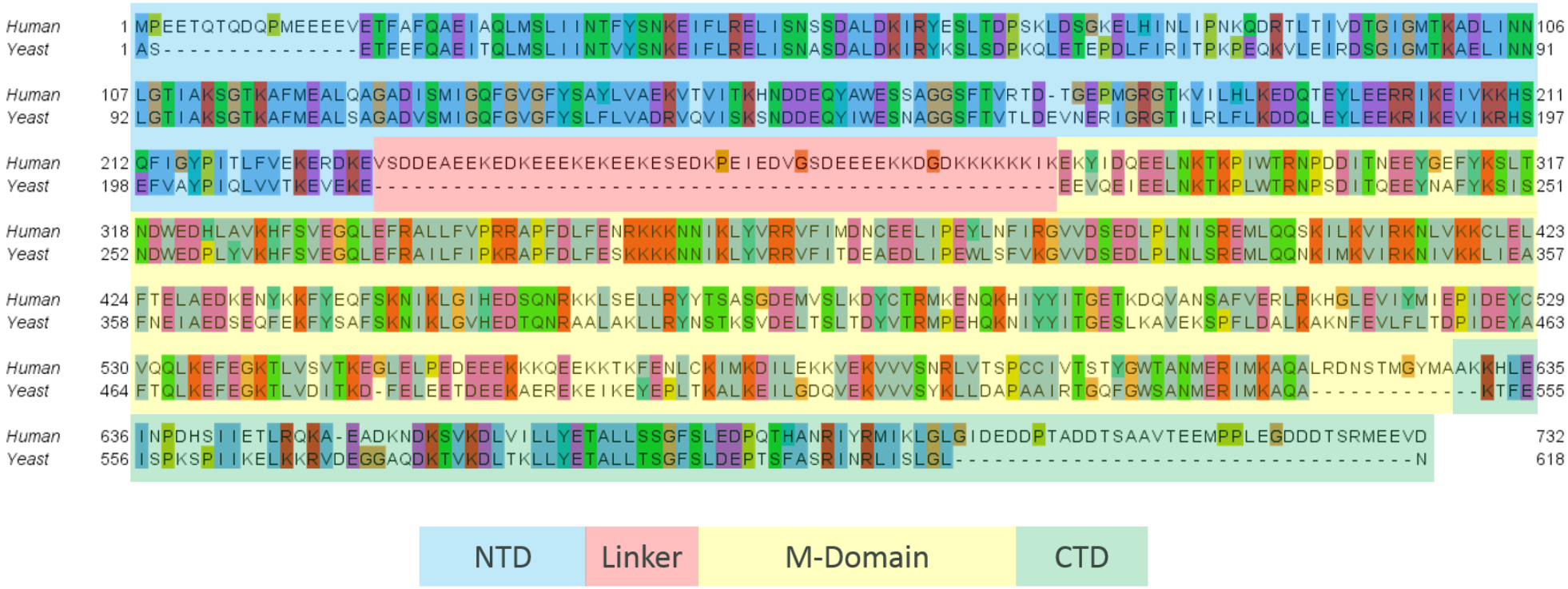


Figure 2-3: Structural sequence alignment between the Hsp90 sequences of human and yeast respectively, showing the suitability of the yeast sequence for use as a structural template, across all three structurally distinct domains (NTD-blue, M-domain-yellow, and CTD-green). While the conservation of these regions is generally high, inserted regions were observed at the extreme ends of the NTD (blue) and CTD (green), as well as a large 51 residue insert in the linker region (red)

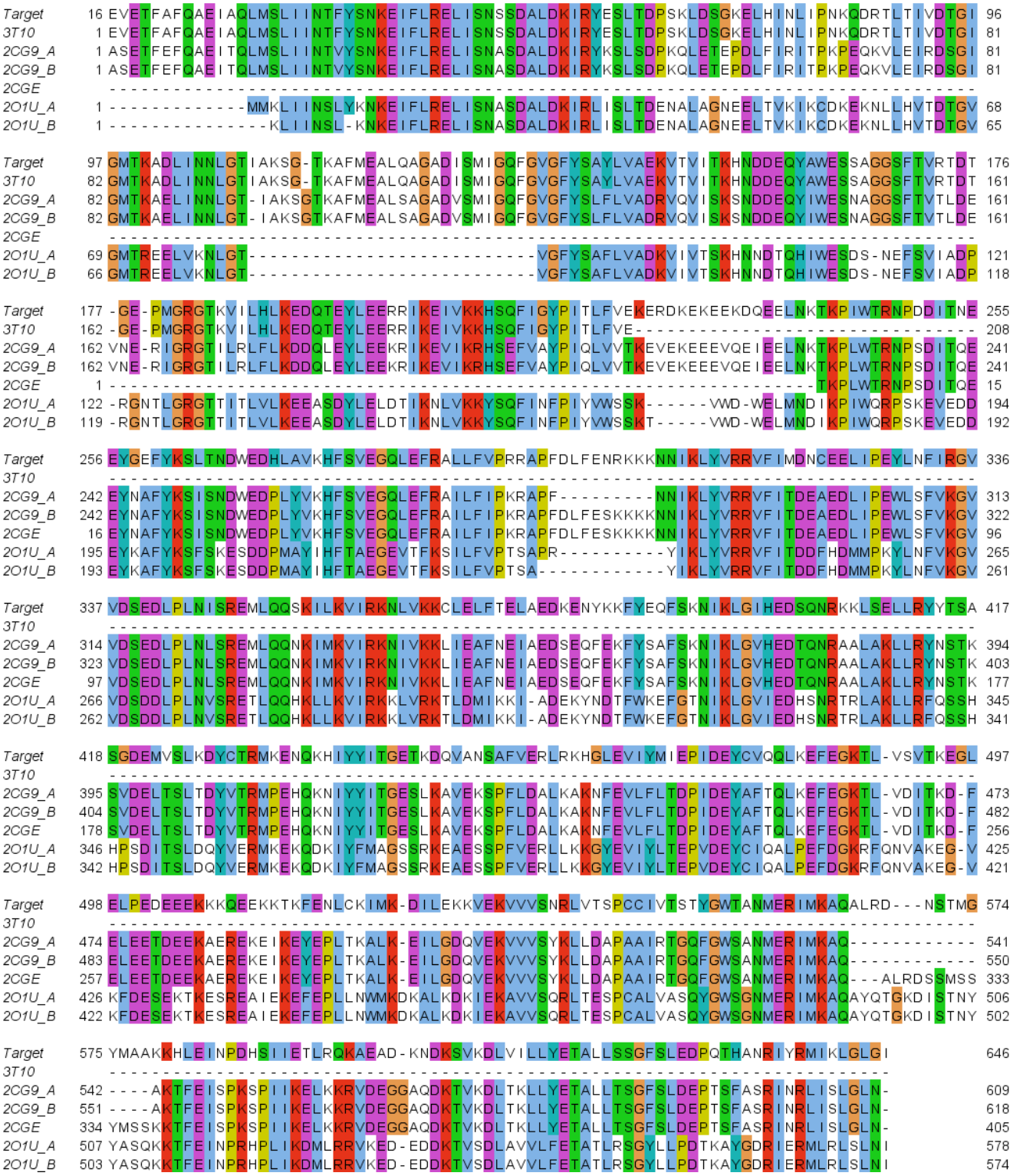


Figure 2-4: Multiple sequence alignment of all templates and target sequence, indicating how each was used to supply structural data for specific inserted regions. Alignment made using PROMALS3D and edited in Jalview

2.7.2 Alignment file construction and model building

In this Section the modelling procedures used for constructing models of human Hsp90 in dimerized closed conformation and dimerized open conformation are discussed in detail. It is necessary however, for the reader to have a brief outline and overview as to how these separate tasks were tackled.

Seeing as the template structure 2CG9 is representative of full length Hsp90 in closed conformation, the procedure for modelling the human homolog in this conformation was made fairly straight forward. The only complication being the addition of several other templates to cover gapped regions. Modelling the open conformation counterpart however, proved to be far more challenging, there being no known suitable template in this particular conformation. A simple homology modelling approach in this instance was not going to be a viable option. Seeing as Hsp90 is a homodimer however, it is possible to model a single monomer of the protein in open conformation. Using this monomer, protein-protein docking techniques could be employed to dock it to a copy and obtain an open conformation model. Although both procedures require the same homology modelling techniques, the significant differences between a monomer in closed conformation and a monomer in open conformation, deem it necessary to construct separate alignment files and thus carry out separate modelling experiments.

2.7.2.1 Construction of closed conformation alignment file

Despite its low quality, the template structure 2CG9 provided an excellent template for modelling human Hsp90 in its closed conformation, and as such it was used as the primary template. To improve the quality of the final model however, the template structure 2CGE was used as an additional model, contributing to the structural data available for the latter half of the middle domain and most of the CTD (Figure 2-5). The insert present at N-terminal was duly covered using the template structure 3T10, and being resolved at a mere 1.47 Å resolution, the quality of the entire domain was also improved. Although the 13 residue insert present in the CTD was well covered by 2CGE, this region was later found to be poorly modelled with steric clashing occurring between the two chains (see Section 2.5.2.2). To solve this problem, the template 2O1U was added to the list of templates, providing improved spatial restrains for this particular segment.

The alignment in Figure 2-5 shows how the modelling alignment file (PIR file) was constructed for chain A only. Being a homodimer the actual PIR file (see Appendix A) was constructed such that the structural data presented here was duplicated using chain breaks in each template. By doing this, it was possible to model both chains simultaneously.

Figure 2-5: Graphical representation of the PIR file alignment for a single chain of the homodimer in closed conformation, showing how each template was used to map structural data to the target sequence (red).

2.7.2.2 Model building and evaluation of Hsp90 in closed conformation

Initially the modelling of human Hsp90 in closed conformation was completed using three out of the four templates earlier reported, namely 2CG9, 2CGE and 3T10, with the exclusion of 2O1U. Up to this point in the study, it was thought that these templates suitably covered the full target sequence, bar that of the large middle domain insert. Upon model building however, as is always the case in the real world, the quality of the modelling for the residue segment 598-610 was very poor indeed. This particular segment appeared to represent a loop region and as such, is expected to have a highly variable placement. The image in Figure 2-6 however shows particularly poor placement, with the loop segment in chain B (cyan) extending and threading through a loop region in chain A (green), essentially knotting the two monomers together. Knotting in any protein is completely unnatural and as such this model was deemed unacceptable.

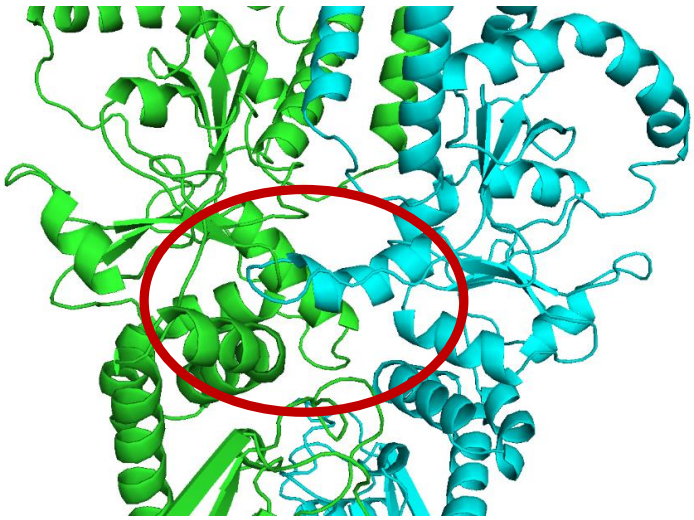


Figure 2-6: Schematic showing knotting (red circle) observed between chains A (green) and B (cyan) of human Hsp90 homology models during the early stages of the model building process

In order to improve the quality of this loop segment such that knotting would not occur, 2O1U was included in the alignment file. One will notice however that 2O1U, based on sequence identities, does not provide a much better quality template than 2CGE. However after remodelling the protein with this template included, the loop segments were placed further apart such that loop segments from either chain could not intertwine, preventing the monomers from knotting. While this solution may seem rather "lucky", the results obtained can be easily explained. By introducing a second template of similar quality to that of the initial template, one

is essentially providing MODELLER with two sets of spatial constraints for this particular region. Being unbiased, MODELLER adds these restraints together and uses and average of the two, which in this case provided a suitable improvement such that knotting between the two monomers was prevented.

Model generation through MODELLER is a statistical approach to homology modelling, and as such, 100 homology models were generated in total, in an attempt to improve the likelihood of producing the best model possible. Each of these models were evaluated by calculating its respective DOPE-Z score, and the three lowest scoring models selected for further analysis and evaluation. The graph in Figure 2-7 shows the variability of the DOPE-Z scores between the different models, where the most negative models (7, 71 and 92) represent the three best scoring models. Each of these three models were submitted to the MetaMQAPII webserver for a more thorough statistical analysis. Visual analysis of the PDB files returned by MetaMQAPII (Figure 2-12 A) colouring the structures according to a colour spectrum based on specific B-factor scores showed that all three models were of similar quality. The Global Distance Test, Test Score (GDT_TS) summarized the rigorous statistical analysis performed by MetaMQAPII and was used as an added criteria when comparing the three models.

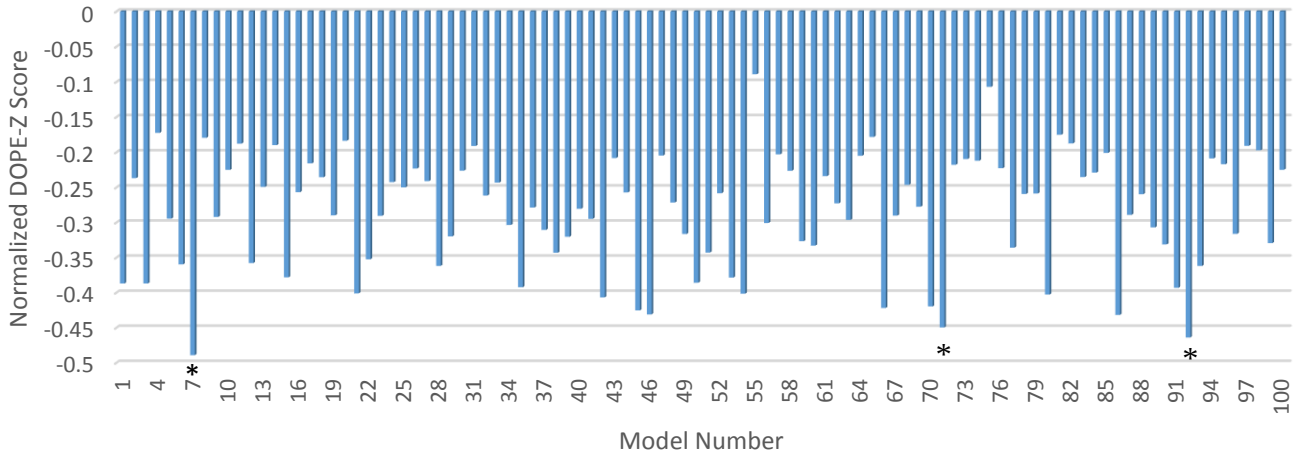


Figure 2-7: Graphical representation of the normalized DOPE-Z scores for each of the 100 homology models, showing the variance in score per model. The three highest scoring models are clearly identified (*).

The findings of this evaluation and validation analysis are summarized in Table 2-2, where both the DOPE-Z and GDT_TS scores are reported. The data shows that model 7 had the best DOPE-Z score of -0.48909 and that model 71 had the highest DOPE-Z score of -0.44952. MetaMQAPII analysis however showed model 92 to be the best of the three models, with the highest GDT_TS score of 64%. Having the second best DOPE-Z score and the highest GDT_TS score, model 92 was selected for all further analyses in this study.

Table 2-2: DOPE-Z and MetaMQAPII evaluation scores for closed conformation homology models

Model No	DOPE-Z score	GDT_TS
7	-0.48909	61 %
71	-0.44952	58%
92	-0.46398	64%

2.7.2.3 Construction of open conformation alignment file

As previously mentioned it was only necessary to model a single human HSp90 monomer in open conformation, for the construction of human Hsp90 in a dimerized open conformation. It is important to note that while this procedure is essentially the same as that previously discussed, certain domains, such as the NTD, have completely different conformations when the monomers are arranged in open conformation. To demonstrate this, the N-terminal domains of 2CG9 and 3T10 are aligned to one another in cartoon representation in Figure 2-8, the conformational differences between the two terminals can be clearly observed (red arrows). In closed conformation (green), the residues PHE 124 – ASN 151, of the helical loop segment, appear to fold over the ATP binding pocket. In open conformation (cyan) this is not the case as ATP is yet to bind, and thus this same helical loop segment is folded back exposing the ATP binding pocket. To ensure that the correct conformation was retained in the human homology model, only 3T10 was used as a template (Figure 2-9) for modelling the NTD. As before, both 2CG9 and 2CGE were used for modelling the remainder of the monomer. The template 201U was unnecessary in this instance and was thus left out of the alignment file.

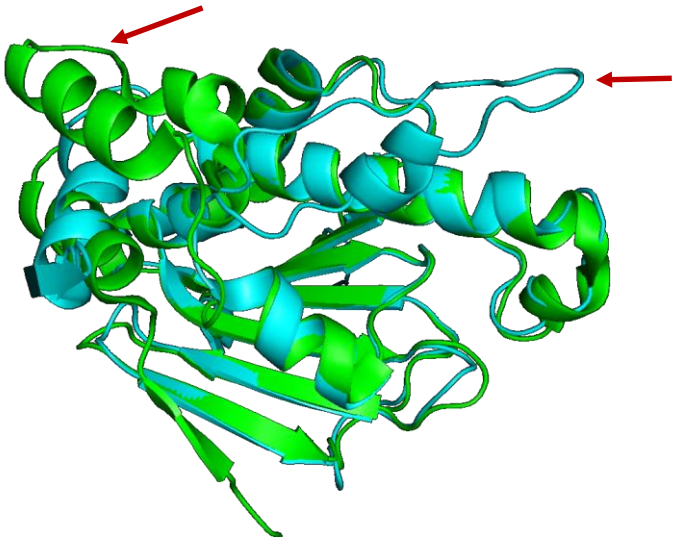


Figure 2-8: Aligned N-terminal domains of closed conformation 2CG9 (green) and open conformation 3T10 (cyan) showing the conformational differences in their respective lid arrangements (red arrows)

```
Target   15 EVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIRYESLTDPSKLDSGKELHINLIPNKQDRTLTIVDTGIGMTKADLINNLGTIAK 111
2CG9A
3T10     16 EVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIRYESLTDPSKLDSGKELHINLIPNKQDRTLTIVDTGIGMTKADLINNLGTIAK 112
2CGE

Target  112 SGTKAFMEALQAGADISMIGQFGVGFYSAYLVAEKVTVITKHNDDEQYAWESSAGGSFTVRTDTGEPMGRGTKVILHLKEDQTEYLEERRIKEIVKK 208
2CG9A     2                                                                 LDENERIGRGTILRLFLKDDQLEYLEEKRIKEVIKR  37
3T10    113 SGTKAFMEALQAGADISMIGQFGVGFYSAYLVAEKVTVITKHNDDEQYAWESSAGGSFTVRTDTGEPMGRGTKVILHLKEDQTEYLEERRIKEIVKK 209
2CGE

Target  209 HSQFIGYPITLFVEKERDKEKEEKDQEELNKTKPIWTRNPDDITNEEYGEFYKSLTNDWEDHLAVKHFSVEGQLEFRALLFVPRRAPFDLFENRKKK 305
2CG9A    38 HSEFVAYPIQLVVTKEVEKEEEVQEIEELNKTKPLWTRNPSDITQEEYNAFYKSISNDWEDPLYVKHFSVEGQLEFRAILFIPKRAPFDLFESKKKK 134
3T10    210 HSQFIGYPITLFVE                                                                                   223
2CGE    273                 TKPLWTRNPSDITQEEYNAFYKSISNDWEDPLYVKHFSVEGQLEFRAILFIPKRAPFDLFESKKKK 338

Target  306 NNIKLYVRRVFIMDNCEELIPEYLNFIRGVVDSEDLPLNISREMLQQSKILKVIRKNLVKKCLELFTELAEDKENYKKFYEQFSKNIKLGIHEDSQN 402
2CG9A   135 NNIKLYVRRVFITDEAEDLIPEWLSFVKGVVDSEDLPLNLSREMLQQNKIMKVIRKNIVKKLIEAFNEIAEDSEQFEKFYSAFSKNIKLGVHEDTQN 231
3T10
2CGE    339 NNIKLYVRRVFITDEAEDLIPEWLSFVKGVVDSEDLPLNLSREMLQQNKIMKVIRKNIVKKLIEAFNEIAEDSEQFEKFYSAFSKNIKLGVHEDTQN 435

Target  403 RKKLSELLRYYTSASGDEMVSLKDYCTRMKENQKHIYYITGETKDQVANSAFVERLRKHGLEVIYMIEPIDEYCVQQLKEFEGKTLVSVTKELELPE 499
2CG9A   232 RAALAKLLRYNSTKSVDELTSLTDYVTRMPEHQKNIYYITGESLKAVEKSPFLDALKAKNFEVLFLTDPIDEYAFTQLKEFEGKTLVDITKDFELEE 328
3T10
2CGE    436 RAALAKLLRYNSTKSVDELTSLTDYVTRMPEHQKNIYYITGESLKAVEKSPFLDALKAKNFEVLFLTDPIDEYAFTQLKEFEGKTLVDITKDFELEE 532

Target  500 DEEEKKKQEEKKTKFENLCKIMKDILEKKVEKVVVSNRLVTSPCCIVTSTYGWTANMERIMKAQALRDNSTMGYMAAKKHLEINPDHSIIETLRQKA 596
2CG9A   329 TDEEKAEREKEIKEYEPLTKALKEILGDQVEKVVVSYKLLDAPAAIRTGQFGWSANMERIMKAQ                    AKTFEISPKSPIIKELKKRV 412
3T10
2CGE    533 TDEEKAEREKEIKEYEPLTKALKEILGDQVEKVVVSYKLLDAPAAIRTGQFGWSANMERIMKAQALRDSSMSSYMSSKKTFEISPKSPIIKELKKRV 629

Target  597 EADKNDKSVKDLVILLYETALLSSGFSLEDPQTHANRIYRMIKLGLG                                                  643
2CG9A   413 DEGAQDKTVKDLTKLLYETALLTSGFSLDEPTSFASRINRLISLGLN                                                  459
3T10
2CGE    630 DEGAQDKTVKDLTKLLYETALLTSGFSLDEPTSFASRINRLISLGLN                                                  676
```

Figure 2-9: Graphical representation of the PIR file alignment used for modelling a single monomer in open conformation, showing how each template was used to map structural data to the target sequence (red).

2.7.2.4 Model building and evaluation of single open conformation monomer

DOPE-Z evaluation of 100 unique homology models indicated the top scoring model to be model 44, with a score of -0.83687, followed by models 49 and 32 (Table 2-3). As before, these models were all submitted to the MetaMQAPII webserver for further analysis. The PDB files returned were visually inspected and critically analysed by b-factor colouring (Figure 2-10). Model 44 again recorded the best score GDT_TS of 72 %, and thus this model was selected for protein-protein docking to generate an open conformation model.

Table 2-3: DOPE-Z and MetaMQAPII evaluation scores for open conformation single monomer homology models

Model No	DOPE-Z score	GDT_TS
32	-0.780890	64 %
44	-0.836873	72 %
49	-0.796888	67 %

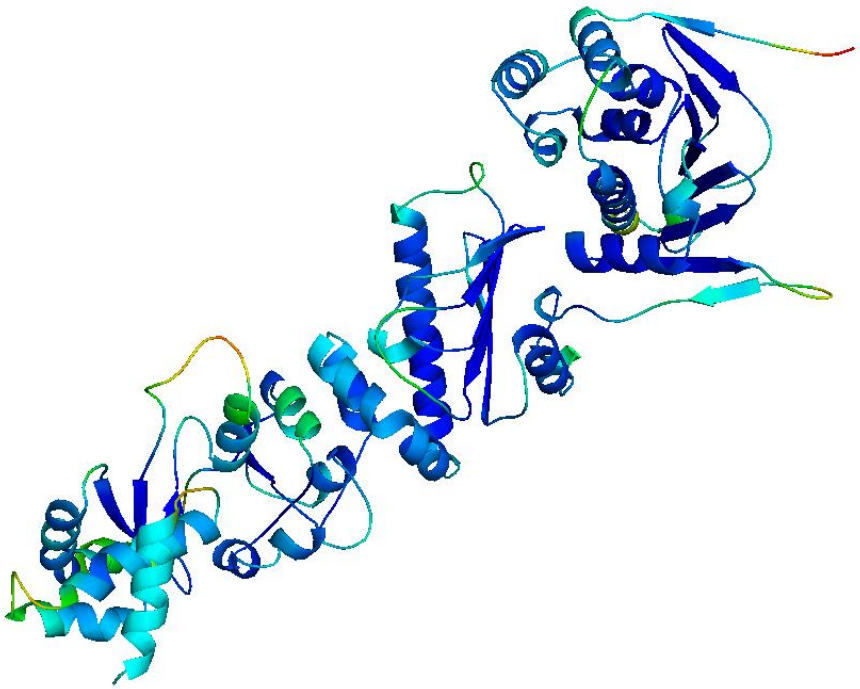


Figure 2-10: MetaMQAPII analysis of model 44, showing good regions (blues) and poorer quality regions (yellow and red)

2.7.2.5 *Protein – Protein docking of single Hsp90 monomers*

In an attempt to produce an accurate representation of Hsp90 in open dimerized conformation, the protein-protein docking server ClusPro 2.0 was used to dock two molecules of model 44. Previous work by Rastelli (2014) reported that the residues His-610, His-622, Ile-626, Tyr-627 and Ile-630 are important dimerization hotspots in the C-terminal domain of Hsp90. In order to dock two proteins to one another in ClusPro 2.0, it is necessary to indicate which residues are attractive in the receptor protein and ligand protein respectively. As such the above mentioned dimerization residues, were listed as the attractive residues in both chains.

Of the 22 clusters returned by ClusPro, only two were selected for further analysis, based solely on their visual orientation, where the criteria was the open "V" like structure (Figure 2-14) reported by Shiau et al. (2006). A summary of these results is shown in Table 2-4. Based on the centred energy score of -1168.2, and despite the inferior number of members, the model represented by cluster 5 was chosen to be the best representation of Hsp90 in open conformation and as such was selected for all further experimentation.

Table 2-4: Summary of ClusPro protein-protein docking results

Cluster	Members	Energy score
2	90	-1138.7
5	55	-1168.2

2.7.3 Energy minimization

All proteins have a net charge, which arises from its specific amino acid composition and in particular the amino acid side chains present. The charge on each amino acid side chain is dependent on the pH of the solution it is in as well as the pK_A of the side chain. It thus stands to reason the localized environment around a side chain will affect its charge. In a pH which is less that the pK_A of the side chain results in its protonated form while a pH higher than the pK_A results in the deprotonated form. Side chains in their protonated from leaves the acidic side chains with a charge closer to 0 and the basic side chains a charge approaching a limit of 1. The converse is

true for side chains in a deprotonated form, with the charge of acidic side chains approaching -1 and basic side chains 0. A protein's charge is thus the sum of charges on the individual amino acid side chains. The tertiary structure and final fold of a protein is highly dependent on its charge composition. Thus far the charge effect on the tertiary structure of the homology models described in this chapter was not taken into account. GROMACS energy minimization was thus introduced at this point in an attempt to neutralize the net charge on both protein models, allowing them to 'relax' into lower energy and more native like states. This was accomplished by placing the protein into a standard water solution, and adding an equal number of ions of opposite charge to a calculated net charge. By allowing time to pass, the protein structures were given the necessary time required to neutralize their net charge and obtain lower energy states.

The net charge for the closed conformation model was found to be -18.00 while the open conformation model was found to be -34.00. These charge deficits were countered by adding

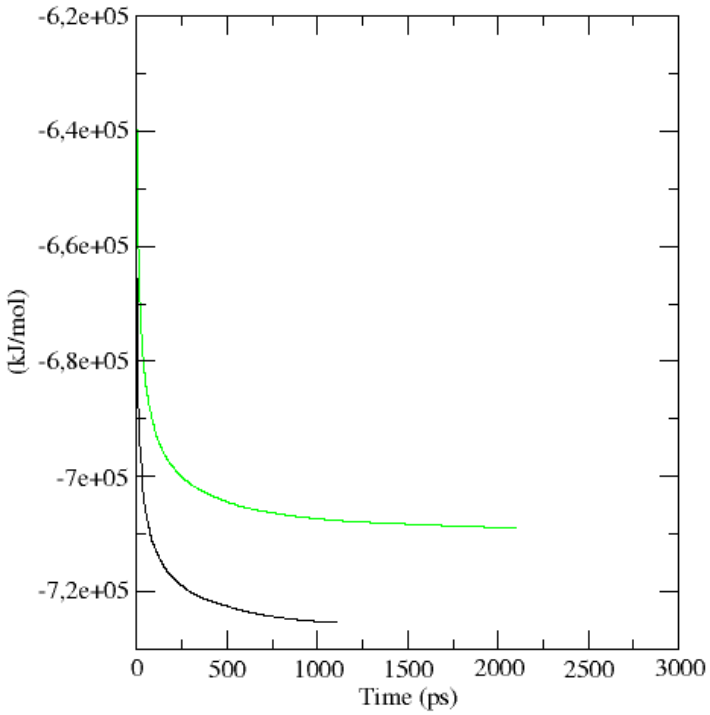


Figure 2-11: Energy potentials of open (green) and closed (black) conformation models plotted over time, showing the relative decrease in energy.

and equal number of positive sodium ions to the solvent. The curves in Figure 2-11, show the exponential decrease in energy for each model over time. Interestingly the closed conformation model achieved a much lower energy (-7.25×10^5) than that of the open conformation model (-7.05×10^5). The final models after this energy minimization process were completely devoid of steric clashes and any other geometric discrepancies in relation to the net charge.

MetaMQAPII analysis of these energy minimized models neatly summarized the structural improvements energy minimization introduced to the respective models, as shown in Figure 2-12. The structures of several helical and loop regions dramatically improved as indicated by the shift in colouring of poor scoring regions in red to relatively well represented regions in green/cyan. In fact recalculation of the DOPE-Z score for the closed conformation model (model 92) revealed an improvement of more than double with the new score sitting at -1.100. A score widely accepted as more than adequate.

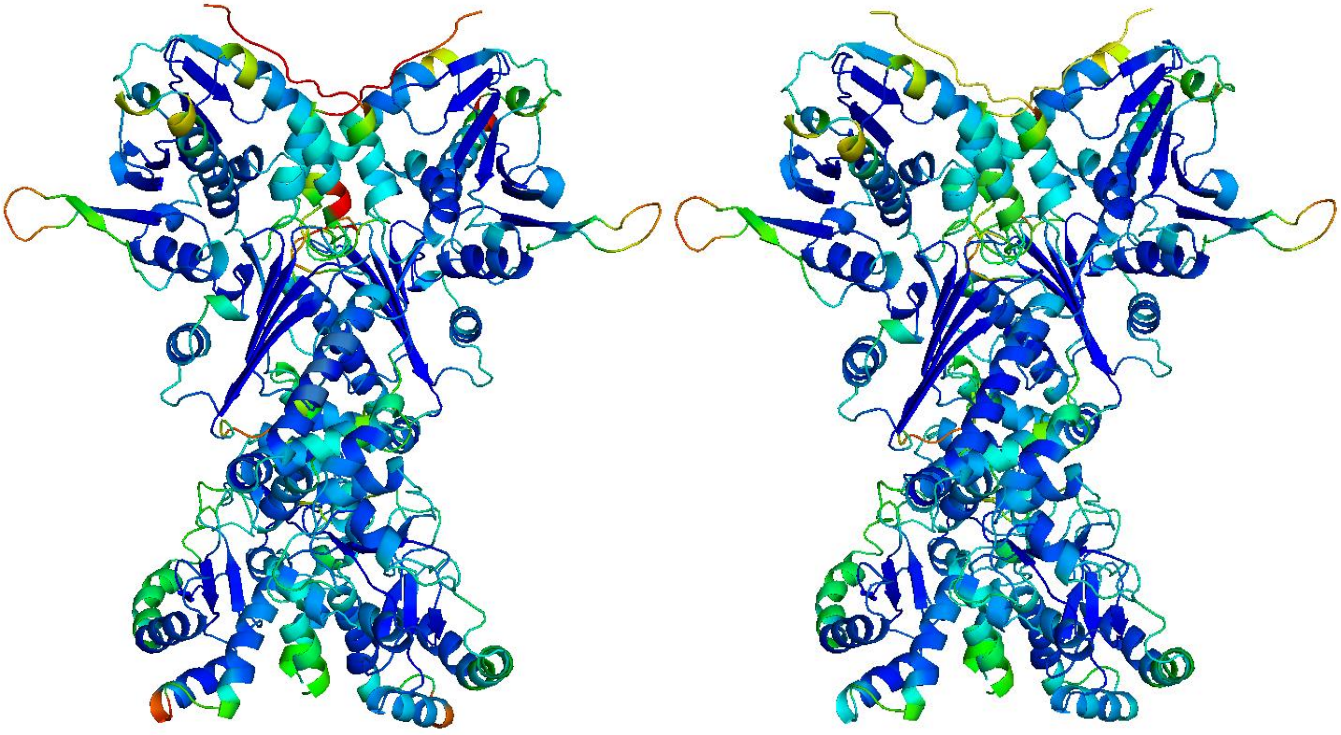


Figure 2-12: MetaMQAPII representation of closed conformation models comparing un-minimized (**A**) and minimized (**B**), showing good quality regions (blues) and poorer quality regions (yellow and red)

2.8 Chapter conclusions

At the commencement of this chapter, several objectives where discussed and laid out. This following Section summarizes the findings for each of these objectives and allows for some final criticism of the work thus far achieved.

As is now clear to the reader, the underlying objective of this chapter was to model human cytosolic Hsp90 in two of its functionally important conformations. Fortunately the full structure of yeast Hsp90 has been solved albeit at a very poor resolution. This structure however provided an excellent starting point and foundation for all the homology modelling to follow. By including several other structural templates to improve the quality of each domain, a full sequence structure of human Hsp90 in closed conformation was constructed bar a 51 residue loop insert (Figure 2-13).

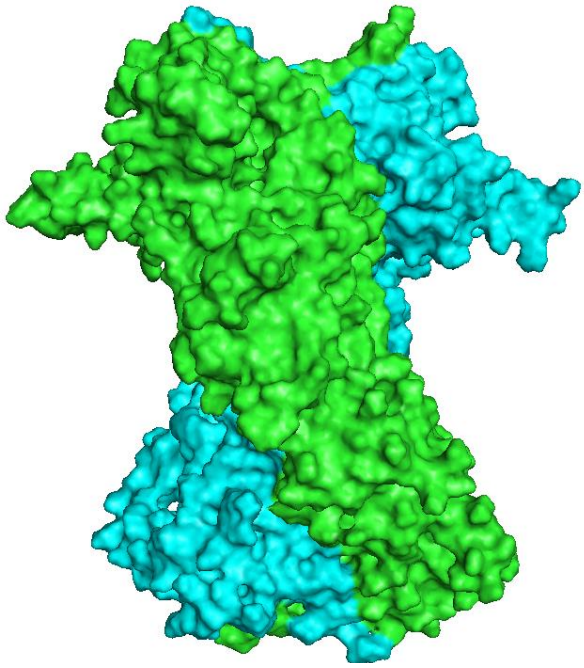


Figure 2-13: Homology model of human Hsp90 in closed nucleotide bound conformation, as depicted in PyMOL using surface representation showing chains A (green) and chain B (cyan)

In accordance with the second objective, this model was fully evaluated and validated using the MetaMQAPII webserver. Overall the results from this analysis showed that the model is statistically significant, and that the only the loop regions showed evidence of poor modelling and statistical inaccuracies. This is not surprising however as loop regions are regarded as being

the most flexible regions of any protein. Having no particular tertiary structure, loop regions are known to flex and rotate allowing the protein to undergo conformational changes. These regions however are very rarely constituents of a protein's active site and their absolute accuracy is thus not necessary. In this closed conformation model, none of the loop regions displayed exceptionally poor quality as is indicated by the positive colouring in Figure 2-12, and all were deemed adequate.

As previously mentioned, this model represents the entire structure of human Hsp90, except for a 51 residue insert located in the middle domain. Despite an in depth search, no structural template could be identified for this insert, and as such modelling of this region was not possible. It would appear from the model (Figure 2-12), that this particular region is most likely a loop region, however there is no way of confirming this surmise, until structural data is obtained. It is interesting to note however that this insert is located in the middle domain, and more importantly within the flexible linker region. It is thus not surprising that all attempts at crystallising this region have failed, being so inherently flexible, these regions are notoriously difficult to crystallise.

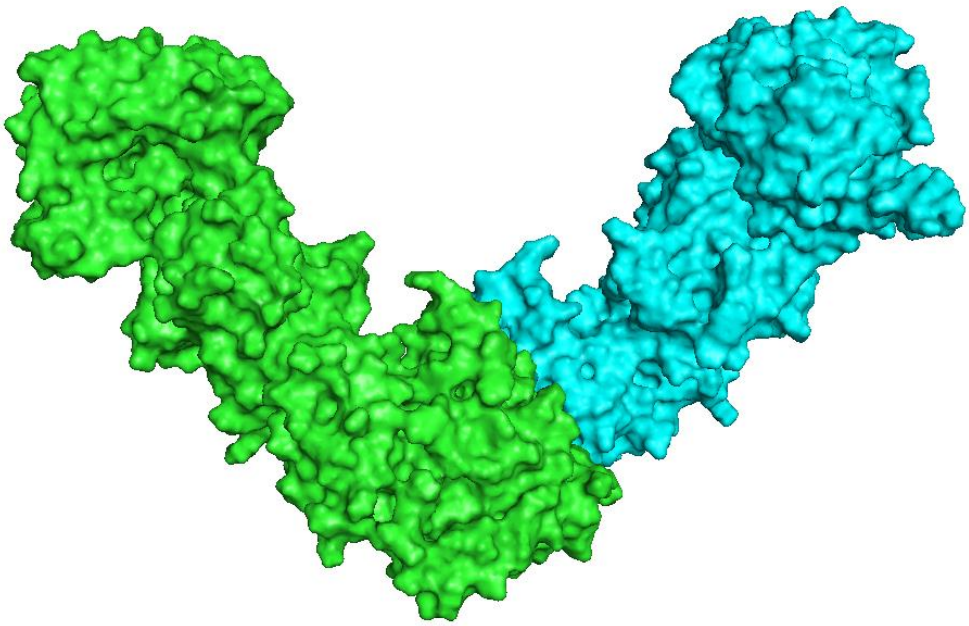


Figure 2-14: Homology model of human Hsp90 in open "v-like" conformation as depicted in PyMOL using surface representation showing chains A (green) and chain B (cyan)

When Hsp90 is in its open conformation, it is regarded to be at its most functionally important stage in its multi-conformational cycle. At this point, the newly synthesised monomers dimerize at the C-terminal domain, the result being an open V-like conformation. In this conformation the interior surface of the respective monomers are exposed and ready to accept client protein binding. This stage is regarded as being the only time Hsp90 accepts new client proteins and as such is critical for its housekeeping function. Given this importance, an open conformation model of human Hsp90 was modelled.

Given the success in modelling Hsp90 in closed conformation using the template 2CG9 as a primary template, it was unfortunate that there exists no structure of Hsp90 in open dimerized conformation. As such the modelling process of the generation of this model was somewhat more complex than that used for the closed conformation. *In vivo* the monomers of Hsp90 dimerize to form active Hsp90. This is a fairly uncomplicated reaction and in theory, given the current bioinformatics techniques available is possible to simulate *in silico*. Homology modelling was used to accurately model a single Hsp90 monomer. The yeast template 2CG9 again provided much of the necessary structural data, and chain B was used as a primary template, as it contained more structural information per residue than chain A. The human N-terminal domain template 3T10 was used as the sole N-terminal template, to ensure the ATP binding domain remained open and fully exposed, this is not the case when in closed conformation (see Figure 2-8). To improve on the poor quality of 2CG9 the template 2CGE was included for a large part of the middle domain and the entire C-terminal domain. It should be noted that, as with the closed conformation, the 51 residue insert in the middle domain couldn't be modelled and as such was removed from the structural alignment. MetaMQAPII analysis provided evidence that the quality of this monomeric model was very similar to that achieved for the dimerized closed model.

Thus with a fairly accurate monomer in hand, the methods and techniques for *in silico* protein-protein docking were successfully carried out using the ClusPro 2.0 webserver docking pipeline (Comeau et al. 2004). To ensure that the monomers interacted at the C-terminal alone, the residues His-610, His-622, Ile-626, Tyr-627 and Ile-630 (Rastelli 2014), were specified as interactive residue pairs between the two chains. After a close inspection and analysis of the

returned clusters a single dimerized model was selected. The accumulative result of this combined homology modelling and protein docking is shown in Figure 2-14.

CHAPTER 3: MOLECULAR DOCKING

The research in this chapter is based on the homology models calculated in Chapter 2. This chapter is preceded by a brief explanation and description of molecular recognition and the key players involved therein. The methodology for whole protein screening and site specific targeted docking experiments is discussed in detail followed by a detailed analysis of the results obtained.

3.1 Introduction

Proteins control a wide array of biological processes by interacting with small molecules or other proteins that bind to them. Understanding these protein-ligand interactions is of great interest, providing an opportunity to better understand protein function and further therapeutic intervention. The interaction between a protein and ligand is known as molecular recognition, and is defined by a complex combination of several factors such as inter-molecular forces between the protein, ligand and surrounding solvent, variation in conformation between binding partners and the thermodynamics of molecular association. Despite the development of experimental and computational techniques to better understand the specific role of these factors, complete understanding of molecular recognition is still a work in process. This chapter includes a brief review of some of the most important aspects of protein-ligand interactions as well as their computational prediction.

3.2 The thermodynamics of protein-ligand binding

The reversible non-covalent binding of a small molecule to a protein in an aqueous environment is best described by the reaction:

$$P_{(aq)} + L_{(aq)} \leftrightharpoons PL_{(aq)}$$

The dissociation constant KD for this reaction, where there is a single ligand binding site on the receptor that is unaffected by any competing binding sites, is defined as being the concentration of the ligand required to saturate half of the given binding sites (Dunn 2010) , as shown in the following reaction:

$$\mathrm{KD} = \frac{[\mathrm{P}] + [\mathrm{L}]}{[\mathrm{PL}]}$$

KD is therefore a measure of the affinity of a ligand towards its respective binding site on a receptor and is measured in molar units (M). A chemical reaction such as this is accompanied by a change in free energy (ΔG), which is defined and influenced by heat content enthalpy (ΔH) and the temperature-independent degree of disorder entropy (ΔS). The relationship of these three quantities is given by the equation:

$$\Delta\mathrm{G} = \Delta\mathrm{H}\text{-}\mathrm{T}\Delta\mathrm{S}$$

From this equation it is clear that ΔG is directly influenced by changes in enthalpy and entropy, changes that are manifested through factors such as ionization effects, electrostatic and van der Waals interactions, conformational changes and the role of the solvent (Dunn 2010). Changes in enthalpy are related to the breaking and formation of non-covalent interactions, such as the loss of protein and ligand solvent hydrogen bonds, and the subsequent formation of protein-ligand hydrogen bonds and hydrophobic contacts, where the strength of these interactions determine whether or not the enthalpy change is favourable (Perozzo et al. 2004). Likewise, the changes in entropy during binding are predominantly influenced by solvent displacement and reduction in conformational freedom, where the burial of lipophilic surfaces results in an increase in entropy and the restriction of protein and ligand side-chains results in the opposite. The effect of one quantity on another is clearly observed where a gain in enthalpy from new bond formations leads to precise interactions that cause structural rigidity and a decrease in entropy, a phenomenon known as enthalpy-entropy compensations (Freire 2008). For a reaction to occur spontaneously ΔG should be negative and at equilibrium it is related to the equilibrium constant in the following equation:

$$\Delta\mathrm{G} = \text{-}\mathrm{RT}\mathit{ln}\mathrm{K}$$

R representing the gas constant and T the absolute temperature. This relationship can be used to calculate the free energy of binding from experimentally determined quantities such as KD. Being dependent on concentration, very low KD constants are desired in drug design as therapeutic drugs often cause harmful side-effects at high concentrations due to off target

interactions and binding affinities within the range of 0.1 to 10 nM are considered to be suitable (Dunn 2010).

3.3 Types of protein-ligand interactions and other contributing factors

Non-covalent binding of ligands to proteins is mediated by a variety of inter-atomic interactions, however electrostatic and van der Walls interactions are the main contributors. Binding affinity also relies heavily on the contribution of entropy, desolvation, receptor protein flexibility, and water molecules present in the binding site (Bissantz et al. 2010). The following is a brief description of the most important protein-ligand interactions as well as some of the other factors that play a part in binding affinity.

3.3.1 Electrostatic interactions

Electrostatic interactions arise from charge attraction and repulsion, and in terms of molecular recognition include; hydrogen bonding, salt bridges, and metal interactions (Chowdhry & Harding n.d.) (Figure 3-1). Hydrogen bonding is considered to be the most important directional interaction in biological macromolecules, and are known to confer both stability and selectivity in protein-ligand interactions (Hubbard & Haider 2010). Hydrogen bonding occurs between two electronegative atoms, one the acceptor with a lone pair of electrons and the other the donor with a covalently bound hydrogen atom, usually a nitrogen or oxygen, and attraction arises from a partial positive charge on the hydrogen atom and a partial negative charge on the acceptor atom. The average bond length of hydrogen bonds is normally between 2.5 and 3.2 Å with angles of 130° to 180° (Hubbard & Haider 2010). The strength of a hydrogen bond depends on its directionality as well as its surroundings, where bonds in the interior of a protein are much stronger than those in solvent exposed regions, and it is the difference in strength between these two environments (solvent and binding interface) that determines to what extent hydrogen bonds contribute to binding affinity (Perozzo et al. 2004).

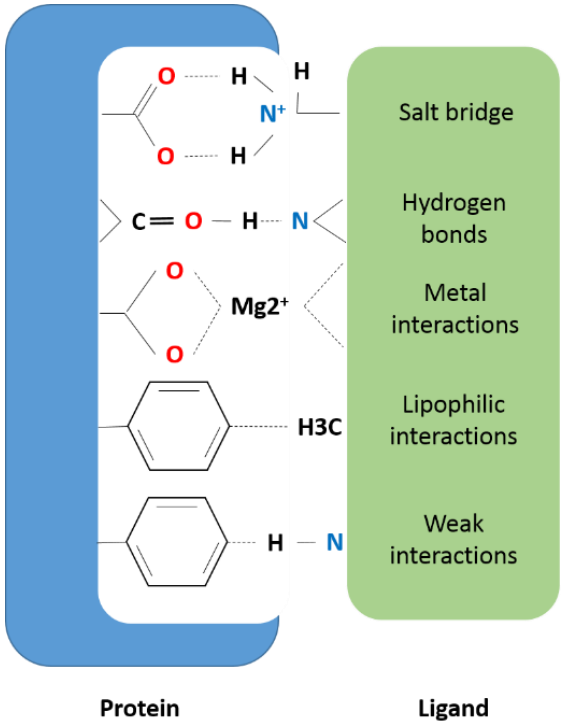


Figure 3-1: Schematic representation of the main non-bonded interaction types between a protein (blue) and ligand (green)

3.3.2 Hydrophobic interactions

Hydrophobic interactions in protein-ligand complexes involve non-polar contact between molecules (Figure 3-2). Upon binding, the interacting non-polar surfaces are buried, causing the surrounding water molecules to be displaced, increasing the entropy of the system (Bissantz et al. 2010). Buried hydrophobic contacts have been found to increase binding affinity by 30 cal mol^{-1} for 1 $Å^2$ of buried lipophilic surface area (Hubbard & Haider 2010). This suggests that optimization of non-polar contacts of ligand atoms within hydrophobic pockets of a receptor protein would result in much tighter binding (Gohlke & Klebe 2002).

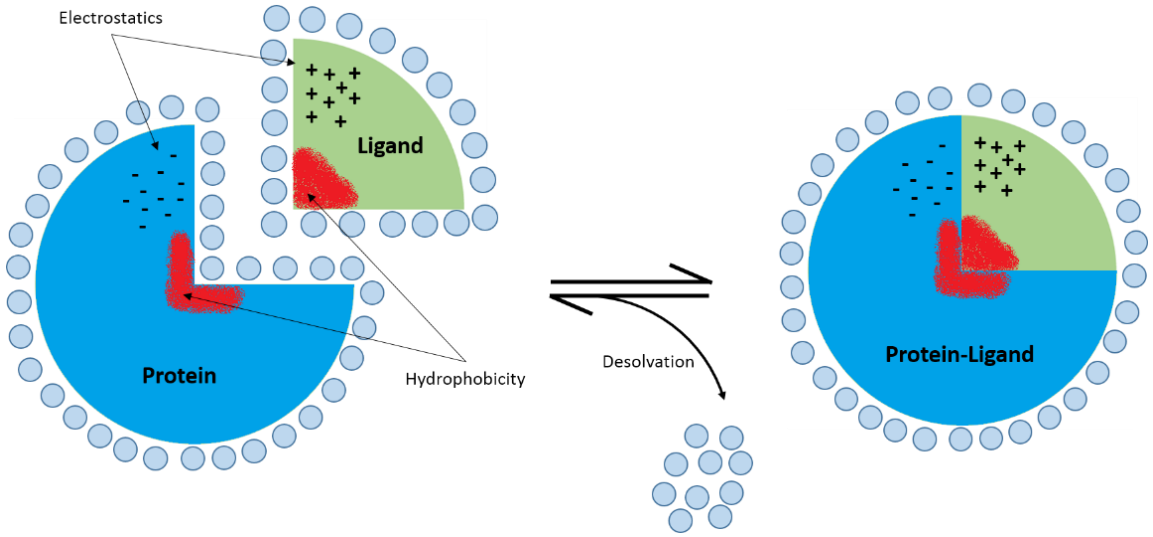


Figure 3-2: Schematic overview protein-ligand interactions between a protein receptor (blue) and ligand (green). Electrostatic, hydrophobic and desolvation factors that guide the induced protein-ligand fit are shown

3.3.3 Protein binding site water molecules

Water molecules present in the binding site of a protein play an import role in maintaining the structural integrity of interactions between biomolecules. Before ligand binding, water molecules present in the binding site must first be displaced, and the contribution of this displacement to binding affinity depends on how tightly the water molecules are bound to the receptor and how efficiently the inherent loss in enthalpy can be compensated by subsequent interactions between the receptor and ligand molecule (Bissantz et al. 2010).

3.3.4 Solvation and desolvation

Water molecules associated with a receptor protein form an extensive and dynamic hydrogen bond network in which each molecule present is involved in approximately 3 to 4 hydrogen bonds at any given time (Gohlke & Klebe 2002). The introduction of small molecule ligands that are non-polar or have non-polar parts, results in a disruption of this network and a subsequent re-organization of the water molecules around the non-polar ligand, and an overall loss in entropy, a discrepancy which must be compensated for by the formation of stronger hydrogen bonds within the water molecules. The binding of non-polar molecules in water is thus mainly by driven

the release of water molecules from the binding interface, causing an increase in the entropy of the system, a process known as the classical hydrophobic effect (Bissantz et al. 2010).

3.3.5 Protein flexibility

Another important consideration for molecular recognition is conformational flexibility. Depending on their biological function, many proteins are inherently flexible and are capable of transitioning through different conformational arrangements. When a ligand binds to a protein receptor, the binding site can undergo various conformational rearrangements ranging from side-chain side-chain rearrangements to full loop movements. In some cases, a protein can even undergo complete re-organization on ligand binding, such as nucleotide binding to Hsp90 (Shiau et al. 2006).

3.4 Prediction of protein-ligand interactions

Molecular recognition is made inherently complex given the number of contributing factors and parameters. Despite this, various computational methods and approaches for the prediction and modelling of these interactions have been developed over the past decade. Due to the complexity if of the phenomenon, the task of accurately modelling protein-ligand interactions is an extremely challenging task. Some degree of accuracy has however been achieved with predictions methods that have been developed to couple thermodynamic principles with statistical mechanics, and employ full-scale molecular dynamics simulations. These various approaches and methods can be succinctly summarized into three main groups; Free energy methods, Molecular Mechanics with Poisson-Boltzmann and Surface Area (MM-PBSA) and Molecular Mechanics with Generalized Born and Surface Area (GBSA) methods, and Docking and scoring methods (Mobley & Dill 2009). The following is a brief discretion of each, in order of decreasing accuracy and computational demand.

3.4.1 Free energy calculations

The free energy calculation approach is often referred to as “computational alchemy”, in that they evaluate the difference between the binding energy of two similar ligands. Pathways are used to calculate the change in energy when one ligand is changed to another within the same

solvated binding site (Tembe & McCammon 1984). In free energy calculation approaches, either the absolute or relative binding energies can be calculated. Absolute binding free energies are considered to be more accurate but are computationally expensive, requiring individual separate simulation runs for the solvated protein, the ligand and the protein-ligand complex. The method doesn't require any prior information about structure or binding affinity of the complex (Gohlke & Klebe 2002). In relative free energy calculations, known structures for the complex are required for use as a reference, and the difference in the binding free energy is calculated for the ligand of interest.

3.4.2 MM-PBSA and MM-GBSA methods

MM-PBSA and MM-GBSA were first implemented in 1990 and have since been extensively developed (Gohlke & Klebe 2002). Both methods are based on the assumption that free energy can be decomposed into several individual terms, each of which describe different important contributions to protein-ligand binding (Gohlke & Klebe 2002). The overall sum of these energetic terms (intra-molecular, van der Walls, electrostatic interactions and solvation) is separately calculated for the protein, ligand and protein-ligand complex using molecular mechanics force-fields. The Poission-Boltzmann (PB) and Generalized Born (GB) approaches are implemented in consideration of implicit solvent. PB is the more rigorous of the two, but is also more computationally expensive. The GB methods are based on an approximation to the PB equation (Feig & Brooks 2004).

3.4.3 Docking and scoring

Docking and scoring methods were designed for high throughput computation of binding affinities, where the degree of accuracy is traded for an increase in computational efficiency. This methodology involves the generation of a set of different poses for a given ligand that fit into a specific binding site, which are then ranked according to a specific scoring function (Mobley & Dill 2009). A number of different scoring functions have been developed over the past few years and can be divided into the empirical, knowledge-based and force-field based methods.

3.4.3.1 Empirical scoring

Empirical scoring functions are based on the addition of all the factors (terms) contributing to the total binding enthalpy (Gohlke & Klebe 2002) and are fitted to reproduce experimentally determined energies. These terms include the important contributions made by hydrogen bonds, hydrophobic and ionic interactions and an entropic term that accounts for any loss in conformational freedom. Weighted coefficients for each parameter are derived using regression analysis from experimentally determined binding affinities. In general a scoring scheme would take the following form:

$$\Delta G = \sum_i f_i \Delta G_i$$

Where f and G respectively denote the coefficient and free energy associated with the interaction term i. While these functions have the advantage of having a very simple functional form, the additivity of terms pose a major disadvantage in that larger ligands will always get a higher score than smaller ligands (Schulz-Gasch & Stahl 2004). A well-established example of this scoring function is see in the software Chemscore (Eldridge et al. 1997).

3.4.3.2 Knowledge-based scoring

The knowledge-based soring functions are rooted in inverse Boltzmann law. This scoring function focuses on reproducing actual structures rather than particular energies. They are based on atomic interaction-pair potentials derived from contacts observed in ligand-protein complex structures available in databases such PDB, giving rise to the fundamental premise that the frequency of a particular structural arrangement between two atom types is directly related to its specific energy. While these methods tend to be simple to compute efficiently, a drawback is that the number of protein-ligand complexes available to derive parameters is limited and it is therefore necessary to find a balance between well-defined atom-types and chemical diversity of interactions (Kitchen et al. 2004). DrugScore (Gohlke et al. 2000) is a very well established example of knowledge-based scoring.

3.4.3.3 Force-field scoring

Force-field scoring doesn't require the need for any specific parameterization, and makes use of well-established molecular mechanics force fields to estimate the binding energy of non-bonded protein-ligand interactions (Kitchen et al. 2004). An example of force-field scoring is seen in AutoDock (Morris et al. 1998) and G-Score (Kramer et al. 1999). Non-bonded energy terms are pre-calculated on a grid and interpolated to positions in the model where atoms in protein-ligand complexes are located. The effects of solvent on the system is approximated by applying a distance-dependent dielectric constant and long-range shielding of electrostatic interactions are accounted for by combining the scoring function with PBSA or GBSA approaches.

3.5 Chapter Objectives

The overriding objective of this chapter was the screening of natural compounds isolated from organisms indigenous to South Africa, against human Hsp90 using *in silico* molecular docking techniques. The work described herein covers several key sub-objectives, each of which contributed to the main objective; 1) A mini database of small molecule natural compounds was constructed, based solely on published literature, each compound being carefully optimized, geometrically and energetically for accurate used in *in silico* screening experiments; 2) A blind docking protocol was designed to implement whole protein screening of the aforementioned compound database against the homology models of human Hsp90; 3) Putative novel target sites on the surface of Hsp90, in open and closed conformation, were identified using quantitative techniques, and 5) Refined targeted docking studies and subsequent analysis of these putative target sites was carried out, to provide a detailed description and tentative characterization of the binding pocket; 6) Bound ligands at target sites were further analysed for the identification of potential lead compound candidates for future drug development.

3.6 Methodology

3.6.1 Construction of mini South African compound database

The construction of this mini database was done in conjunction with the construction of a full scale database, South African Natural Compounds Database (SANCDB), by Professor Tastan Bishop's research group (RUBi). SANCDB is a fully referenced database and as such the 316 compounds in the mini database and subsequently added to SANCDB (including 100 compounds personally added), were fully referenced being extracted directly from literature. The major criterion for compound addition was that the source organism was indigenous to South Africa, and compound addition to the database followed several steps/stages in a stage-by-stage completion manner (Figure 3-3).

The first step was compound identification in literature and the extraction its published name, with was then submitted to SciFinder (Wagner 2006) in the second step to obtain an accurate 2D chemical structure, which was saved as an image file. For compounds that had no published name, the referencing publication title was submitted to SciFinder and the chemical structure identified and named using SciFinder's naming convention.

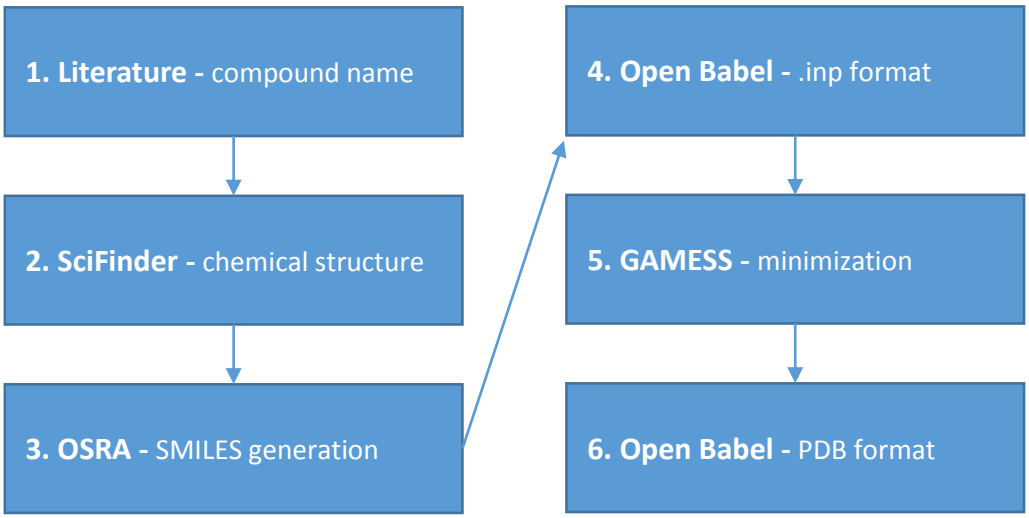


Figure 3-3: Overview of sequential steps required for compound database addition. A compound that failed any of these stages was rejected

In step 3 the saved chemical structure images were submitted to the Optical Structure Recognition Software tool OSRA (Filippov & Nicklaus 2009), which was used to generate an appropriate SMILE for each respective compound. The compound SIMLES were then converted to GAMESS input format using Open Babel (O'Boyle et al. 2011) in step 4. The GAMESS (Schmidt et al. 1993) minimization software was used to perform energy minimization of each compound in step 5 and the final energy minimized compound structure saved in PDB file format using Open Babel in the final step.

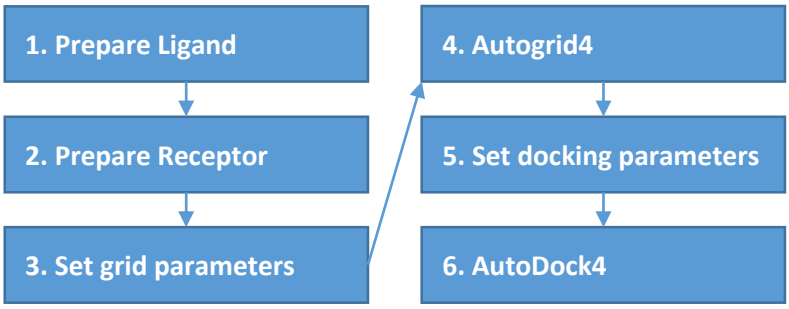


Figure 3-4: Overview of ligand and receptor protein preparation for AutoDock4

3.6.2 Molecular docking procedures

All molecular docking experiments were completed using AutoDock4 and the AutoDock tools suite (ADT) (Morris et al. 2009). In all instances the technique of rigid docking was used, whereby the protein was held rigid and the ligand allowed conformational flexibility. Rigid body docking was utilized due to the large number of ligands as well as the large protein size. Allowing for protein flexibility would demand an excessive computational cost. Molecular docking with AutoDock4 can be broken down into six steps (Figure 3-4).The two preparation steps (1 and 2 Figure 3-4) were completed using a custom made python script which incorporated the ADT scripts prepare_receptor4.py and prepare_ligand4.py, to convert the human homology models of Hsp90 in open and closed conformation as well as each compound ligand in the compound database to PDBQT file format. In preparation for Autogrid4 (step 3), a grid parameter file was defined using the ADT script prepare_gpf4.py and to account for as many ligand types as possible, the ligand types parameter was included and defined as follows; H, HD, C, A, N, NA, P, S, SA, Br, I, OA and CL. In preparation for the docking run with AutoDock4 (step 6), docking parameters were define using the ADT script prepare_dpf42.py and user defined parameters passed in TXT file format (Appendix B-1).

3.6.2.1 Blind docking parameters

As there is no prior knowledge of any active site in blind docking, the grid was centred on the receptor's center of mass using the parameter 'auto', a technique established by Hetényi & van der Spoel (2002). To ensure that the whole receptor was encapsulated within the grid box the spacing parameter was set to 1 Å and the number of points X = 126, Y = 126, Z = 126. Specific docking parameters were used in all blind docking runs; population = 50, evaluations = 10 million, number of generations = 10 million, and number of runs = 256.

3.6.2.2 Targeted docking parameters

Given the prior knowledge of the location of putative binding pockets, grid centres and grid parameters for each targeted docking run could be specified according to the size and location of the binding pocket under investigation (Table 3-1). The docking parameters in these runs were increased from those used in the blind docking screening; population = 150, evaluations = 10 million, number of generations = 10 million and number of runs 100.

3.6.3 Control docking procedure

The N-terminal crystal structure of human Hsp90 (PDB code 3T10) was retrieved from the protein data bank. The ligand molecule ACP was separated from protein receptor using PyMOL, saving each molecule being in separate PDB files. All water molecules and co-factors were removed from the receptor PDB in a text editor. The full human Hsp90 homology model was edited in PyMOL such that only the N-terminal domain was present, saving this structure to a separate PDB file. The 3D orientation of the homology modelled N-terminal domain was matched to that of the experimentally solved N-terminal crystal structure 3T10 by aligning the two proteins in PyMOL. The blind docking procedure previously mentioned was carried out on each of the two apo-proteins, using the known inhibitor ligand ACP as a positive control. The only change to the aforementioned methodology being the grid box dimensions, which were reduced relative to the smaller N-terminal domain. Spacing = 0.375 Å, number of points X = 90, Y = 90, Z = 90.

3.6.4 Docking analysis and scoring

3.6.4.1 Blind docking analysis

Whole protein blind docking screening generated a single docking log file (DLG) for each ligand. Each log file was summarized accordingly using the write_lowest_energy_ligand.py and write_largest_cluster_ligand.py ADT scripts and the resulting ligands visually inspected in conjunction with the respective receptor using the PyMOL molecular viewer and the ADT PyMOL plug-in (Seeliger & de Groot 2010; Trott & Olson 2010).

3.6.4.2 Targeted docking analysis and scoring

Ligands docked to specific target sites were filtered according to their binding energy scores alone, in a series of filtering steps (Figure 3-10). The ADT script summarize_docking.py was used to generate a summary text file for each ligand, containing the energy scores for each geometric pose calculated by AutoDock4. A custom python script was used to sort and extract the best scoring ligand pose for each ligand, generating a list of best scoring poses for each ligand. From this list only the 10 best scoring ligands were retained for further analysis with LigPlot+ (Wallace et al. 1995) and X-Score (Wang et al. 2002).

In preparation for analysis with LigPlot+, each ligand was written to PDB file format using the ADT scripts write_lowest_energy_ligand.py and pdbqt_to_pdb.py. Each ligand PDB was viewed with its respective receptor in PyMOL and a PDB file containing the receptor and ligand in complex written. X-score preparation required the conversion of each ligand from PDB file format to MOL2 file format, which was achieved using Open Babel. The parameters used for X-score analysis can be seen in Appendix B-2. The reference ligand-receptor complex was omitted and all other parameters set to the default settings.

3.7 Results and discussion

3.7.1 South African natural compounds database

In conjunction with colleagues from RUBi, 350 natural compounds isolated from indigenous South African organisms, were added to the mini-database. Only compounds less than 600 KDa were selected for further study, reducing the total number investigated to 316 compounds. The chemical geometry for each compound was defined using SMILES format obtained from chemical drawings with OSRA. The SMILES format can be converted to PDB file format and viewed graphically in a molecular viewer such as PyMOL and was done to critically analyse each compound, checking for any inaccuracies or inconsistencies, such as stereochemistry and protonation. Even after manual optimization, the chemical description of each compound needed to be calculated to ensure the compounds were optimized in terms of energy and global minima. The minimization software GAMESS was used for these calculations, and refinement was calculated at the semi-empirical level (AM1). Atom charges were calculated using single point calculations at the HF/6-31G(d) level. The fairly large dataset of 316 compounds meant that higher levels of theory would have been too computationally demanding. Of the 316 compounds selected for screening, 311 were successfully modelled and optimized for later use in this study.

3.7.2 Control study for the validation of blind docking procedure

In this experiment, the docking procedures mentioned in Section 3.7.2 were tested and validated in a controlled docking study, which provided a suitable platform to test and gauge how well AutoDock4 performs in finding a putative binding pocket on a receptor without prior knowledge. It also provided an opportunity to further validate our Hsp90 homology models, the primary target in this project.

The N-terminal human Hsp90 crystal structure 3T10, a template structure previously used in the homology modelling procedure (see Section 2.2.1), was co-crystalized with the ATP binding domain inhibitor ACP. As mentioned in the previous chapter, 3T10 was solved with a high level of detail at a resolution of 1.47 Å. Given the presence of a co-crystalized inhibitor and the good quality, this crystal structure provided an excellent model for a control study. The ACP inhibitor

was removed from complex and, using a blind docking technique, re-docked to the apo-protein. The resulting complex was then visually compared to the original crystal structure for evaluation. The re-docked experiment showed that AutoDock4 is indeed capable of finding the ATP active site on Hsp90.

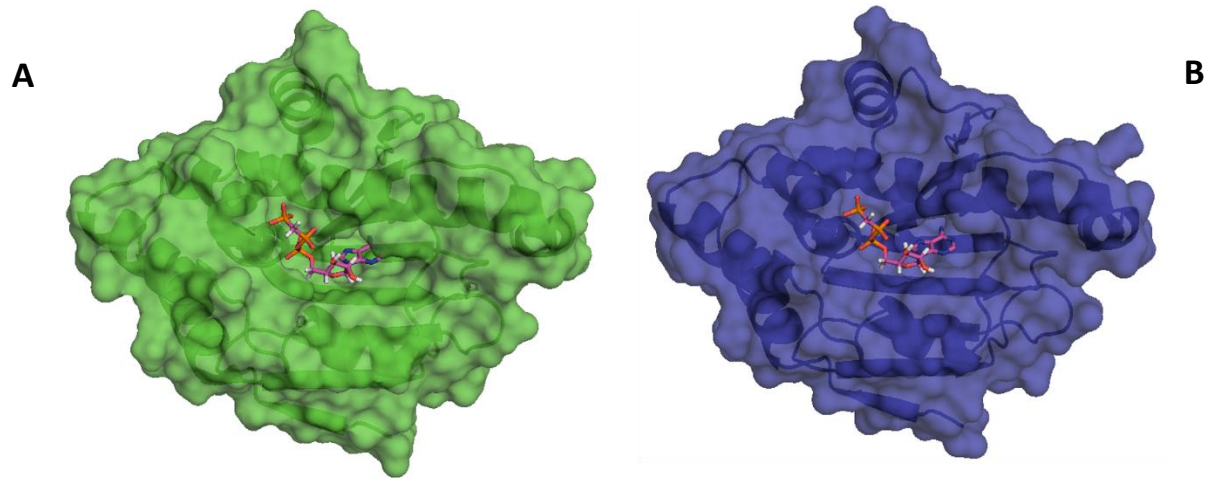


Figure 3-5: Comparison of blind docking of ATPase inhibitor ACP against the NTD ATP binding pocket of human Hsp90 (3T10) (**A**) and human homology model in open conformation (**B**). Ligand binding in both showed near identical binding conformations

The next step in this experiment was to repeat the methodology using the N-terminal domain of our human Hsp90 homology model in open conformation, as the receptor molecule. Being a homology model however, this structure lacked any water molecules or cofactors, and thus these factors were removed from the original crystal structure and the inhibitor re-docked as before. Visual analysis again indicated that the inhibitor in both cases bound to the active site and in similar orientations to boot (Figure 3-5). LigPlot+ analysis gave an indication as to which residues in the active site were involved in binding. Comparing the LigPlot+ results from both the control and the homology model showed identical sets of interacting residues between the two, bar the additional Glu32 in the homology model (Figure 3-6).

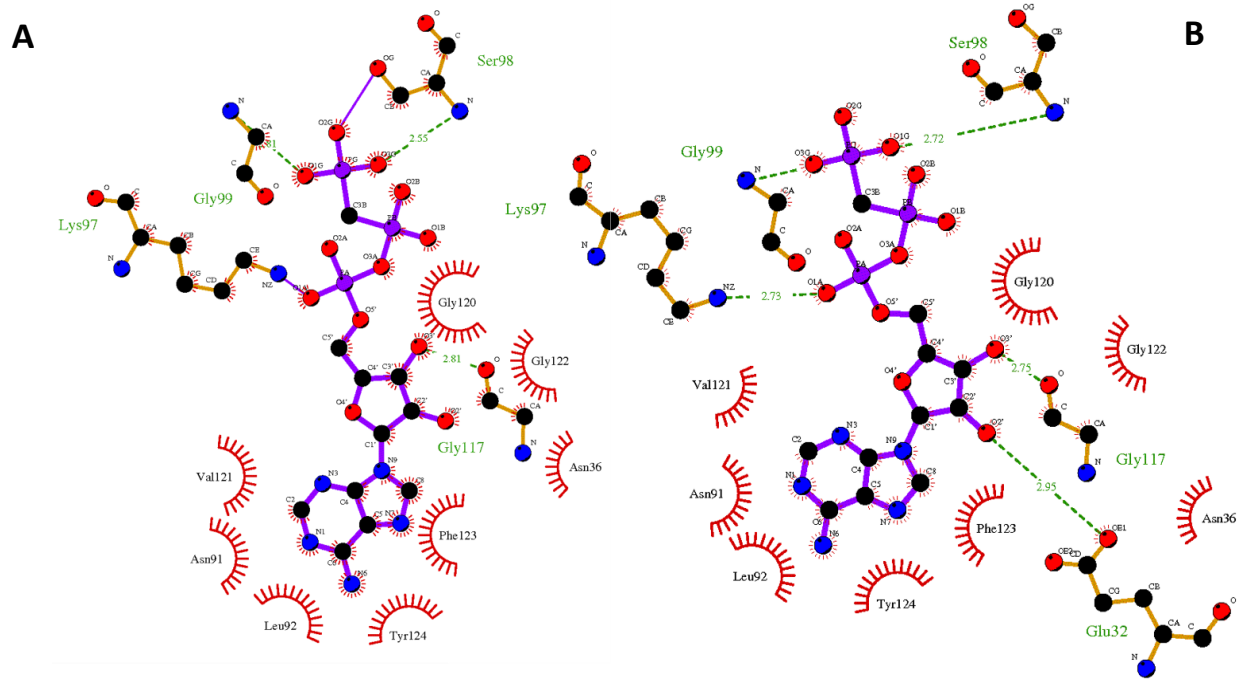


Figure 3-6: Comparison of LigPlot+ 2D interaction maps of the control docking (**A**) and homology model docking (**B**) with inhibitor ACP. The map clearly shows near identical poses and interacting residues bar Glu32 in the homology model (**B**)

3.7.3 Whole protein virtual screening

As mentioned in both previous chapters, Hsp90 cycles through several conformational states, each of which are important for a particular phase in its functional life cycle. Two of the most important phases being the fully extended or open phase and the closed ATP bound phase. During the open phase, the middle domain is completely exposed providing a suitable client protein binding site. In the closed phase, the client protein has already bound along with ATP at the N-terminal domain and it is in this phase that Hsp90 aids in guiding protein folding. Given the functional importance of these two particular conformational phases, both were investigated for potential novel target sites in this study.

Whole protein virtual screening was used in this investigation in an attempt to identify any potential target sites on the protein's surface. By centring the grid box on the protein's center of mass, and setting the dimensions to maximum, it was possible to encapsulate the whole protein within the simulation space (Figure 3-7). The simulation space in this instance is the specific docking environment defined by the boundaries of the grid box. It within this space that the test ligand will randomly search for a binding site on the receptors surface.

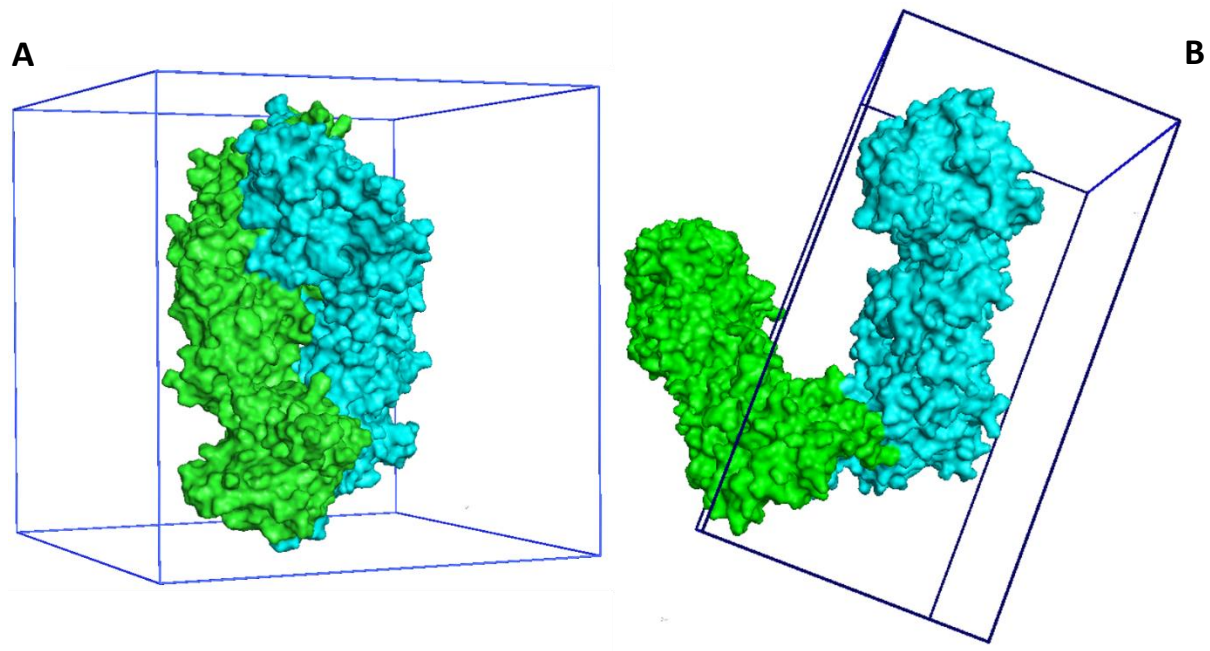


Figure 3-7: Schematic representation of the simulation space used in blind docking of Hsp90 in closed (**A**) and open (**B**) conformations as defined by their respective gridboxes

All 311 optimized database compounds were individually screened against both open and closed human Hsp90 homology models, using the same procedure described in the previous Section. The Lamarckian Genetic Algorithm was used for a conformation search being the most efficient method, and is typically effective for systems with up to 10 rotatable bonds in the ligand (Morris et al. 2009). Using this algorithm AutoDock4 calculated several conformational dockings for each ligand, scoring each according to an internal empirical free energy scoring function. Python scripting was used to filter the results of all 311 docking experiments according to conformational energy scores. The ligands with the lowest energy and largest cluster number were extracted from the results and written to PDB file format.

Thus for each docked compound, two energy scored conformations were retained, taking the total number of docked ligands per protein to 622. By viewing all 622 docked ligand conformations simultaneously, it was possible to obtain a visual binding consensus for the protein of interest. Binding regions on the proteins surface that were conserved over multiple docked ligand conformations could be visually identified by positional conservation and further investigated.

3.7.3.1 Closed conformation screening and target site identification

Full protein screening of Hsp90 in closed conformation and subsequent visual analysis using PyMOL, revealed several highly conserved binding pockets (Figure 3-8 A) excluding the already well-established ATP binding domain (Prodromou et al. 1997). Closer inspection of these binding pockets revealed that two were in very close proximity to residues that are thought to be essential for Hop-Hsp90 interactions, namely residues 239, 243, 250, 252, 254, 363, 364, 404, 405, 414 and 418 (Hatherley et al. 2015).

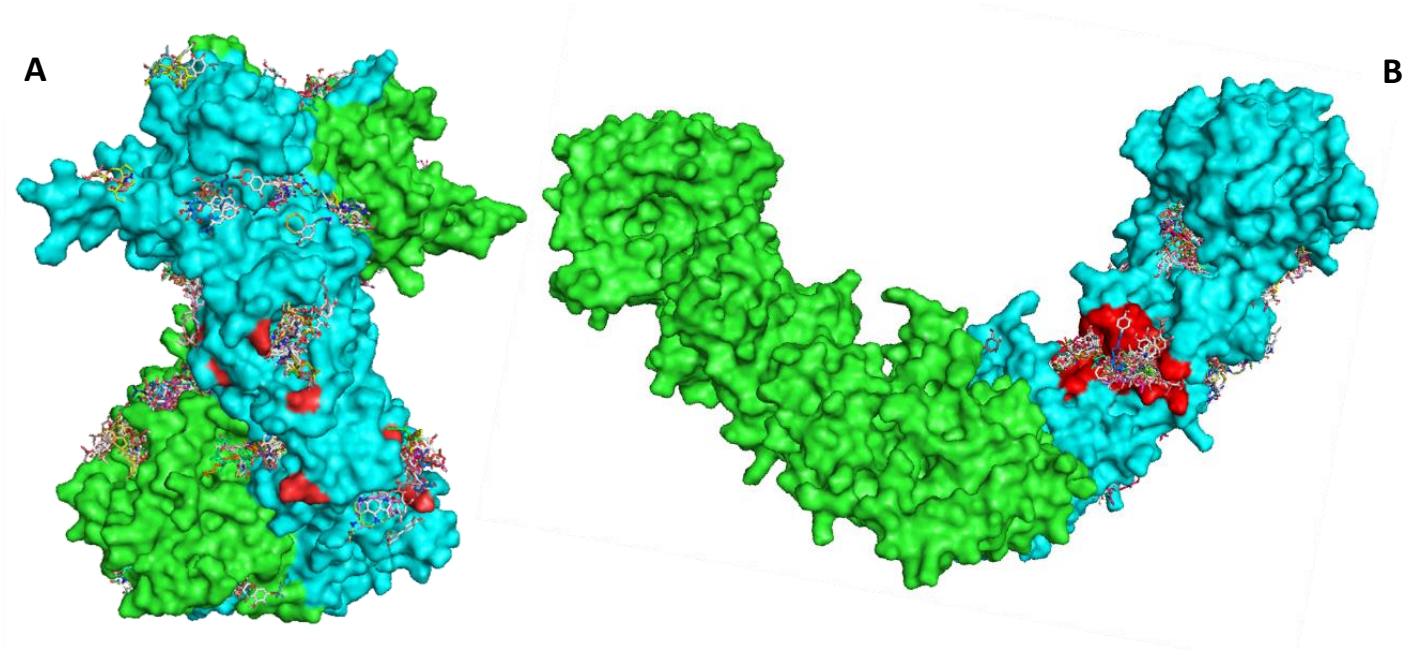


Figure 3-8: Whole protein screening of Closed (**A**) and open (**B**) Hsp90 homology models revealed several putative target sites in the form of conserved binding pockets. Target residues of interest shown in red.

3.7.3.2 Open conformation screening and target site identification

Hsp90's open v-like conformation is not nearly as compact as the closed conformation and as such has a surface area almost double that of its closed counterpart. Virtual screening of this exceptionally large conformation using blind docking techniques is challenging as the simulation space required is excessively large. Seeing as the monomers of Hsp90 are identical, it being a homodimer, this spatial limitation was easily circumvented by setting up a grid box that only fully encapsulated a single monomer. The dimensions were however set to include the C-terminal domain for both monomers as it is considered to be an important site for protein dimerization (Figure 3-7). The screening and subsequent filtering of all 311 compounds, within this simulation space, setting the open conformation human homology model as the receptor, revealed fewer but no less distinct binding pockets in comparison with the closed conformation screenings. Interestingly however, a sizable binding pocket was observed on the open and exposed middle domain (residues; 284-291, 374, 375, 377, 377-388, and 462-468), a region thought to be involved in protein client binding (Prodromou 2012).

3.7.4 Targeted docking studies

As previously discussed, whole protein screening studies revealed several potential binding pockets that could prove to be putative druggable targets. Due to time constraints, only three of targets could be further investigated using targeted docking studies. As the name suggests, these docking studies involve the specific targeting of ligands to particular regions on the receptor's surface. These 'target' regions are defined by the simulation space enclosed within the dimensions of the gridbox, which in most cases has a very limited volume, often only big enough to contain a few key residues. Cartesian coordinates are used to position the gridbox over the center of the putative target site, and the box dimension are accordingly set to encapsulate the whole site. Table 3-1 provides a detailed summary of the coordinates and box dimensions used for each respective target site investigated in this study.

Table 3-1: Gridbox parameters used to define the simulation space for the experimental control and each target site investigated

AutoDock4 run		Grid center			Number of points			Spacing (Å)
		X	Y	Z	X	Y	Z	
Control		-	-	-	90	90	90	0.375
Hop interaction sites	Target 1	74.66	71.33	106.18	60	73	60	0.375
	Target 2	94.66	100.33	95.18	65	65	65	0.375
CTD Target		90.66	73.33	103.18	73	65	65	0.375

3.7.4.1 Targeting Hop-Hsp90 interacting residues

Closed conformation screening of Hsp90 revealed two distinct binding pockets in close association with residues reported in a recent study by Hatherley et al. (2015), to be involved in Hop-Hsp90 interactions. The binding of Hop to Hsp90 is an important event in Hsp90's functional life cycle. Hop is an essential co-chaperone of Hsp90 and plays an important role in the delivery and transfer of client proteins to Hsp90 (Southworth & Agard 2011). Preventing this macromolecular interaction from occurring would potentially disrupt the natural order of events in Hsp90's functional cycle, and possibly even derail its biological function entirely. This experiment was designed to further investigate the two binding pockets in close association with the Hop interaction residues, hereon to be referred to as Target site 1 and 2 respectively (T1 and T2), using targeted docking studies. The outcomes of which would be twofold; 1) to define and characterise each binding pocket in a detailed target site assessment; and 2) to identify well bound ligands within these target sites, for the identification of potential lead compound candidates.

The protein model in Figure 3-9 shows a graphical representation of the gridboxes used to define T1 and T2. The exact coordinates and box dimensions for these targeted gridboxes are listed in Table 3-1. The docking parameters used for these docking runs were set to be as exhaustive as possible, and the population size set to 150. Computational demands however increase proportionately with parameter refinement, but because the simulation space required for these experiments needed not be very large, the high computational demands could be offset by reducing the number of runs to 100. Using these parameters, all 311 compounds in the compound library could be efficiently docked to both T1 and T2 with a good degree of accuracy.

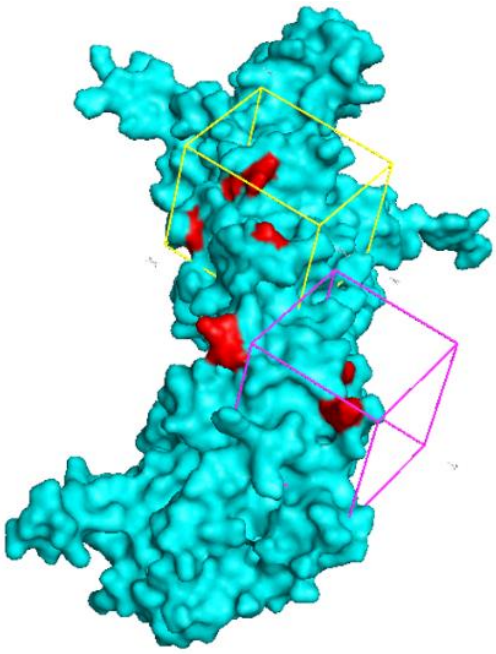


Figure 3-9: Schematic overview of target sites 1 and 2 as defined by the yellow and magenta gridboxes respectively

The docking results for each ligand were analysed separately in a series of filtering and scoring steps, which are outlined in Figure 3-10. Initial filtering of each ligand docking was completed based solely on the conformational empirical free energy scores calculated by AutoDock4. ADT summarize_docking.py script was used to summarize each docking log file, to generate a list of conformational energy scores. Custom Python scripts were designed to sort these energy score lists in decreasing order and extract the conformation with the best free energy score, creating a second list containing the best energy scored conformation for each compound. These energy

scores were subsequently sorted in decreasing order and the top scoring compounds for each target site identified.

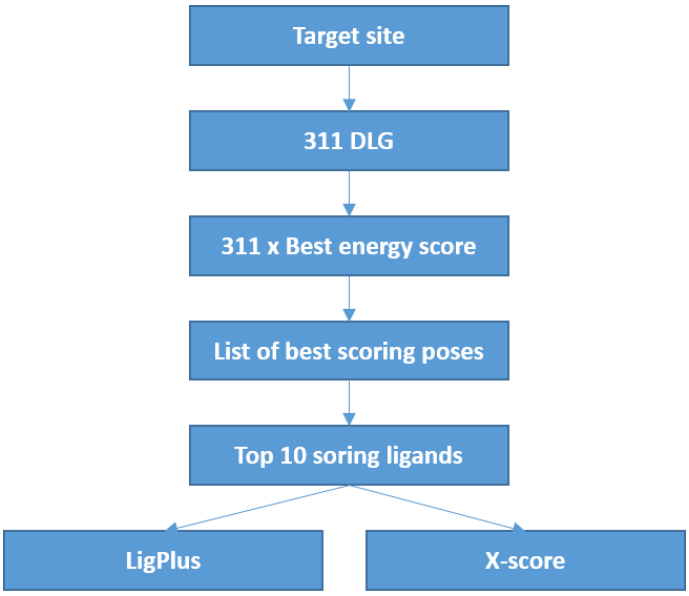


Figure 3-10: Flow diagram summarizing the filtering procedure used to extract the top 10 scoring ligands per target site from the initial 311 compounds

LigPlot+ was used to analyse the protein ligand interactions for each of the best scoring compound conformations in each of the two target sites. Given a protein ligand complex, LigPlot+ calculates and generates a 2D map showing the protein-ligand interactions for each residue involved in binding. These binding interaction maps were used to perform four quantitative analyses. Namely the calculation of; the total number of interacting residues per ligand, the total number of ligands per interacting residue, the number of hydrogen bonds per ligand, and the number of residues involved in hydrogen bonding. An example of a LigPlot+ 2D interaction map can be seen in Figure 3-13 (B).

3.7.4.1.1 Quantitative analysis of target site 1

After conformational filtering, a total of 7 compounds in T1 were selected for further quantitative analysis with LigPlot+ and are shown graphically in Figure 3-11 according to their respective empirical free energy scores as calculated by AutoDock4. The energy scores shown here

represent the highest 7 recorded for T1, the 8th being -8.2. Theoretically these ligands therefore represent the best bound compounds at T1.

Molecular recognition between a small molecule ligand and protein receptor, however is defined by the specific interactions between the ligand and the residues that make up the receptor active site. The greater the number of interacting residues with the ligand, the greater the total number of individual protein-ligand interactions, and therefore the greater the force of attraction between the two. The residue-ligand interaction maps for each of the seven T1 ligands gave an overview of which residues in the putative active site contributed to binding and an amalgamated list of all interacting residues over all seven ligands could be compiled to give a broad overview of all interacting residues.

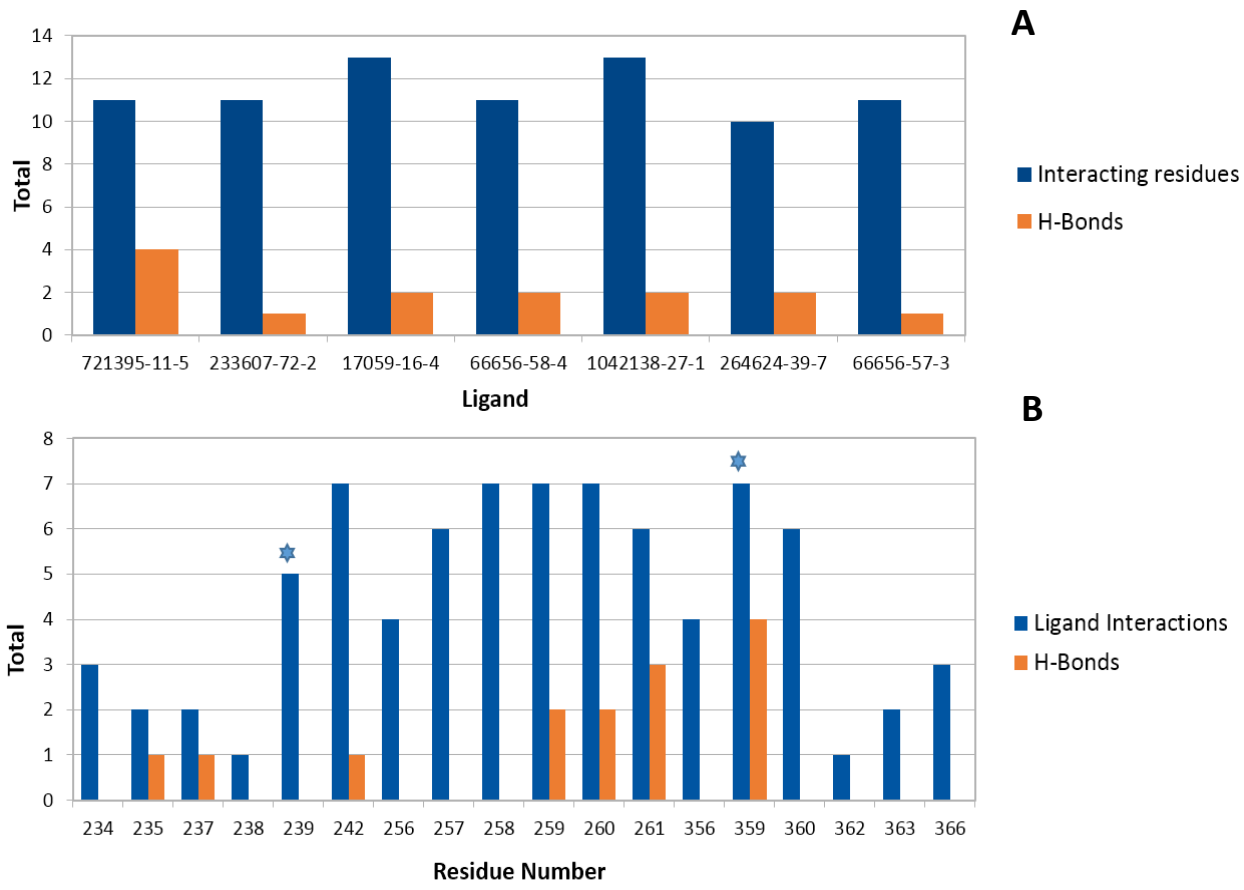


Figure 3-11: Graphical representation of quantitative data collected from LigPlot+ analysis of T1. Showing; (**A**) The total number of residue interactions and hydrogen bonds per ligand, (**B**) The number of ligand interactions and hydrogen bond contributions per residue. Residues directly involved in Hop-Hsp90 interactions are indicated with a ✡

Figure 3-11 presents all the collected ligand-residue data graphically, demonstrating how many residues contributed to binding for each ligand (Figure 3-11 A), as well as the number of times each residue was observed interacting with any of the seven ligands (Figure 3-11 B). From the data shown in Figure 3-11 (A), the average residue contribution per ligand could be calculated and was found to be 10 residues per ligand. Two ligands were observed to have a maximum 13 interacting residues. A minimum of 9 interacting residues was observed once, and together with the high average residue contribution per ligand, the data suggests that the binding pocket of T1 contains a rich source of residues that are capable of contributing to ligand interactions. The data presented in Figure 3-11 shows, that on average each residue contributed to interactions with 4 of the 7 ligands. Of the total 18 interacting residues present in T1, 5 residues (242, 257, 258, 259, 260 and 359) were conserved over all seven ligand interactions, suggesting their putative relative importance. Of the remaining 13 residues, only two were observed to be involved in ligand interactions with a single ligand. Interestingly, residues 239 and 363, which occurred in 70% and 30% of the ligand interactions respectively, are two of the residues previously reported to be directly involved in Hop-Hsp90 interactions (Hatherley et al. 2015).

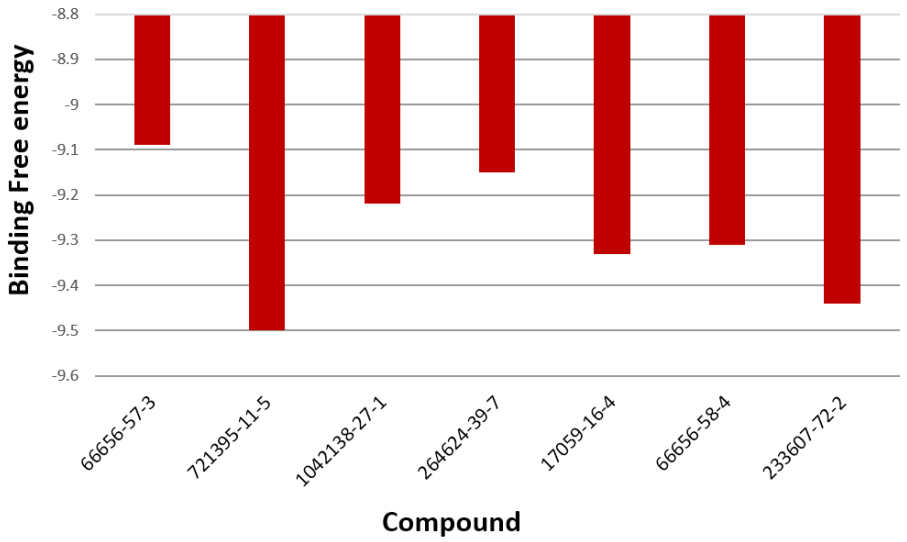


Figure 3-12: Graphical representation of empirical binding free energy scores for each ligand as calculated by AutoDock4

Looking at the empirical free energy scores for each ligand (Figure 3-12), it is clear that ligand 721395-11-5 obtained the best score with a binding free energy of -9.5. One of the most important contributors to binding affinity and a good energy score are Hydrogen bonds.

Hydrogen bonding is a key interaction type in molecular recognition, and as previously mentioned in Section 3.3.1, they are thought to be the most important directional interaction, and significantly contribute to both the stability and selectivity of a bound ligand. It is therefore not surprising to note that ligand 721395-11-5 also had the highest number of recorded hydrogen bonds (Figure 3-11 A). Of the 5 residues conserved over all 7 ligands, 4 contributed to one or more hydrogen bond interactions (Figure 3-11 B), providing further evidence that these residues may be key contributors to protein-ligand interactions in T1. Indeed the hydrogen bond data in Figure 3-11 (A) suggests that ligand interactions in T1 are fairly stable with an average of 2 hydrogen bond interactions observed for each ligand.

By accumulating the data presented here, it is possible to define the binding pocket of T1 with a greater level of detail. PyMOL representation of the binding pocket shown in Figure 3-13 (A), clearly shows which residues are essential to the target site. These residues were selected according to their relative importance over all tested ligands. Residues that contributed to interactions with 6 or more ligands automatically qualified, while those residues that contributed to 2 or more hydrogen bond interactions were also included. The 2D interaction map of ligand 721395-11-5 is shown in Figure 3-13 (B) and describes all the residues presented in Figure 3-13 (A).

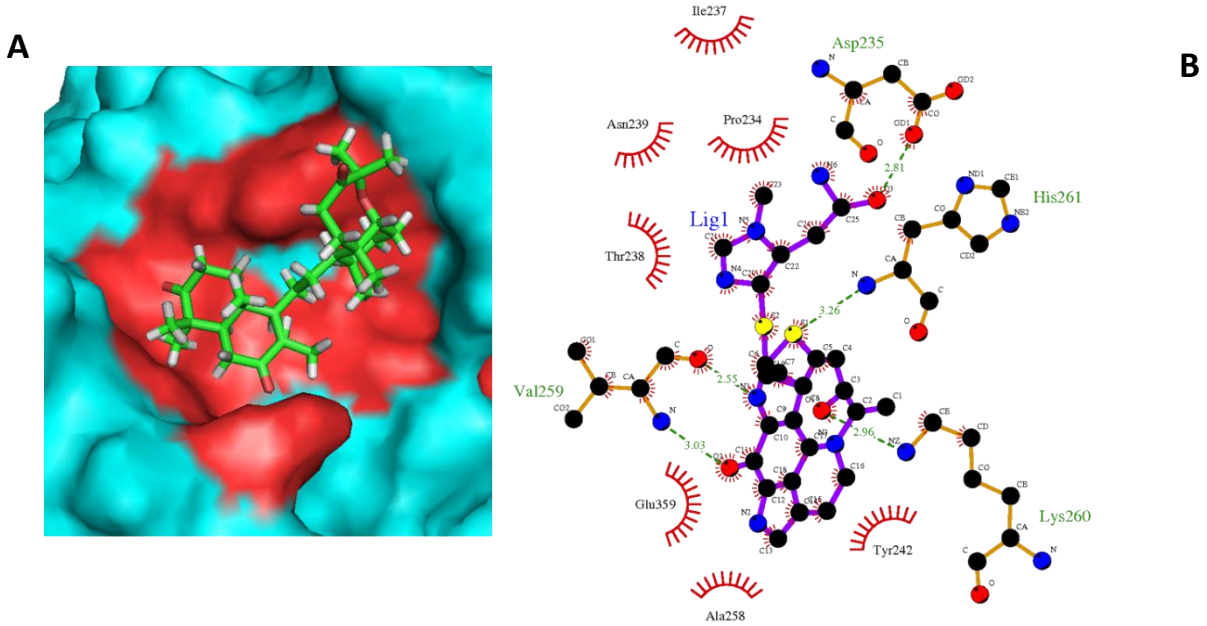


Figure 3-13: Schematic representations of target site one bound by ligand 721395-11-5; (**A**) Pymol visualization with binding pocket in red, (**B**) LigPlot+ 2D interaction map showing interacting residues and hydrogen bonds (dashed line)

3.7.4.1.2 Quantitative analysis of target site 2

A total of 9 ligands bound to target site 2 (T2) were analysed in the same manner as those in T1, and the following observations made. On average, ligands docked in T2 interacted with approximately 10 of the total 21 residues analysed (Figure 3-14 A). The average number of ligand interactions per residue was found to be 4, and a 100% ligand conservation was observed for residues 418 and 421, which were involved in interactions will all 9 ligands (Figure 3-14 B).

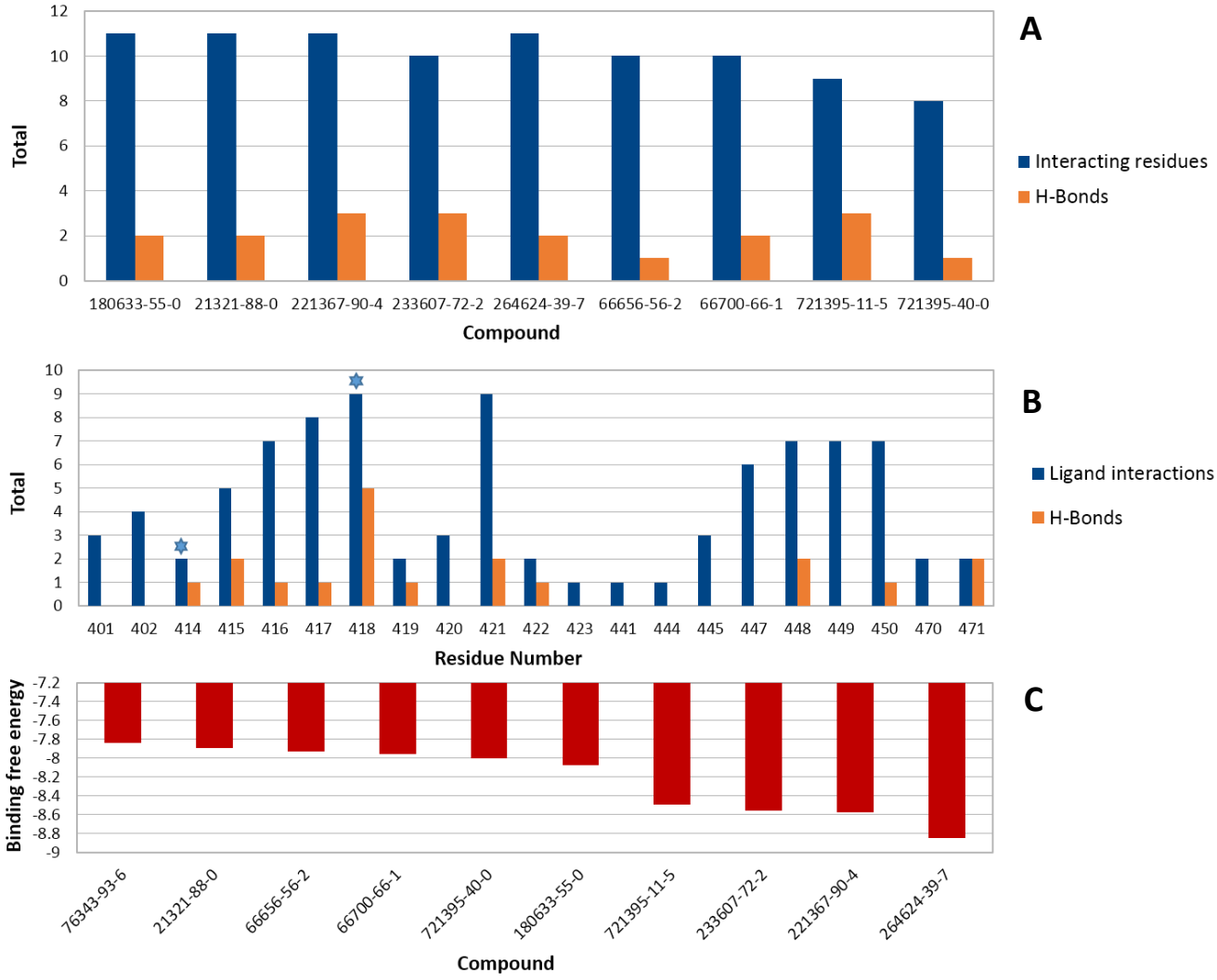


Figure 3-14: Graphical representation of quantitative data collected from LigPlot+ analysis for T2, Showing; (**A**) The total number of residue interactions and hydrogen bonds per ligand, (**B**) The number of ligand interactions and hydrogen bond contributions per residue, with residues directly involved in Hop-Hsp90 interactions are indicated with a ❄, (**C**) Binding free energy scores for ligands bound to target site 2, as calculated by AutoDock4

Interestingly of the 9 interactions residue 418 was involved in, a total of 5 were found to be hydrogen bonds, suggesting further evidence that this particular residue may be a key interacting residue in T2. Indeed, residues 414 and 418 form part of the list of Hop-Hsp90 interaction target residues.

The binding site of T2 was defined and refined by selecting residues that were observed to have had interactions with 6 or more of the 9 ligands analysed, as well as residues that contributed to hydrogen bond interactions with at least 2 different ligands. This binding pocket is shown in Figure 3-15 along with the LigPlot+ results for ligand 264624-39-7, which was recorded to have the lowest empirical free energy score of -8.8 Figure 3-14 (C). The PyMOL representation of this docking reveals that the binding site of T2 is quite clearly defined by a narrow binding groove, which is completely occupied by ligand 264624-39-7. As can be seen in Figure 3-14 (A), 264624-39-7 interacted with a maximum 11 of the 21 recorded residues in and around this binding pocket. Coupled with the 2 recoded hydrogen bonds and low energy score, 264624-39-7 may be a candidate for lead discovery as a potential inhibitor of T2.

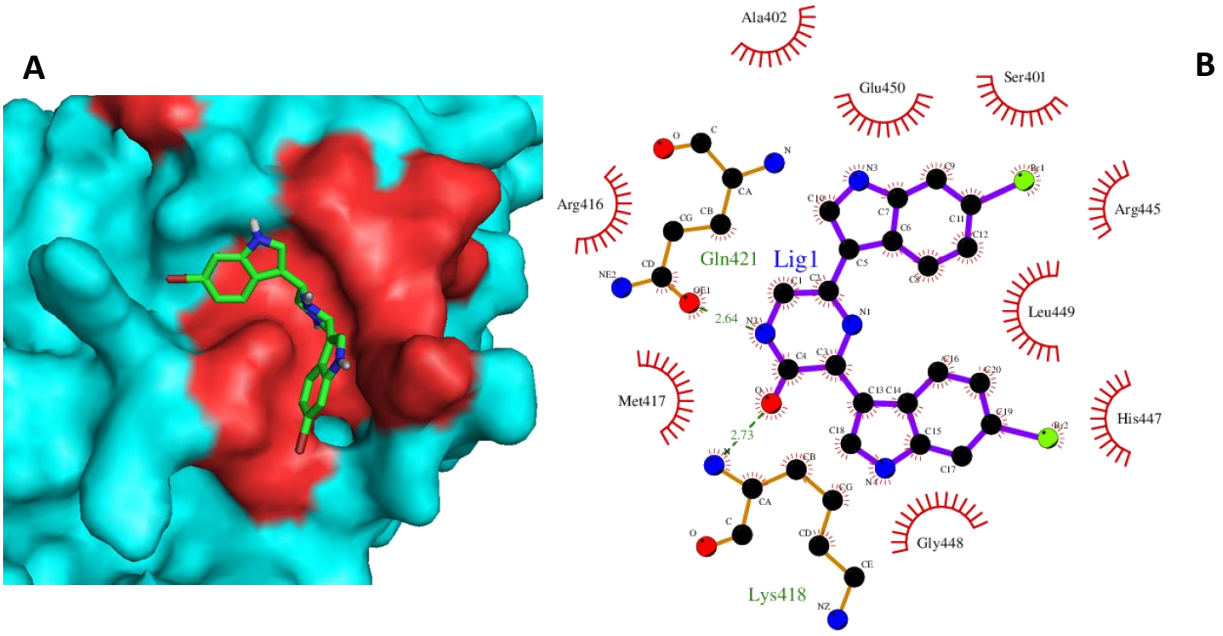


Figure 3-15: Binding groove of target site 2 with bound ligand 264624-39-7. (**A**) Pymol representation with binding pocket in red, (**B**) LigPlot+ 2D interaction map showing all residue interactions including two hydrogen bonds (dashed line)

3.7.4.2 Targeting the CTD dimerization site

The CTD of Hsp90 provides the primary site for Hsp90 monomer dimerization (Rastelli 2014). Before Hsp90 can become biologically active, two identical monomers dimerize, and are arranged in an open v-like conformation. Preventing Hsp90 from dimerizing would, to all intents and purposes, render the protein non-functional. A recent study by Rastelli (2014), reported several residues in the CTD to be crucial for Hsp90 dimerization, suggesting a putative drug target site. In this experiment, a CTD target site (T3), was defined using these target residues (610, 622, 626, 627, and 630), and screened as before, using all 311 of our optimized compounds in a third and final targeted docking experiment.

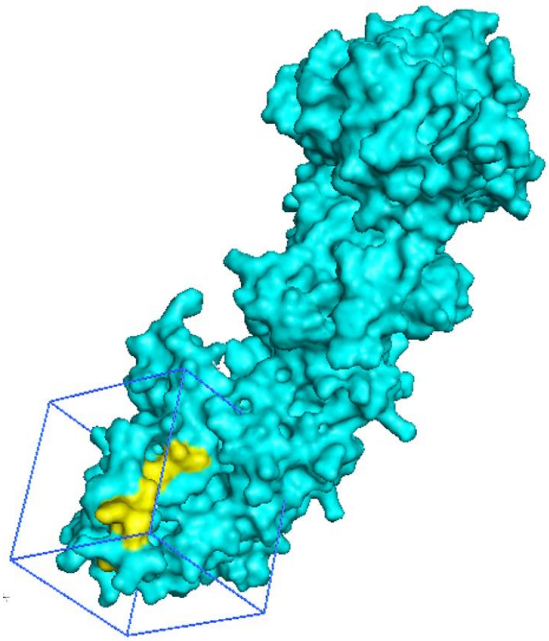


Figure 3-16: CTD dimerization target site (T3) defined by the gridbox (blue) and target residues (yellow) as depicted in PyMOL

The grid box presented schematically in Figure 3-16 was designed to encapsulate all the target residues (yellow), and the dimensions used are listed in Table 3-1. No changes to the previously used docking parameters were made in this experiment, and all 311 compounds were successfully screened against T3. The results generated were handled in an identical manner to the previous two experiments and the following observations made.

After the initial filtering, a total of 9 ligands were selected for further analysis each having an empirical energy score of less than -6.00 (Figure 3-17 C). LigPlot+ analysis revealed a total of 30 residues involved in interactions with these ligands. Based on the data presented in Figure 3-17 (A), the average numbers of 9 interacting residues and 2 hydrogen bonds per ligand could be calculated. The individual residue data shown in Figure 3-17 (B) shows that, excluding residues involved with a single ligand interaction, each residue had on average 4 different ligand interactions. Of the total 30 residues, 607 and 610 showed the highest conservation of ligand interactions, each interacting with 8 of the 9 ligands.

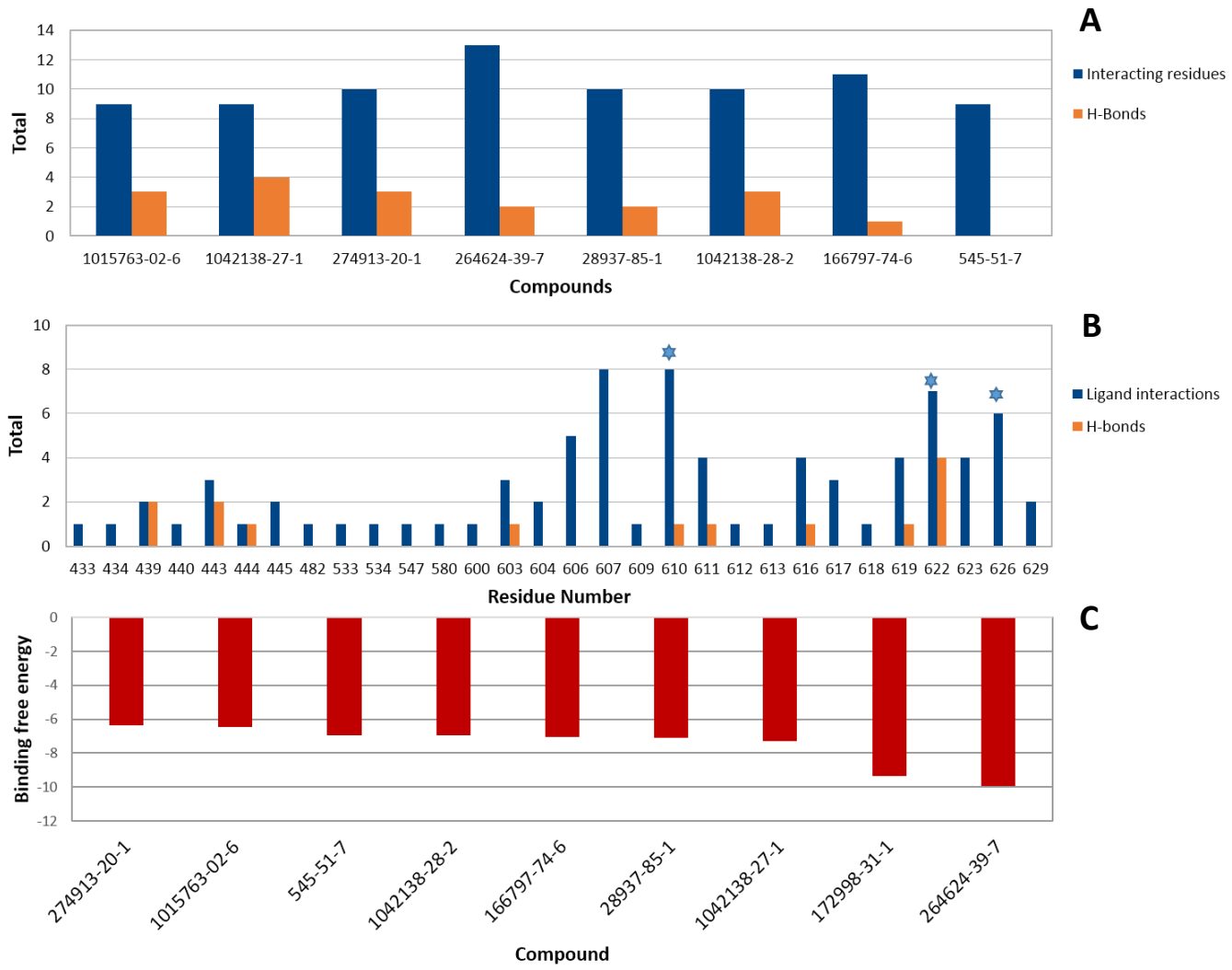


Figure 3-17: Graphical representation of quantitative data collected from LigPlot+ analysis for T3, Showing; (**A**) The total number of residue interactions and hydrogen bonds per ligand, (**B**) The number of ligand interactions and hydrogen bond contributions per residue, with residues directly involved in Hop-Hsp90 interactions are indicated with a ✵, (**C**) Binding free energy scores for ligands bound at the target site, as calculated by AutoDock4

Based on this data, a detailed binding pocket was defined by residues that either interacted with a minimum of 5 of the 9 ligands, or contributed to hydrogen bonding with 2 or more ligands (shown coloured in yellow in Figure 3-18 A, occupied by the docked ligand 264624-39-7). Interestingly, ligand 264624-39-7, which was identified as being the best bound ligand in T2, scored the lowest empirical free energy score in T3 as well, with a score of -10.0. A score likely attributed to by the maximum 13 interacting residues, the highest number seen across all 9 ligands. The LigPlot+ data shown in Figure 3-18 (B) shows that 264624-39-7 also had additional stabilizing support of 2 hydrogen bonds.

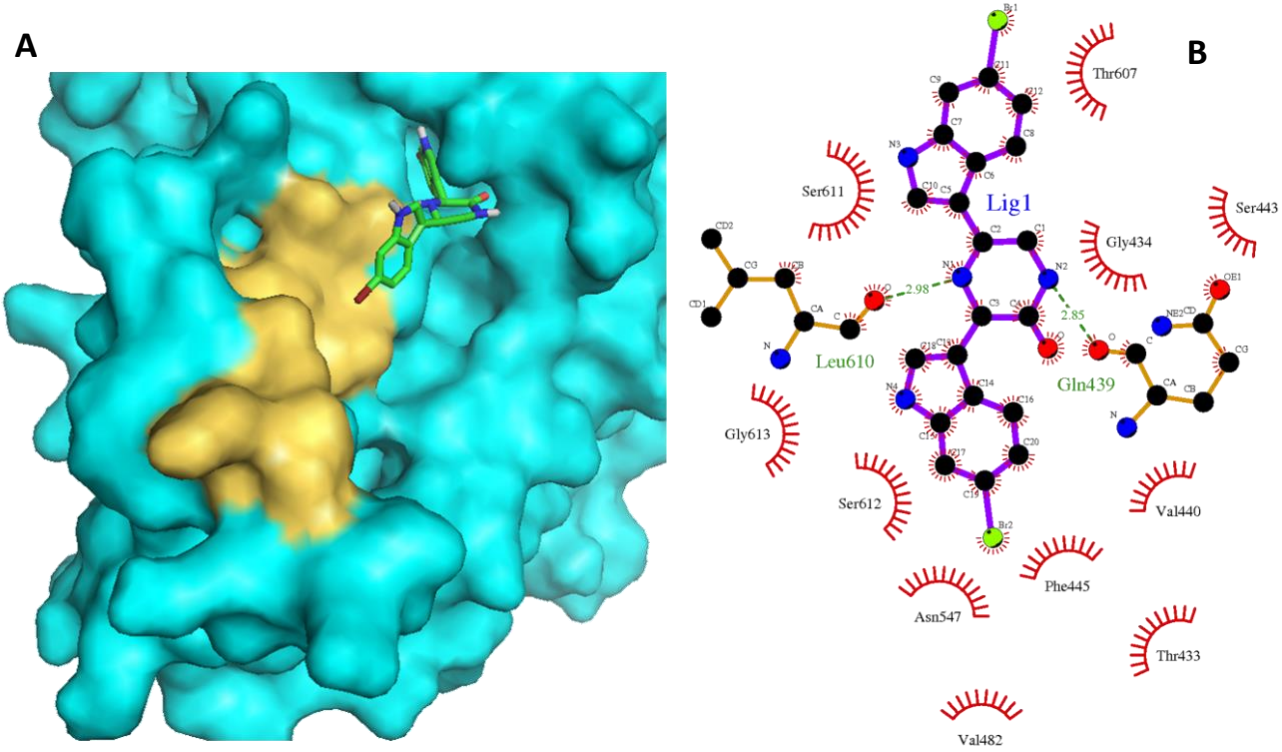


Figure 3-18: The CTD dimerization site bound by ligand 264624-39-7. (**A**) Pymol representation of binding pocket with target residues in yellow, (**B**) LigPlot+ 2D interaction map of all interacting residues with ligand 264624-39-7, including two hydrogen bonds (dashed line).

3.7.5 Evaluation of putative lead compounds

Each of the afore mentioned targeted docking studies, produced a best bound ligand based on the number and type of ligand interactions, and calculated free energy score. These ligands were further investigated for future use as putative lead compounds according to Lipinski's rule of 5 (Lipinski et al. 1997), which were used to evaluate each compound's druglikeness. Lipinski's rules, also known as the Pfizer's rule of five are used to assess whether a chemical compound has

suitable properties for use as a potential orally administered drug. The rules essentially describe the molecular properties that are important for a drug's pharmacokinetics within the human body (Lipinski 2004), such as absorption rate, distribution, metabolism, and excretion.

X-Score is a scoring function that computes the binding affinities of a given ligand with a target protein, and was used in this study to analyse the chemical makeup ligands 721395-11-5 and 264624-39-7. Additionally, X-Score can calculate the molecular formula, molecular weight and Log *P* values for a docked ligand, and were used along with 2D imaging of each compound, to assess each ligand in terms of Lipinski's rule of five. This data is shown in Table 3-2.

Table 3-2: Summary of analysis of best scoring ligands according to Lipinski's rule of 5

Ligand	Target site	MW	Log P	H-bond donors	H-bond acceptors
721395-11-5	T1	535.62	1.56	5	10
264624-39-7	T2 and T3	488.17	5.41	4	5

Lipinski's rules define four simple physicochemical parameter ranges, MWT ≤ 500, Log *P* ≤ 5, H-bond donors ≤ 5, and H-bond acceptor ≤ 10. Looking at the data in Table 3-2, both ligands appear to have some potential as lead drug candidates. Ligand 721395-11-5 is slightly over the maximum 500 KDa with a weight of 535.62. The log *P* value for this ligand however is excellent at a mere 1.56. The H-bond donors and acceptors are at the maximum values of 5 and 10 respectively. Ligand 264624-39-7 is a very interesting putative candidate in that it was recoded as being the best scoring ligand it two separate target sites (T2 and T3). Its MW is well below the upper limit at 488.17 KDa. Its Log *P* however was calculated at 5.41, just over the maximum 5. The number of H-bond donors and acceptors were well within range at 4 and 5 respectively. These data suggest that that both ligands have potential as lead compounds both only failing one of the Lipinski rules.

3.8 Chapter conclusions

At the outset of this chapter, several objectives relevant to the aim of this project were laid out. The following Section allows for a revisit of these objectives relative to findings described in the previous Section.

A comprehensive search of publically available literature revealed that there is a vast quantity of naturally derived compounds indigenous to South Africa, which have been defined and classified. It was thus possible to construct a mini-database of 350 such compounds in preparation for a compound screening library. Using freely available software in the public domain, it was possible to optimize 311 of these compounds in preparation for use in *in silico* docking studies.

Whole protein screening of Hsp90 homology models revealed that it is possible to use a blind docking screening approach to identify putative binding sites on a proteins surface, and in conjunction with literature it was possible to identify 3 of these binding pockets as putative drug target sites.

Targeting these putative target sites using the conventional targeted docking approach, where the simulation space was restricted to the region of interest, it was possible to generate data, such that quantitative analysis of each target site was possible. This analysis revealed in depth detail about each target site as well as the identification of two putative lead compound candidates.

CHAPTER 4: CONCLUDING REMARKS

The ultimate goal of protein modelling is to predict the structure of a protein based on its sequence with an accuracy that is comparable to experimentally determined data (Krieger et al. 2003). Likewise the aim of molecular docking is to accurately predict the structure of a ligand bound to a receptor binding site and to correctly estimate the strength of binding (Waszkowycz et al. 2011). The experimental work described in this thesis has addressed both these techniques in parallel, in an attempt provide a structural platform for human Hsp90 research and tentatively pave the way for the rational drug design of next generation Hsp90 inhibitors.

4.1 Homology modelling of human Hsp90

The homology modelling techniques used in this study, we believe are a novel approach to modelling full length human Hsp90. The amalgamation of multiple structural templates for the homology modelling of human Hsp90 has never been attempted before. Despite the success in modelling a monomer in open conformation and a homodimer in closed conformation, the work described here is still limited by a lack of experimental structural data. The 51 residue insert in the human Hsp90 amino acid sequence is still unaccounted for and could not be modelled in any way in this study. The remaining 681 residues however have been well accounted for and their structure as accurately modelled as possible.

The successful arrangement of homology modelled Hsp90 monomers in the open v-like state, demonstrated the potential of combining homology modelling techniques with protein-protein docking techniques. Providing evidence that if done carefully and accurately and basing the docking on prior binding site knowledge, protein subunits can be fitted together relatively accurately.

The outcome of this Section was the accurate protein modelling of human Hsp90 in its nucleotide bound closed states and unbound open v-like state. Both models providing an excellent platform and starting point for in depth structural Hsp90 analyses.

4.2 Structure-based targeted drug discovery

The molecular docking studies conducted in this study were used to great success in a number of different investigations. The implementation of blind docking techniques in an attempt to locate potential binding pockets provided further model validation after successfully completing a controlled docking experiment. Whole protein screening revealed several sites on the surface of both open and closed homology models that demonstrated high binding affinities for ligands, the conservation of which was used as a basis for identifying putative binding pockets. Further analysis of these binding sites, some of which were found to be in close proximity to key co-chaperone interaction residues, revealed a greater level of detail with respect to residue-ligand interactions. The data of which could be used to characterize and define the binding pocket. Targeted docking studies at three separate target sites, revealed that natural compounds still much to offer in novel drug identification, with two compounds from the library of 311 natural compound showing excellent binding affinities at two of the investigated target sites.

The work discussed in this chapter demonstrates the potential of structure-based drug discovery techniques, showing that homology models have much to offer in terms of a structural base for these studies and that if used correctly, large scale screening of natural compounds and their derivatives provides a fast and relatively cheap alternative for drug discovery. Indeed three distinct target sites were identified in this study. The two Hop-interaction target sites T1 and T2 fit well with the observations made in Section 1.10, rather than interfering with the Hsp90s ATPase cycle, as seen by most therapeutic Hsp90 inhibitors thus far, these target sites provide a means by which potential inhibitors can interfere with essential co-chaperone protein interactions, rather than completely disabling the Hsp90 chaperone. T3 on the other hand provides and interesting binding pocket at the CTD dimerization interface only present on the monomeric form of the chaperone. Successful targeting of this site will no doubt render the protein non-functional but still provides great potential for the development of next generation CTD inhibitors.

Analysis of ligands 721395-11-5 and 264624-39-7 according to Lipinski's rule of five, revealed that these two ligands may be suitable candidates as lead compounds in a detailed drug discovery study.

4.3 Future work and prospects

The homology modelling research described in this thesis provides a suitable platform and starting point for a plethora of different *in silico* structural studies. While the homology models introduced here provide further structural insight into human Hsp90, the perspective of the protein models is restricted to a single point in time. Whole protein *in silico* molecular dynamics studies on these homology models would provide a means of further investigating the dynamics of this highly flexible and mobile protein, capturing structural data for each conformational stage. Notable experimentation should also include the interaction between Hsp90 and co-chaperone Hop at the interaction interface described by the "hot spot" residues mentioned in this research, and reported by Hatherley et al. (2015). A study of this nature would provide valuable insights regarding the involvement of these residues, and those that make up the rest of the investigated binding pockets (T1 and T2), in Hsp90-Hop interactions and their relative suitability as druggable target sites. Additionally the binding mode and strength of compounds 721395-11-5 and 264624-39-7 at target sites T1 – T3 could be further using molecular dynamics simulations, allowing for better scoring methodologies such as free energy calculations and MM-PBSA and MM-GBSA scoring functions, and a more in depth description of the ligand binding over time.

Perturbation-Response Scanning (PRS), is a toolkit developed by Atilgan & Atilgan (2009) based on Linear Response Theory (LRT) for the study of the origins of structural changed undergone by protein molecules. PRS involves the systematic application of forces at singly selected residues and recording the relative linear response for the whole protein. This technique could be used to provide further insight and understanding as which residues in Hsp90 are crucial for maintaining conformational changes. Indeed, the technique could also be applied to each of the three target sites reported in this study (T1, T2, and T3), to further elucidate and characterise the respective binding pockets.

The future work described here sheds light on the prospects a working 3D model of human Hsp90 has to offer. The aforementioned studies are the mere tip of the iceberg in relation to the plethora of possible *in silico* experimental work that could follow.

Indeed, the next potential step for selected target sites and hit compounds, such as the putative targets sites and hit compounds discussed here, should at some point be further investigated *in vitro* and *in vivo*. In experiments that would further facilitate and validate the hypothetical *in silico* analysis. Such experiments would include protein activity and toxicity assays, and mutagenesis and knock down analyses.

REFERENCES

Ali, M.M.U., Roe, S.M., Vaughan, C.K., Meyer, P., Panaretou, B., Piper, P.W., Prodromou, C. & Pearl, L.H., 2006. Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature*, 440(7087), pp.1013–7.

American Cancer Society, 2015. Cancer Facts & Figures 2015.

Anon, 2014. Growth of GenBank and WGS. *National Center for Biotechnology Information*. Available at: http://www.ncbi.nlm.nih.gov/genbank/statistics [Accessed October 14, 2014].

Anon, 2015. RCSB Protein Data Bank - RCSB PDB. Available at: http://www.rcsb.org/pdb/home/home.do [Accessed January 20, 2015].

Arnold-Schild, D., Hanau, D., Spehner, D., Schmid, C., Rammensee, H., de la Salle, H. & Schild, H., 1999. Cutting edge: receptor-mediated endocytosis of heat shock proteins by professional antigen-presenting cells. *Journal of Immunology*, (162), pp.3757–3760.

Atilgan, C. & Atilgan, A.R., 2009. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS computational biology*, 5(10), p.e1000544.

Bardwell, J. & Craig, E., 1988. Ancient heat shock gene is dispensable. *Journal of bacteriology*, (170), pp.2977–2983.

Barril, X., Brough, P., Drysdale, M., Hubbard, R.E., Massey, A., Surgenor, A. & Wright, L., 2005. Structure-based discovery of a new class of Hsp90 inhibitors. *Bioorganic & medicinal chemistry letters*, 15(23), pp.5187–91.

Berendsen, H.J.C., van der Spoel, D. & van Drunen, R., 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3), pp.43–56.

Binart, N., Chambraud, B., Dumas, B., Rowlands, D., Bigogne, C., Levin, M., Garnier, J., Baulieau, E. & Catelli, M., 1989. The cDNA-derived amino acid sequence of chicken heat shock protein Mr 90,000 (hsp 90) reveals a “DNA like” structure: potential site of interaction with steroid receptors. *Biochimica et Biophysica*, (159), pp.140–147.

Bissantz, C., Kuhn, B. & Stahl, M., 2010. A Medicinal Chemist’s Guide to Molecular Interactions. *Journal of Medicinal Chemistry*, 53(14), pp.5061–5084.

Boniecki, M., Rotkiewicz, P., Skolnick, J. & Kolinski, A., 2004. Protein fragment reconstruction using various modeling techniques. , pp.725–738.

Brown, M., Zhu, L., Schmidt, C. & Tucker, P., 2007. HSP90 – from signal transduction to cell transformation. *Biochemical and Biophysical Research Communications*, (363), pp.241–246.

Buchner, J., 1996. Supervising the fold: functional principles of molecular chaperones. *Federation of American Societies for Experimental Biology*, (10), pp.10–19.

Chen, S., Sullivan, W., Toft, D. & Smith, D., 1998. Differential interactions of p23 and the TPR-containing proteins Hop, Cyp40, FKBP52 and FKBP51 with hsp90 mutants. *Cell Stress Chaperones*.

Chowdhry, B. & Harding, S., Protein-ligand interactions and their analysis.

Clayton, Turkes, A., Navabi, H., Mason, H. & Tabi, Z., 2005. Induction of heat shock proteins in B-cell exosomes. *Journal of Cell Science*, (118), pp.3631–3638.

Comeau, S.R., Gatchell, D.W., Vajda, S. & Camacho, C.J., 2004. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic acids research*, 32(Web Server issue), pp.W96–9.

Comeau, S.R., Gatchell, D.W., Vajda, S. & Camacho, C.J., 2003. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, 20(1), pp.45–50.

Csermely, P., Schnaider, T., Csaba, S., Nardai, G., Amily, A. & Prohászka, Z., 1998. The 90-kDa Molecular Chaperone Family : Structure , Function , and Clinical Applications. , 79(2), pp.129–168.

Dollins, D.E., Warren, J.J., Immormino, R.M. & Gewirth, D.T., 2007. Structures of GRP94-nucleotide complexes reveal mechanistic differences between the hsp90 chaperones. *Molecular cell*, 28(1), pp.41–56.

Dominguez, C., Boelens, R. & Bonvin, A., 2003. HADDOCK: a protein– protein docking approach based on biochemical or biophysical information. *Journal American Chemistry Scociety*, 125, pp.1731–1737.

Dunn, M., 2010. Protein-Ligand Interactions: General Description. *Encyclopedia of Life Sciences*.

Dutta, R. & Inouye, M., 2000. GHKL, an emergent ATPase/kinase superfamily. *Trends in biochemical sciences*, 25(1), pp.24–8.

Eiben, C.B., Siegel, J.B., Bale, J.B., Cooper, S., Khatib, F., Shen, B.W., Players, F., Stoddard, B.L., Popovic, Z. & Baker, D., 2012. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature Biotechnology*, 30(2), pp.190–192.

Eldridge, M., Murray, C., Auton, T., Paolini, G. & Mee, R., 1997. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design*, 11(5), pp.425–445.

Eramian, D., Eswar, N. & Shen, M., 2008. How well can the accuracy of comparative protein structure models be predicted ? , pp.1881–1893.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M., Pieper, U. & Sali, A., 2001. Comparative Protein Structure Modeling Using MODELLER. In *Current Protocols in Protein Science*. John Wiley & Sons, Inc.

Feig, M. & Brooks, C., 2004. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Current opinion in structural biology*, 14(2), pp.217–224.

Filippov, I. V & Nicklaus, M.C., 2009. Optical Structure Recognition Software To Recover Chemical Information : OSRA , An. , pp.740–743.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J. & Punta, M., 2014. Pfam: The protein families database. *Nucleic Acids Research*, 42.

Freeman, B. & Morimoto, R., 1996. The human cytosolic molecular chaperones hsp90, hsp70 (hsc70) and hdj-1 have distinct roles in recognition of a non-native protein and protein refolding. *European Molecular Biology Organization*, (15), pp.2969–2979.

Freire, E., 2008. Do enthalpy and entropy distinguish first in class from best in class? *Drug Discovery Today*, 13, pp.869–874.

Gabb, H., Jackson, R. & Sternberg, M., 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of molecular biology*, 272, pp.106–120.

Garnier, C., Lafitte, D., Tsvetkov, P.O., Barbier, P., Leclerc-Devin, J., Millot, J.-M., Briand, C., Makarov, A.A., Catelli, M.G. & Peyrot, V., 2002. Binding of ATP to heat shock protein 90: evidence for an ATP-binding site in the C-terminal domain. *The Journal of biological chemistry*, 277(14), pp.12208–14.

Garrett, J. & Arteaga, C., 2011. Resistance to HER2-directed antibodies and tyrosine ki- nase inhibitors: mechanisms and clinical implications. *Cancer Biology Therapy*, (11), pp.793–800.

Goetz, M.P., Toft, D.O., Ames, M.M. & Erlichman, C., 2003. The Hsp90 chaperone complex as a novel target for cancer therapy. *Annals of oncology : official journal of the European Society for Medical Oncology*, 14(8), pp.1169–76.

Gohlke, H., Hendlich, M. & Klebe, G., 2000. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of molecular biology*, 259(2), pp.337–356.

Gohlke, H. & Klebe, G., 2002. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie-International Edition*, 41(15), pp.2645–2676.

Gray, J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. & Baker, D., 2003. Protein–protein docking with simultaneous opti- mization of rigid-body displacement and side-chain conformations. *Journal of molecular biology*, 331, pp.281–299.

Gupta, R., 1995. Phylogenetic analysis of the 90 kDa heat shock family of protein sequences and an examination of the relationship amoung animals, plants and fungi species. *Molecular Biology and Evolution*, (12), pp.1063–1073.

Hammerman, P., Janne, P. & Johnson, B., 2009. Resistance to epidermal growth factor receptor tyrosine kinase inhibitors in non-small cell lung cancer. *Clinical Cancer Research*, (15), p.75.

Harris, S., Shiau, A. & Agard, D., 2004. The crystal structure of the carboxy- terminal dimerization domain of htpG, the Escherichia coli Hsp90, reveals a potential substrate binding site. *Structure*, (12), pp.1087–1097.

Hartl, F., 1996. Molecular chaperones in cellular protein folding. *Nature*, (381), pp.571–580.

Hatherley, R., Clitheroe, C.-L., Faya, N. & Tastan Bishop, Ö., 2015. Plasmodium falciparum Hop: Detailed analysis on complex formation with Hsp70 and Hsp90. *Biochemical and biophysical research communications*, 456(1), pp.440–5.

Hetényi, C. & van der Spoel, D., 2002. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein science*, 11(7), pp.1729–37.

Hieronymus, H., Lamb, J., Ross, K.N., Peng, X.P., Clement, C., Rodina, A., Nieto, M., Du, J., Stegmaier, K., Raj, S.M., Maloney, K.N., Clardy, J., Hahn, W.C., Chiosis, G. & Golub, T.R., 2006. Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer cell*, 10(4), pp.321–30.

Hubbard, R. & Haider, M., 2010. *Hydrogen Bonds in Proteins: Role and Strength*, John Wiley & Sons, Inc.

Hubbard, T.J.P., Ailey, B., Brenner, S.E., Murzin, A.G. & Chothia, C., 1999. SCOP: A structural classification of proteins database. *Nucleic Acids Research*, 27, pp.254–256.

Huth, J.R., Park, C., Petros, A.M., Kunzer, A.R., Wendt, M.D., Wang, X., Lynch, C.L., Mack, J.C., Swift, K.M., Judge, R.A., Chen, J., Richardson, P.L., Jin, S., Tahir, S.K., Matayoshi, E.D., Dorwin, S.A., Ladror, U.S., Severin, J.M., Walter, K.A., Bartley, D.M., Fesik, S.W., Elmore, S.W. &

Hajduk, P.J., 2007. Discovery and design of novel HSP90 inhibitors using multiple fragment-based design strategies. *Chemical biology & drug design*, 70(1), pp.1–12.

Iannotti, A., Rabideau, D. & Dougherty, J., 1988. Characterization of purified avian 90, 000-Da heat shock protein. *Archives of Bochemistry and Biophysiology*, (264), pp.54–60.

Illergård, K., Ardell, D.H. & Elofsson, A., 2009. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins*, 77(3), pp.499–508.

Jackson, S.E., 2013. Hsp90: structure and function. *Topics in current chemistry*, 328, pp.155–240.

Jakob, U., Lilie, H., Meyer, I. & Buchner, J., 1995. Transient interaction of Hsp90 with early unfolding intermediates of citrate synthase. Implications for heat shock in vivo. *Journal of Biological Chemistry*, (270), pp.7288–7294.

Jayaram, B., Dhingra, P., Mishra, A., Kaushik, R., Mukherjee, G., Singh, A. & Shekhar, S., 2014. Bhageerath-H: A homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. *BMC Bioinformatics*, 15(Suppl 16), p.S7.

Jego, G., Hazoumé, A., Seigneuric, R. & Garrido, C., 2013. Targeting heat shock proteins in cancer. *Cancer letters*, 332(2), pp.275–85.

Jhaveri, K., Ochiana, S.O., Dunphy, M.P., Gerecitano, J.F., Corben, A.D., Peter, R.I., Janjigian, Y.Y., Gomes-DaGama, E.M., Koren, J., Modi, S. & Chiosis, G., 2014. Heat shock protein 90 inhibitors in the treatment of cancer: current status and future directions. *Expert opinion on investigational drugs*, 23(5), pp.611–28.

Jhaveri, K., Taldone, T., Modi, S. & Chiosis, G., 2012. Advances in the clinical development of heat shock protein 90 (Hsp90) inhibitors in cancers. *Biochimica et biophysica acta*, 1823(3), pp.742–55.

Kamal, A., Thao, L., Sensintaffar, J., Zhang, L., Boehm, M., Fritz, L. & Burrows, F., 2003. A high-affinity conformation of Hsp90 confers tumour selectivity on Hsp90 inhibi- tors. *Nature*, (425), pp.407–410.

Kitchen, D.B., Decornez, H., Furr, J.R. & Bajorath, J., 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews. Drug discovery*, 3(11), pp.935–49.

Koga, F., Kihara, K. & Neckers, L.E.N., 2009. Inhibition of Cancer Invasion and Metastasis by Targeting the Molecular Chaperone Heat-shock Protein 90. , 808, pp.797–807.

Koonin, E. V, 2002. The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes. *Antenna*, pp.1–6.

Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S.E., Xia, B., Hall, D.R. & Vajda, S., 2013. How good is automated protein docking? *Proteins*, 81(12), pp.2159–66.

Kramer, B., Rarey, M. & Lengauer, T., 1999. Evaluation of the flexx incremental construction algorithm for protein-ligand docking. *Proteins*, 37(2), pp.228–241.

Krieger, E., Nabuurs, S. & Vriend, G., 2003. Homology modeling. In P. E. Bourne & H. Weissig, eds. *Structural Bioinformatics*. pp. 507–521.

Krishnamoorthy, B. & Tropsha, a., 2003. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 19(12), pp.1540–1548.

Kumar, S. & Weaver, V., 2009. Mechanics, malignancy, and metastasis: the force journey of a tumour cell. *Cancer Metastasis Reviews*, (28), pp.113–127.

Langton, J.M., Blanch, B., Drew, A.K., Haas, M., Ingham, J.M. & Pearson, S.-A., 2014. Retrospective studies of end-of-life resource utilization and costs in cancer care using health administrative data: A systematic review. *Palliative medicine*.

Laskey, R.A., Honda, B.M., Mills, A.D. & Finch, J.T., 1978. Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. *Nature*, 275(5679), pp.416–420.

Lees-Miller, S. & Anderson, C., 1989. Two human 90-kDa heat shock proteins are phosphoryl-ated in vivo at conserved serines that are phosphorylated in vitro by casein kinase II. *Journal of Biological Chemistry*, (264), pp.2431–2437.

Li, J., Sun, L., Xu, C., Yu, F., Zhou, H., Zhao, Y., Zhang, J., Cai, J. & Mao, C., 2012. Structure insights into mechanisms of ATP hydrolysis and the activation of human. *Biochimica et biophysica acta*, (February), pp.300–306.

Li, W., Li, Y., Guan, S., Fan, J., Cheng, C., Bright, A., Chinn, C., Chen, M. & Woodley, D., 2007. Extracellular heat shock protein-90alpha: linking hypoxia to skin cell motility and wound healing. *European Molecular Biology Organization*, (26), pp.1221–1233.

Li, W., Sahu, D. & Tsen, F., 2012. Secreted heat shock protein-90 (Hsp90) in wound healing and cancer. *Biochimica et biophysica acta*, 1823(3), pp.730–41.

Li, W., Tsen, F., Sahu, D., Bhatia, A., Chen, M. & Multhoff, G., 2013. Extracellular Hsp90 (eHsp90) as the Actual Target in Clinical Trials: Intentionally or Unintentionally. *International Review of Cell and Molecular Biology*, (303), pp.203–235.

Lin, K., May, A.C.W. & Taylor, W.R., 2002. Threading using neural nEtwork (TUNE): the measure of protein sequence-structure compatibility. *Bioinformatics*, 18(10), pp.1350–1357.

Lipinski, C.A., 2004. Lead- and drug-like compounds: the rule-of-five revolution. *Drug discovery today. Technologies*, 1(4), pp.337–41.

Lipinski, C.A., Lombardo, F., Dominy, B.W. & Feeney, P.J., 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1-3), pp.3–25.

Louvion, J., Warth, R. & Picard, D., 1996. Two eukaryotespecific regions of hsp82 are dispensable for viability and signal transduction functions in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, (93), pp.13937–13942.

Lundgren, K., Holm, C. & Landberg, G., 2007. Hypoxia and breast cancer: prognostic and therapeutic implications. *Cellular and Molecular Life Sciences*, (64), pp.3233–3247.

MacKerell, A.D., Banavali, N. & Foloppe, N., 2001. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, 56(4), pp.257–65.

Mahanta, S., Pilla, S. & Paul, S., 2013. Design of novel Geldanamycin analogue hsp90 alpha-inhibitor in silico for breast cancer therapy. *Medical hypotheses*, 81(3), pp.463–9.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y. & Bryant, S.H., 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic acids research*, 30, pp.281–283.

Marcu, M.G., Schulte, T.W. & Neckers, L., 2000. Novobiocin and related coumarins and depletion of heat shock protein 90-dependent signaling proteins. *Journal of the National Cancer Institute*, 92(3), pp.242–8.

Meyer, P., Prodromou, C., Liao, C., Hu, B., Roe, S., Vaughan, C., Vlasic, I., Panaretou, B., Piper, P. & Pearl, L., 2004. Structural basis for recruitment of the ATPase activator Aha1 to the Hsp90 chaperone machinery. *European Molecular Biology Organization*, (23), pp.1402–1410.

Minami, Y., Kawasaki, H., Miyata, Y., Suzuki, K. & Yahara, I., 1991. Analysis of native forms and isoform compositions of the mouse 90-kDa heat shock protein, HSP90. *Journal of Biological Chemistry*, (266), pp.10099–10103.

Minami, Y., Kimura, H., Kawasaki, H., Suzuki, K. & Yahara, I., 1994. The carboxy-terminal region of mammalian HSP90 is required for its dimerization and function in vivo. *Molecular Cell Biology*, (14), pp.1459–1464.

Mobley, D. & Dill, K., 2009. Binding of Small-Molecule Ligands to Proteins: “What You See” Is Not Always “What You Get.” *Structure*, 17(4), pp.489–498.

Moore, S., Kozak, C., Robinson, E., Ulrich, S. & Appella, E., 1989. Murine 86- abd 84-kDa heat shock proteins,cDNA sequences, chromosome assignments, and evolutionary origin. *Journal of Biological Chemistry1*, (264), pp.5343–5351.

Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R. & Olson, A., 1998. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of computational chemistry*, 19(14), pp.1639–1662.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S. & Olson, A.J., 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16), pp.2785–91.

Nardai, G., Schnaider, T., Sti, C., Ryan, M., Hoj, P. & Csermely, P., 1996. Characterization of the 90 kDa heat shock protein (hsp90)-associated ATP/GTP-ase. *Journal of Biological Science*, (21), pp.179–190.

Neckers, L., 2003. Development of small molecule Hsp90 Inhibitors. *Current Opinion in Medical Chemistry*, (10), pp.733–739.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T. & Hutchison, G.R., 2011. Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1), p.33.

Obermann, W.M., Sondermann, H., Russo, A.A., Pavletich, N.P. & Hartl, F.U., 1998. In vivo function of Hsp90 is dependent on ATP binding and ATP hydrolysis. *The Journal of cell biology*, 143(4), pp.901–10.

Park, H., Kim, Y.-J. & Hahn, J.-S., 2007. A novel class of Hsp90 inhibitors isolated by structure-based virtual screening. *Bioorganic & medicinal chemistry letters*, 17(22), pp.6345–9.

Pawlowski, M., Gajda, M.J., Matlak, R. & Bujnicki, J.M., 2008. MetaMQAP: a meta-server for the quality assessment of protein models. *BMC bioinformatics*, 9, p.403.

Pearl, L., Prodromou, C. & Workman, P., 2008. The Hsp90 molecular chaperone: an open and shut case for treatment. *Journal of Biochemistry*, (410), pp.439–453.

Pei, J., Kim, B.-H. & Grishin, N. V, 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research*, 36(7), pp.2295–300.

Perozzo, R., Folkers, G. & Scapozza, L., 2004. Thermodynamics of protein-ligand interactions: History, presence, and future aspects. *Journal of Receptors and Signal Transduction*, 24, pp.1–52.

Plescia, J., Salz, W., Xia, F., Pennati, M., Zaffaroni, N., Daidone, M.G., Meli, M., Dohi, T., Fortugno, P., Nefedova, Y., Gabrilovich, D.I., Colombo, G. & Altieri, D.C., 2005. Rational design of shepherdin, a novel anticancer agent. *Cancer cell*, 7(5), pp.457–68.

Ponting, C.P., Schultz, J., Milpetz, F. & Bork, P., 1999. SMART: Identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Research*, 27, pp.229–232.

Prodromou, C., 2012. The “active life” of Hsp90 complexes. *Biochimica et biophysica acta*, 1823(3), pp.614–23.

Prodromou, C., Roe, S., O’Brien, R., Ladbury, J., Piper, P. & Pearl, L., 1997. Identification and structural characterization of the ATP/ADP-binding site in the hsp90 molecular chaperone. *Cell*, (90), pp.65–75.

Rastelli, G., 2014. Dimerization hot spots in the structure of human Hsp90. *Medicinal Chemistry Communications*, 5(6), p.797.

Rose, D., Wettenhall, R., Kudlicku, W., Kramer, G. & Hardesty, B., 1987. The 90-kilofalton peptide of the heme-regulated eIF-2-α kinase has sequence similarity with the 90-kilodalton heat shock protein. *Biochemistry*, (26), pp.6583–6587.

Sahu, D., Zhao, Z., Tsen, F., Cheng, C., Ryan, P., Fan, J., Dai, J., Eginli, A., Shams, S., Chen, M., Conti, P., Woodley, D. & Li, W., 2012. Identification of a novel tumor epitope in secreted Hsp90α for HIF-1α-overexpressing breast cancer. *Molecular Cell Biology*.

Sali, A. & Blundell, T., 1993. Comparitive Protein Modelling by Satisfaction of Spacial Restraints. *Journal of molecular biology*, 234, pp.799–815.

Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S., Jensen, J.H., Koseki, S., Matsunaga, N., Nguyen, K.A., Su, S., Windus, T.L., Dupuis, M. & Montgomery, J.A., 1993. General atomic and molecular electronic structure system. *Journal of Computational Chemistry*, 14(11), pp.1347–1363.

Schmitt, E., Gehrmann, M., Brunet, M., Multhoff, G. & Garrido, C., 2007. Intracellular and extracellular functions of heat shock proteins: repercussions in cancer therapy. *Journal of leukocyte biology*, 81(1), pp.15–27.

Schrodinger LLC, 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

Schulz-Gasch, T. & Stahl, M., 2004. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discovery Today: Technologies*, 1(3), pp.231–239.

Seeliger, D. & de Groot, B.L., 2010. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *Journal of computer-aided molecular design*, 24(5), pp.417–22.

Semenza, G., 2007. Evaluation of HIF-1 inhibitors as anticancer agents. *Drug discovery today*, (12), pp.853–859.

Sgobba, M., Forestiero, R., Degliesposti, G. & Rastelli, G., 2010. Exploring the binding site of C-terminal hsp90 inhibitors. *Journal of chemical information and modeling*, 50(9), pp.1522–8.

Shen, M. & Sali, A., 2006. Statistical potential for assessment and prediction of protein structures. *Protein science*, pp.2507–2524.

Shiau, A.K., Harris, S.F., Southworth, D.R. & Agard, D. a, 2006. Structural Analysis of E. coli hsp90 reveals dramatic nucleotide-dependent conformational rearrangements. *Cell*, 127(2), pp.329–40.

Simon, M. & Keith, B., 2008. The role of oxygen availability in embryonic development and stem cell function,. *Nature reviews. Molecular Cell biology*, (9), pp.285–296.

Söding, J., 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21, pp.951–960.

Söding, J., Biegert, A. & Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, 33(Web Server issue), pp.W244–8.

Song, X., Wang, X., Zhuo, W., Shi, H., Feng, D., Sun, Y., Liang, Y., Fu, Y., Zhou, D. & Luo, Y., 2010. The regulatory mechanism of extracellular Hsp90{alpha} on matrix metallopro- teinase-2 processing and tumor angiogenesis. *Journal of Biological Chemistry*, (285), pp.40039–40049.

Söti, C., Rácz, A. & Csermely, P., 2002. A Nucleotide-dependent molecular switch controls ATP binding at the C-terminal domain of Hsp90. N-terminal nucleotide binding unmasks a C-terminal binding pocket. *The Journal of biological chemistry*, 277(9), pp.7066–75.

Söti, C., Vermes, A., Haystead, T.A.J. & Csermely, P., 2003. Comparative analysis of the ATP-binding sites of Hsp90 by nucleotide affinity cleavage: a distinct nucleotide specificity of the C-terminal ATP-binding site. *European journal of biochemistry*, 270(11), pp.2421–8.

Southworth, D.R. & Agard, D. a, 2011. Client-loading conformation of the Hsp90 molecular chaperone revealed in the cryo-EM structure of the human Hsp90:Hop complex. *Molecular cell*, 42(6), pp.771–81.

Sreedhar, A., Soti, C. & Csermely, P., 2004. Inhibition of Hsp90: a new strategy for inhibiting protein kinases. *Biochimica et Biophysica Acta - Molecular Cell Research*, (1697), pp.233–242.

Stellas, D., El Hamidieh, A. & Patsavoudi, E., 2010. Monoclonal antibody 4C5 prevents activation of MMP2 and MMP9 by disrupting their interaction with extracellular HSP90 and in- hibits formation ofmetastatic breast cancer cell deposits. *Cell Biology*, (11), p.51.

Supko, J.G., Hickman, R.L., Grever, M.R. & Malspeis, L., 1995. Preclinical pharmacologic evaluation of geldanamycin as an antitumor agent. *Cancer Chemotherapy and Pharmacology*, 36, pp.305–315.

Tbarka, N., Richard-Mereau, C., Formstecher, P. & Dautrevaux, M., 1993. Biochemical and immunological evidence that an acidic domain of hsp90 is involved in the stabilization of untransformed glucocorticoid receptor complexes. *FEBS Letters*, 322(128-128).

Tembe, B. & McCammon, J., 1984. Ligand-receptor interactions. *Computational chemistry*, 79(8), pp.281–283.

Terasawa, K., Minami, M. & Minami, Y., 2008. Constantly updated knowledge of Hsp90. *Journal of Biochemistry*, (137), pp.443–447.

Toft, D.O., 1998. Recent Advances in the Study of hsp90 Structure and Mechanism of Action. , 9(6), pp.238–243.

Trepel, J., Mollapour, M., Giaccone, G. & Neckers, L., 2010. Targeting the dynamic HSP90 complex in cancer. *Nature Reviews Cancer*, (10), pp.537–549.

Trott, O. & Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2), pp.455–61.

Tsutsumi, S. & Neckers, L., 2007. Extracellular heat shock protein 90: a role for a molecular chaperone in cell motility and cancer metastasis. *Cancer science*, 98(10), pp.1536–9.

Vaidya, S., Ghosh, K. & Vundinti, B., 2011. Recent developments in drug resistance mechanism in chronic myeloid leukemia: a review. *Eukatyotic Journal of Haematology*, (87), pp.381–393.

Venclovas, C., Zemla, A., Fidelis, K. & Moult, J., 2003. Assessment of progress over the casp experiments. *Proteins*, 53, pp.585–595.

Verkhivker, G.M., Dixit, A., Morra, G. & Colombo, G., 2009. Structural and computational biology of the molecular chaperone Hsp90: from understanding molecular mechanisms to computer-based inhibitor design. *Current topics in medicinal chemistry*, 9(15), pp.1369–85.

Wagner, a Ben, 2006. SciFinder Scholar 2006: an empirical analysis of research topic query processing. *Journal of chemical information and modeling*, 46(2), pp.767–74.

Wallace, A.C., Laskowski, R.A. & Thornton, J.M., 1995. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering*, 8(2), pp.127–134.

Wallner, B. & Elofsson, A., 2006. Identification of correct regions in protein models using structural , alignment , and consensus information. , pp.900–913.

Wang, R., Lai, L. & Wang, S., 2002. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design*, 16(1), pp.11–26.

Waszkowycz, B., Clark, D. & Gancia, E., 2011. Outstanding challenges in protein–ligand docking and structure-based virtual screening. *WIREs Computational*, 1, pp.229–259.

Welch, W. & Brown, C., 1996. Influence of molecular and chemical chaperones on protein foldingNo Title. *Cell Stress Chaperones*, (1), pp.109–115.

Whitesell, L. & Lin, N., 2012. HSP90 as a platform for the assembly of more effective cancer chemotherapy. *Biochimica et biophysica acta*, 1823(3), pp.756–66.

Whitesell, L. & Lindquist, S., 2005. HSP90 and the Chaperoning of Cancer. *Nature Reviews Cancer*, 10, pp.761–772.

Wiech, H., Buchner, J., Zimmermann, R. & Jakob, U., 1992. Hsp90 chaperones protein folding in vitro. *Nature*, (358), pp.169–170.

Wiederstein, M. & Sippl, M.J., 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic acids research*, 35(Web Server issue), pp.W407–10.

Xu, D. & Zhang, Y., 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, 80(7), pp.1715–1735.

Yamamoto, M., Takahashi, Y., Inano, K., Horigome, T. & Sugano, H., 1991. Characterization of the hydrophobic region of heat shock protein 90. *Journal of Biochemistry*, (110), pp.141–145.

Young, J., Schneider, C. & Hartl, F., 1997. In vitro evidence that hsp90 contains two independent chaperone sites. *FEBS Letters*, (418), pp.139–143.

Young, J.C., Moarefi, I. & Hartl, F.U., 2001. Hsp90: a specialized but essential protein-folding tool. *The Journal of cell biology*, 154(2), pp.267–73.

Zhang, T., Hamza, A., Cao, X., Wang, B., Yu, S., Zhan, C.-G. & Sun, D., 2008. A novel Hsp90 inhibitor to disrupt Hsp90/Cdc37 complex against pancreatic cancer cells. *Molecular cancer therapeutics*, 7(1), pp.162–70.

APPENDIX A

A-1: Example PIR file format used in homology model building

```
>P1;Target
sequence:gi_154146191_ref_NP_00533:15:.:640:.:::::
PMEEEEVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIRYESLTDPSKLDSGKELHINL
IP
NKQDRTLTIVDTGIGMTKADLINNLGTIAKSGTKAFMEALQAGADISMIGQFGVGFYSAYLVAEKVTVIT
KH
NDDEQYAWESSAGGSFTVRTD-
TGEPMGRGTKVILHLKEDQTEYLEERRIKEIVKKHSQFIGYPITLFVEKE
RDKEVSIKEKYIDQEELNKTKPIWTRNPDDITNEEYGEFYKSLTNDWEDHLAVKHFSVEGQLEFRALLFV
PR
RAPFDLFENRKKKNNIKLYVRRVFIMDNCEELIPEYLNFIRGVVDSEDLPLNISREMLQQSKILKVIRKN
LV
KKCLELFTELAEDKENYKKFYEQFSKNIKLGIHEDSQNRKKLSELLRYYTSASGDEMVSLKDYCTRMKEN
QK
HIYYITGETKDQVANSAFVERLRKHGLEVIYMIEPIDEYCVQQLKEFEGKTLVSVTKEGLELPEDEEEKK
KQ
EEKKTKFENLCKIMKDILEKKVEKVVVSNRLVTSPCCIVTSTYGWTANMERIMKAQALKKHLEINPDHSI
IE
TLRQKA-EADKNDKSVKDLVILLYETALLSSGFSLEDPQTHANRIYRMIKLGLG/
PMEEEEVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIRYESLTDPSKLDSGKELHINL
IP
NKQDRTLTIVDTGIGMTKADLINNLGTIAKSGTKAFMEALQAGADISMIGQFGVGFYSAYLVAEKVTVIT
KH
NDDEQYAWESSAGGSFTVRTD-
TGEPMGRGTKVILHLKEDQTEYLEERRIKEIVKKHSQFIGYPITLFVEKE
RDKEVSIKEKYIDQEELNKTKPIWTRNPDDITNEEYGEFYKSLTNDWEDHLAVKHFSVEGQLEFRALLFV
PR
RAPFDLFENRKKKNNIKLYVRRVFIMDNCEELIPEYLNFIRGVVDSEDLPLNISREMLQQSKILKVIRKN
LV
KKCLELFTELAEDKENYKKFYEQFSKNIKLGIHEDSQNRKKLSELLRYYTSASGDEMVSLKDYCTRMKEN
QK
HIYYITGETKDQVANSAFVERLRKHGLEVIYMIEPIDEYCVQQLKEFEGKTLVSVTKEGLELPEDEEEKK
KQ
EEKKTKFENLCKIMKDILEKKVEKVVVSNRLVTSPCCIVTSTYGWTANMERIMKAQALKKHLEINPDHSI
IE
TLRQKA-EADKNDKSVKDLVILLYETALLSSGFSLEDPQTHANRIYRMIKLGLG*


>P1;Template1
structure:2CG9:2:A:677:A::::
-----
ASETFEFQAEITQLMSLIINTVYSNKEIFLRELISNASDALDKIRYKSLSDPKQLETEPDLFIRITP
KPEQKVLEIRDSGIGMTKAELINNLGTIAKSGTKAFMEALSAGADVSMIGQFGVGFYSLFLVADRVQVIS
KS
NDDEQYIWESNAGGSFTVTLDEVNERIGRGTILRLFLKDDQLEYLEEKRIKEVIKRHSEFVAYPIQLVVT
KE
```

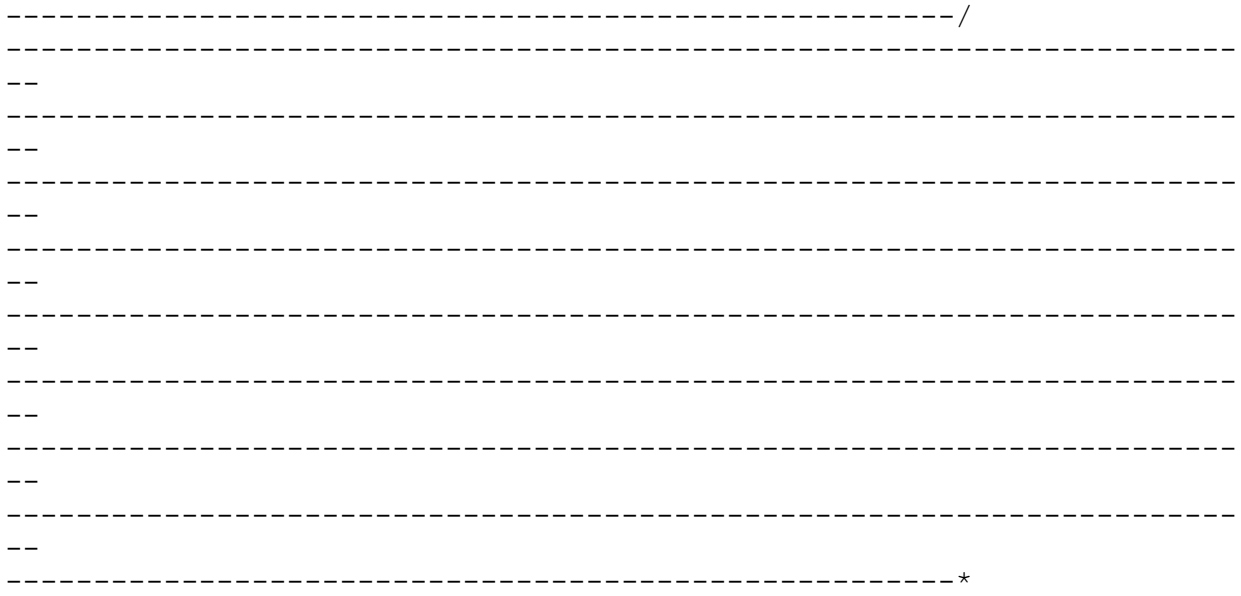
```
VEKE----
EEVQEIEELNKTKPLWTRNPSDITQEEYNAFYKSISNDWEDPLYVKHFSVEGQLEFRAILFIPK
RAPF---------
NNIKLYVRRVFITDEAEDLIPEWLSFVKGVVDSEDLPLNLSREMLQQNKIMKVIRKNIV
KKLIEAFNEIAEDSEQFEKFYSAFSKNIKLGVHEDTQNRAALAKLLRYNSTKSVDELTSLTDYVTRMPEH
QK
NIYYITGESLKAVEKSPFLDALKAKNFEVLFLTDPIDEYAFTQLKEFEGKTLVDITKD-
FELEETDEEKAER
EKEIKEYEPLTKALKEILGDQVEKVVVSYKLLDAPAAIRTGQFGWSANMERIMKAQA--
KTFEISPKSPIIK
ELKKRVDEGGAQDKTVKDLTKLLYETALLTSGFSLDEPTSFASRINRLISLGLN/
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
----------------------------------------------------------*
>P1;Template2
structure:2CG9:2:B:677:B::::
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
----------------------------------------------------------/
------------------------------------------------------------------------
--
-----
ASETFEFQAEITQLMSLIINTVYSNKEIFLRELISNASDALDKIRYKSLSDPKQLETEPDLFIRITP
KPEQKVLEIRDSGIGMTKAELINNLGTIAKSGTKAFMEALSAGADVSMIGQFGVGFYSLFLVADRVQVIS
KS
```

```
NDDEQYIWESNAGGSFTVTLDEVNERIGRGTILRLFLKDDQLEYLEEKRIKEVIKRHSEFVAYPIQLVVT
KE
VEKE----
EEVQEIEELNKTKPLWTRNPSDITQEEYNAFYKSISNDWEDPLYVKHFSVEGQLEFRAILFIPK
RAPFDLFESKKKKNNIKLYVRRVFITDEAEDLIPEWLSFVKGVVDSEDLPLNLSREMLQQNKIMKVIRKN
IV
KKLIEAFNEIAEDSEQFEKFYSAFSKNIKLGVHEDTQNRAALAKLLRYNSTKSVDELTSLTDYVTRMPEH
QK
NIYYITGESLKAVEKSPFLDALKAKNFEVLFLTDPIDEYAFTQLKEFEGKTLVDITKD-
FELEETDEEKAER
EKEIKEYEPLTKALKEILGDQVEKVVVSYKLLDAPAAIRTGQFGWSANMERIMKAQA--
KTFEISPKSPIIK
ELKKRVDEGGAQDKTVKDLTKLLYETALLTSGFSLDEPTSFASRINRLISLGLN*


>P1;Template3
structure:3T10:11:A:224:A::::
PMEEEEVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIRYESLTDPSKLDSGKELHINL
IP
NKQDRTLTIVDTGIGMTKADLINNLGTIAKSGTKAFMEALQAGADISMIGQFGVGFYSAYLVAEKVTVIT
KH
NDDEQYAWESSAGGSFTVRTD-
TGEPMGRGTKVILHLKEDQTEYLEERRIKEIVKKHSQFIGYPITLFVEK-
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
-----------------------------------------------------/
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
-----------------------------------------------------*
>P1;Template4
structure:3T10:11:A:224:A::::
```

```
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
-----------------------------------------------------------/
PMEEEEVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIRYESLTDPSKLDSGKELHINL
IP
NKQDRTLTIVDTGIGMTKADLINNLGTIAKSGTKAFMEALQAGADISMIGQFGVGFYSAYLVAEKVTVIT
KH
NDDEQYAWESSAGGSFTVRTD-
TGEPMGRGTKVILHLKEDQTEYLEERRIKEIVKKHSQFIGYPITLFVEK-
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
-----------------------------------------------------------*

>P1;Template5
structure:2CG9:325:B:342:B::::
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
-K
RAPFDLFESKKKKNNIK----------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
------------------------------------------------------------------------
--
```



A-2: Example of MODELLER script used for executing model building

```
#!/usr/bin/python
# Homology modelling by the automodel class

from modeller import *
#from modeller.parallel import *
from modeller.automodel import *    # Load the automodel class

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

# directories for input atom files
env.io.atom_files_directory =
'/home/david/Project/Modelling/Models/Files'

a = automodel(env,
              alnfile  = '../../PIR10.pir',     # alignment filename
              knowns   =
('Template1','Template2','Template3','Template4','Template5'),
# codes of the templates
              sequence = 'Target')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                # index of the last model
                                    # (determines how many models to
calculate)
a.final_malign3d = True          #generate superimposed template

a.md_level = refine.very_slow

#j = job()
#for i in range(60):
#    j.append(local_slave())

#a.use_parallel_job(j)

a.make()                            # do the actual homology mode
```

APPENDIX B

B-1: Example of docking parameter file used for AutoDock4 runs

```
ga_pop_size 50
ga_num_evals 5000000
ga_num_generations 10000000
ga_elitism 1
ga_mutation_rate 0.02
ga_crossover_rate 0.8
ga_window_size 10
ga_cauchy_alpha 0.0
ga_cauchy_beta 1.0
set _ga
ga_run 50
```

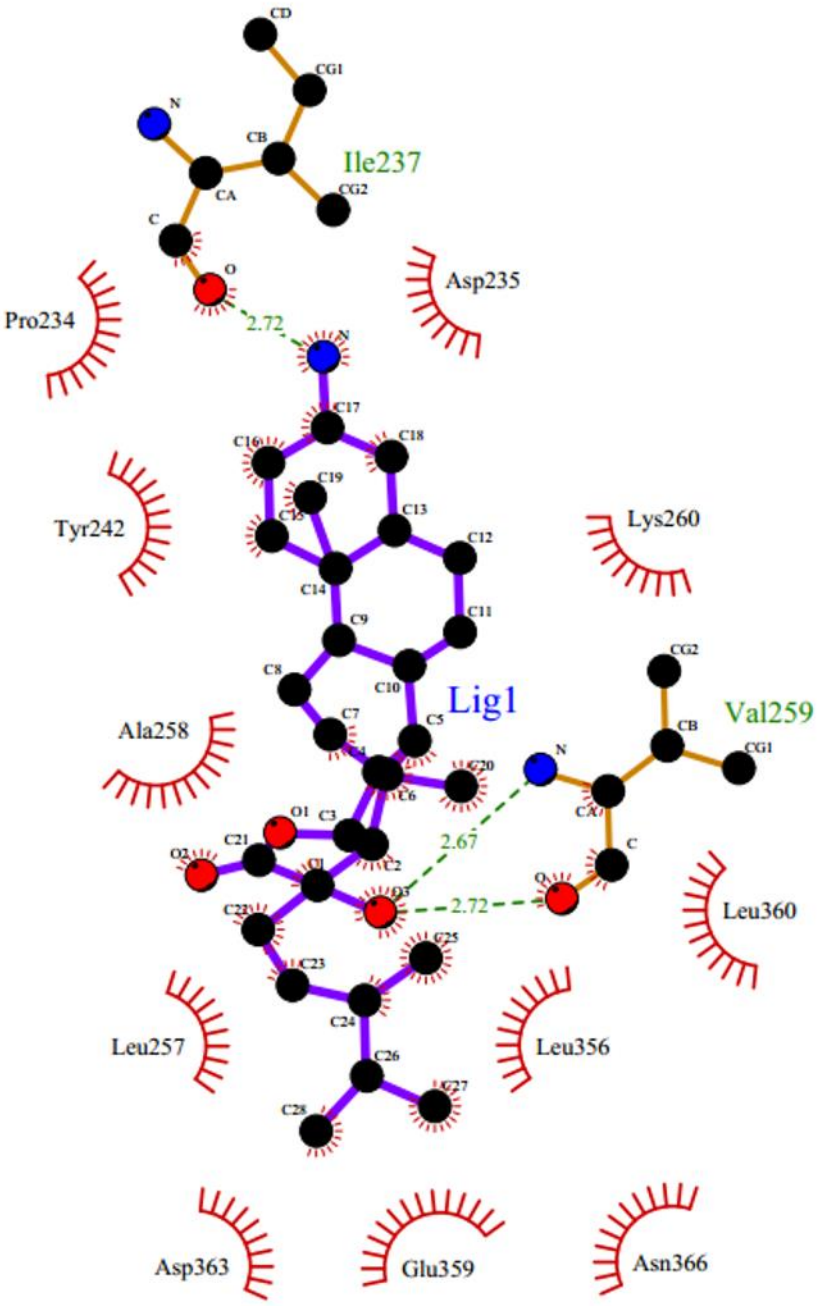
B-2: X-score parameter file

```
#########################################################################
#                              XTOOL/SCORE                              #
#########################################################################
###
FUNCTION   SCORE
###
### set up input and output files -----------------------------------
###
#
RECEPTOR_PDB_FILE    ./HM_Chain_B.pdb
#REFERENCE_MOL2_FILE  ./
#COFACTOR_MOL2_FILE  none
LIGAND_MOL2_FILE     ./ligand1.mol2
#
OUTPUT_TABLE_FILE    ./xscore.table
OUTPUT_LOG_FILE      ./xscore.log
###
### how many top hits to extract from the LIGAND_MOL2_FILE?
###
NUMBER_OF_HITS       5
HITS_DIRECTORY       ./hits.mdb
###
### want to include atomic binding scores in the resulting Mol2 files?
###
SHOW_ATOM_BIND_SCORE  YES        [YES/NO]
###
### set up scoring functions ----------------------------------------
###
APPLY_HPSCORE         YES             [YES/NO]
     HPSCORE_CVDW  0.004
     HPSCORE_CHB   0.053
     HPSCORE_CHP   0.011
     HPSCORE_CRT  -0.061
     HPSCORE_C0    3.448
APPLY_HMSCORE         YES             [YES/NO]
     HMSCORE_CVDW  0.004
     HMSCORE_CHB   0.094
     HMSCORE_CHM   0.394
     HMSCORE_CRT  -0.099
     HMSCORE_C0    3.585
APPLY_HSSCORE         YES        [YES/NO]
     HSSCORE_CVDW  0.004
     HSSCORE_CHB   0.069
     HSSCORE_CHS   0.004
     HSSCORE_CRT  -0.092
     HSSCORE_C0    3.349
###
### set up chemical rules for pre-screening ligand molecules ---------
###
APPLY_CHEMICAL_RULES    NO            [YES/NO]
```

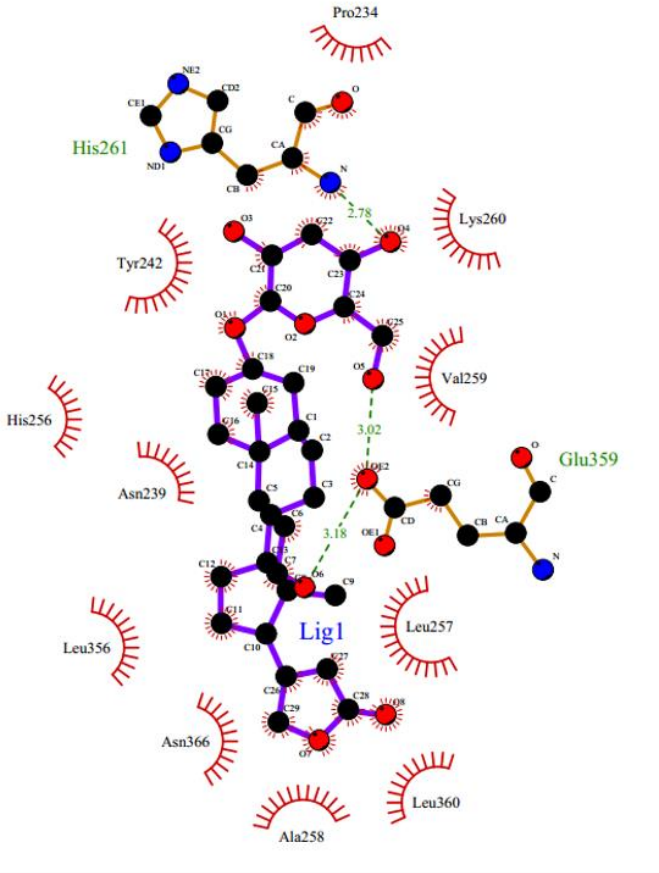
```
     MAXIMAL_MOLECULAR_WEIGHT      600.0
     MINIMAL_MOLECULAR_WEIGHT      200.0
     MAXIMAL_LOGP                  6.00
     MINIMAL_LOGP                  1.00
     MAXIMAL_HB_ATOM               8
     MINIMAL_HB_ATOM               2
###
###
```

APPENDIX C

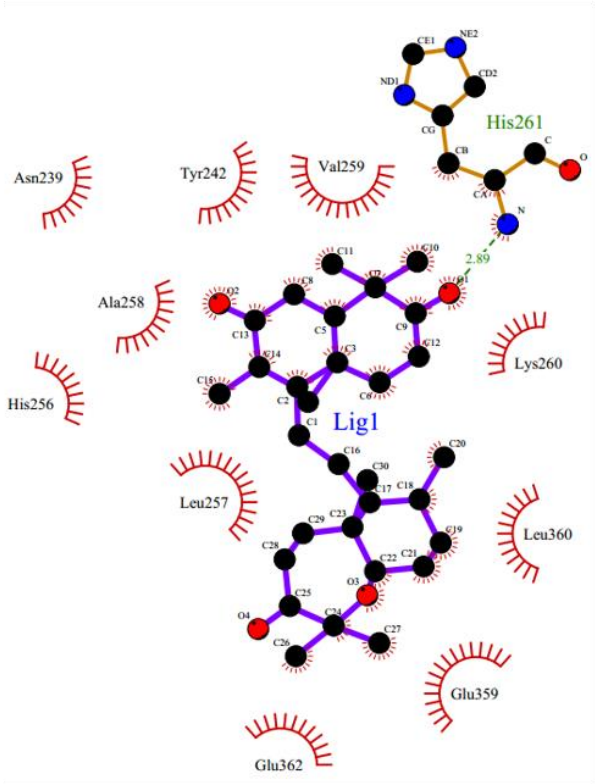
C-1: LigPlot+ 2D interaction maps for all ligands bound at Target site 1



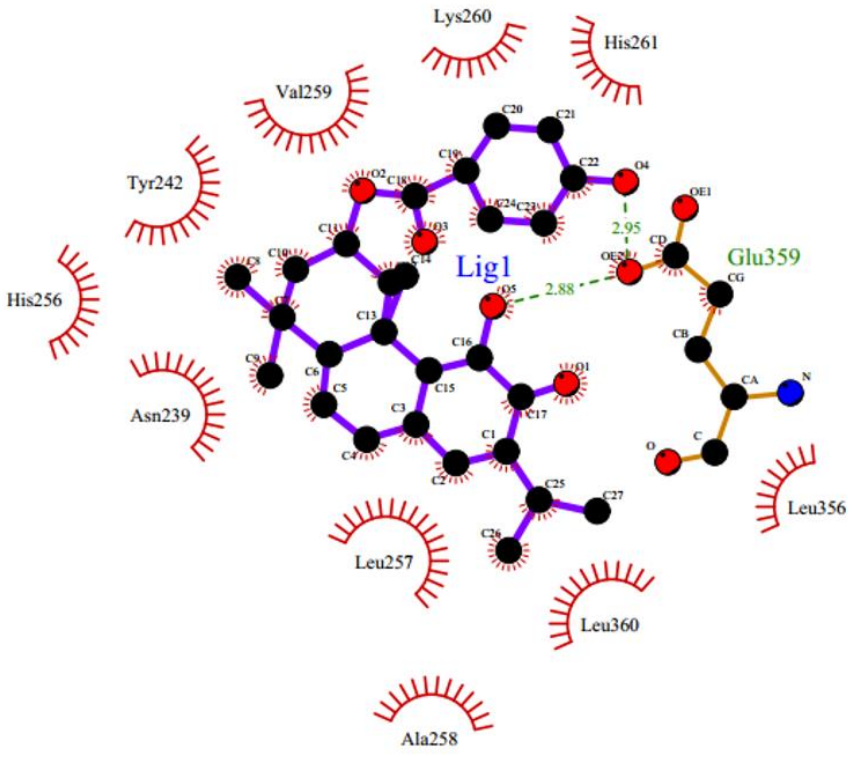
1042138-27-1



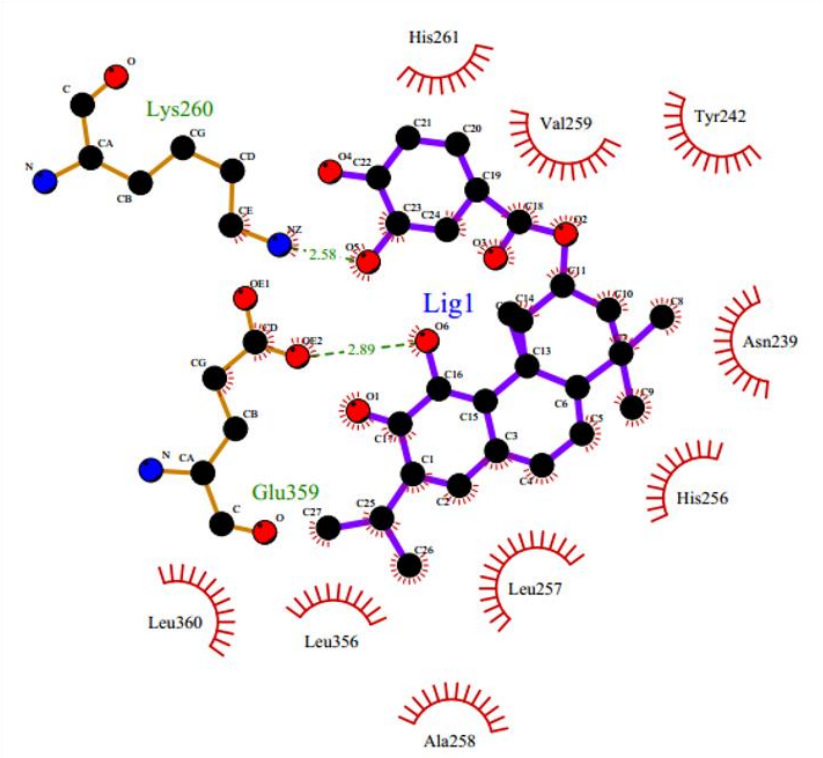
17059-16-4



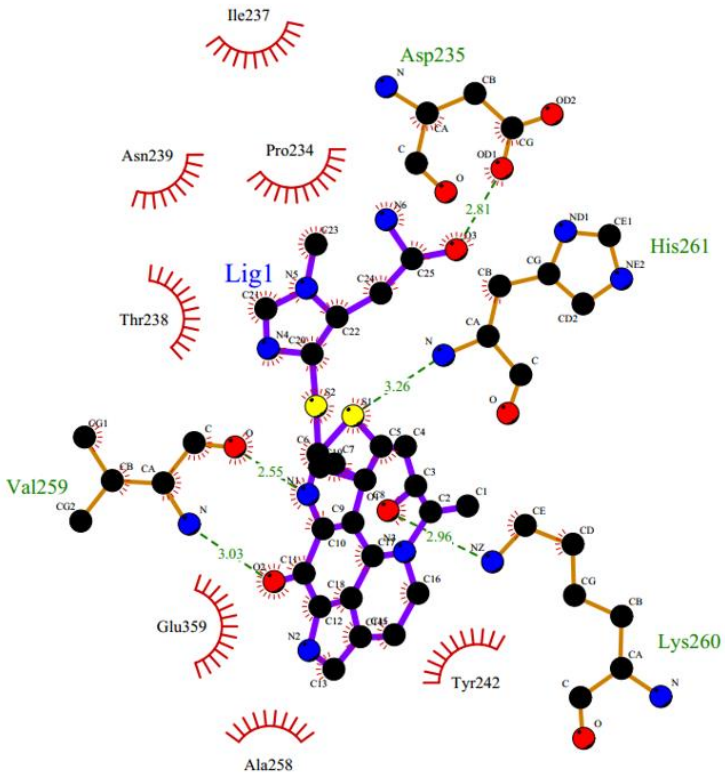
233607-72-2



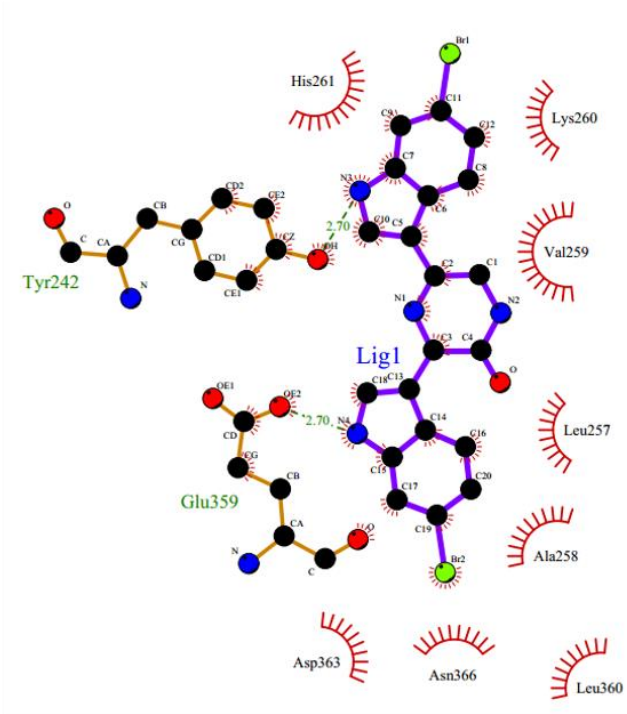
66656-57-3



66656-58-4

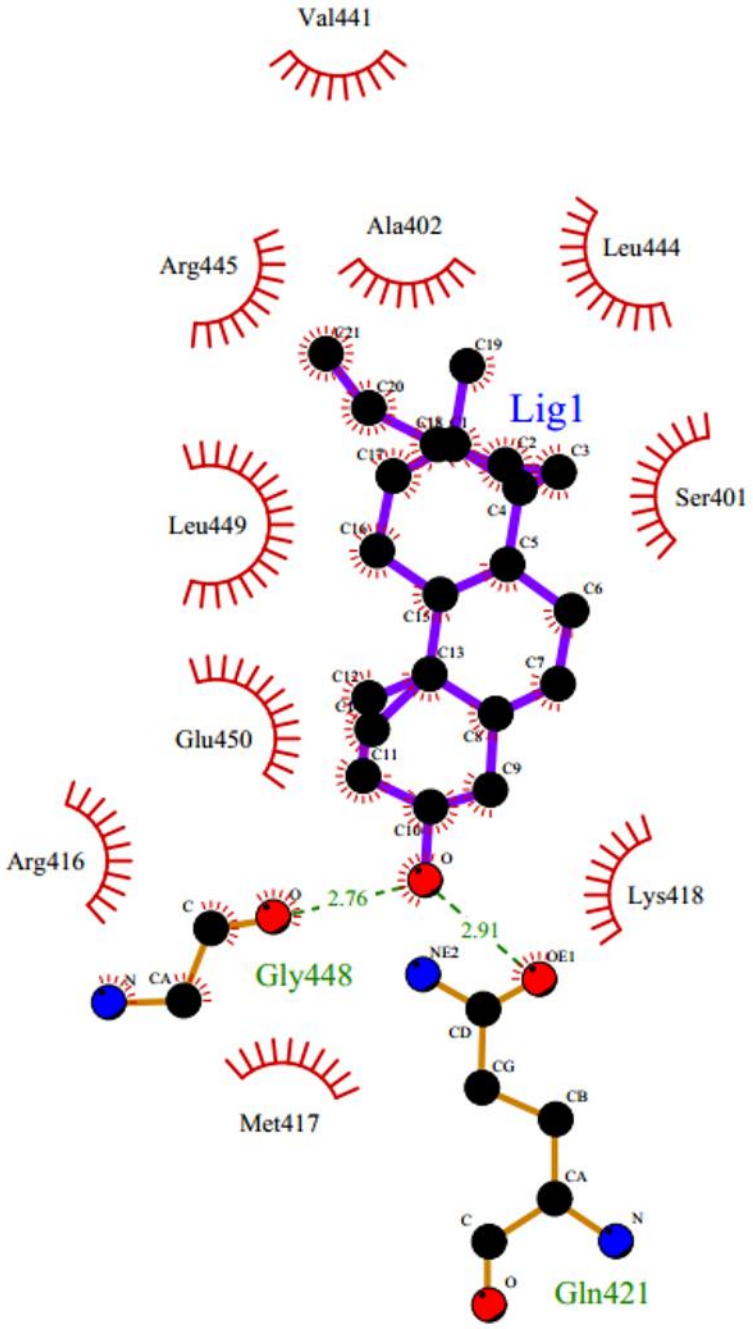


721395-11-5

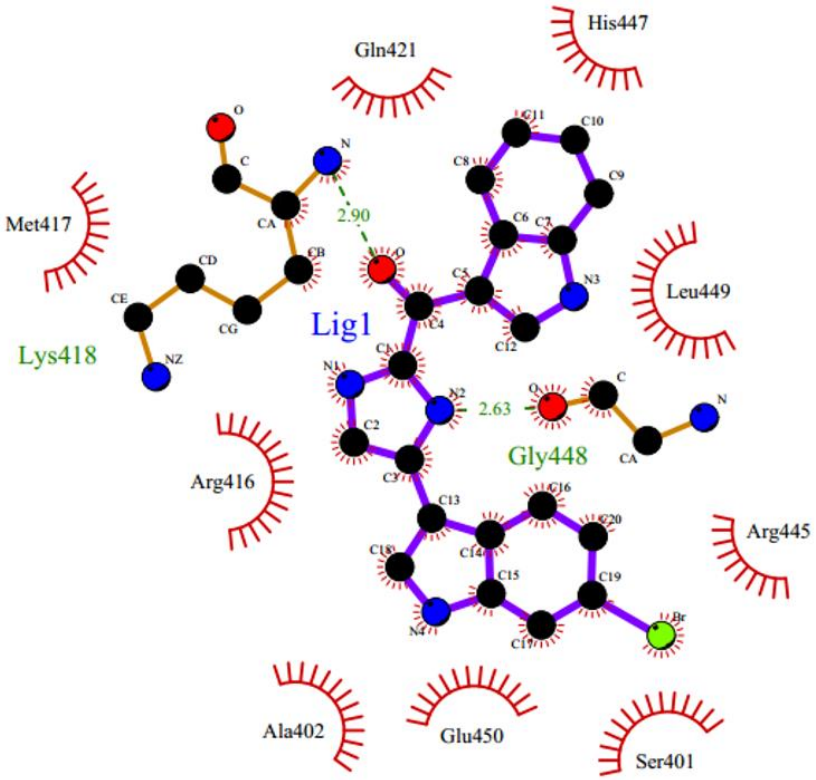


264624-39-7

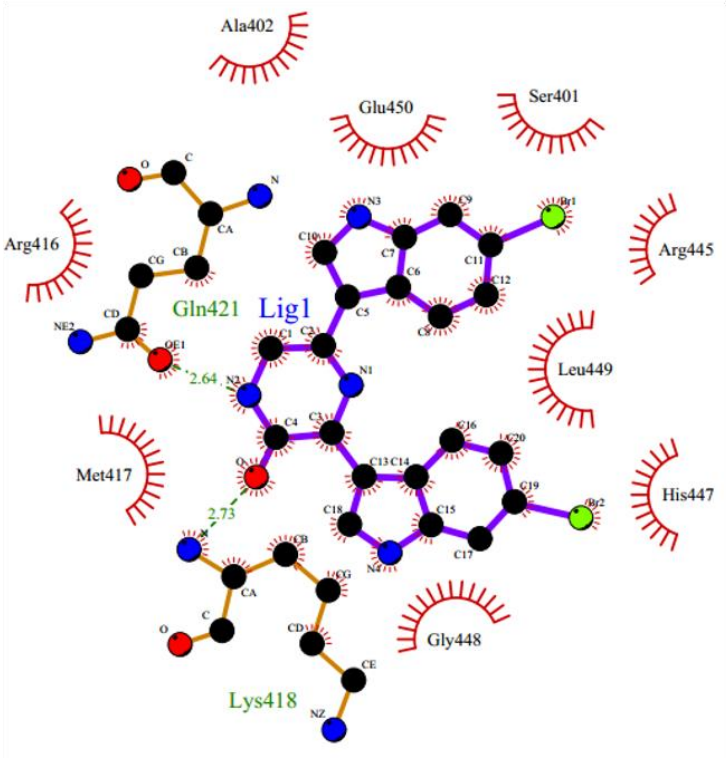
C-2: LigPlot+ 2D interaction maps for all ligands bound at Target site 2



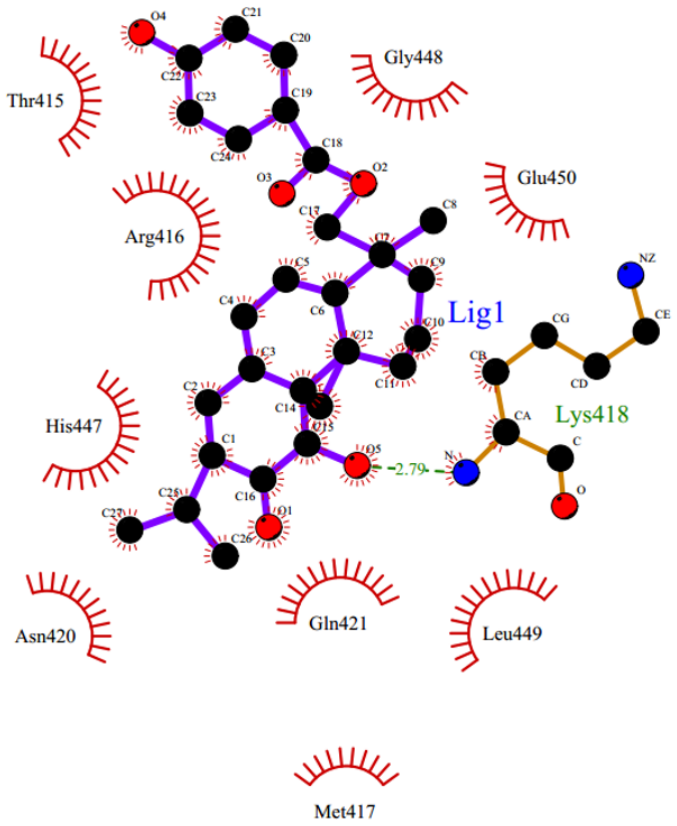
21321-88-0



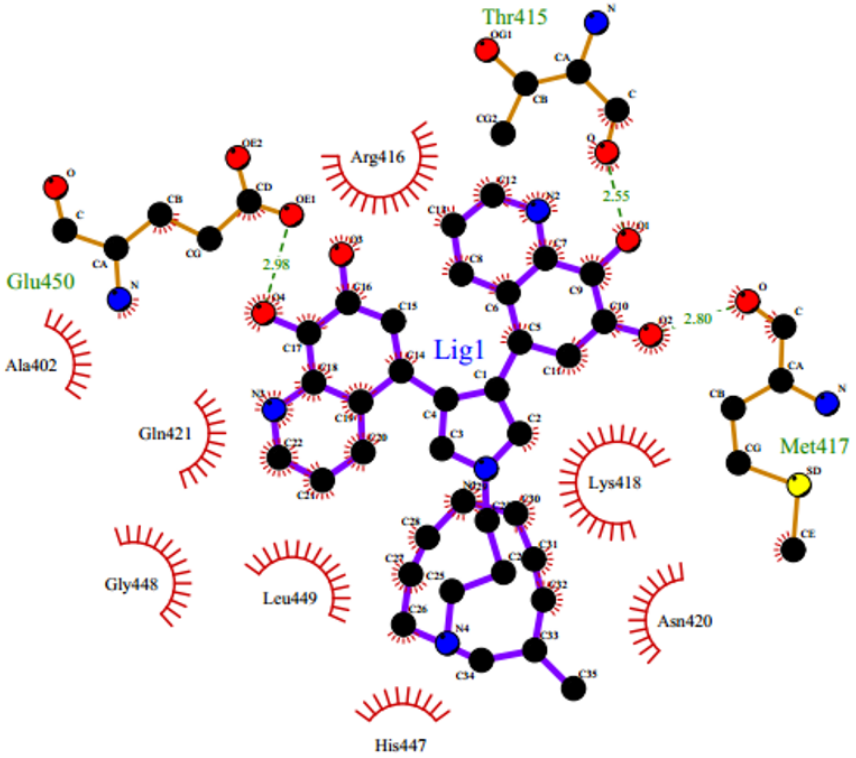
180633-55-0



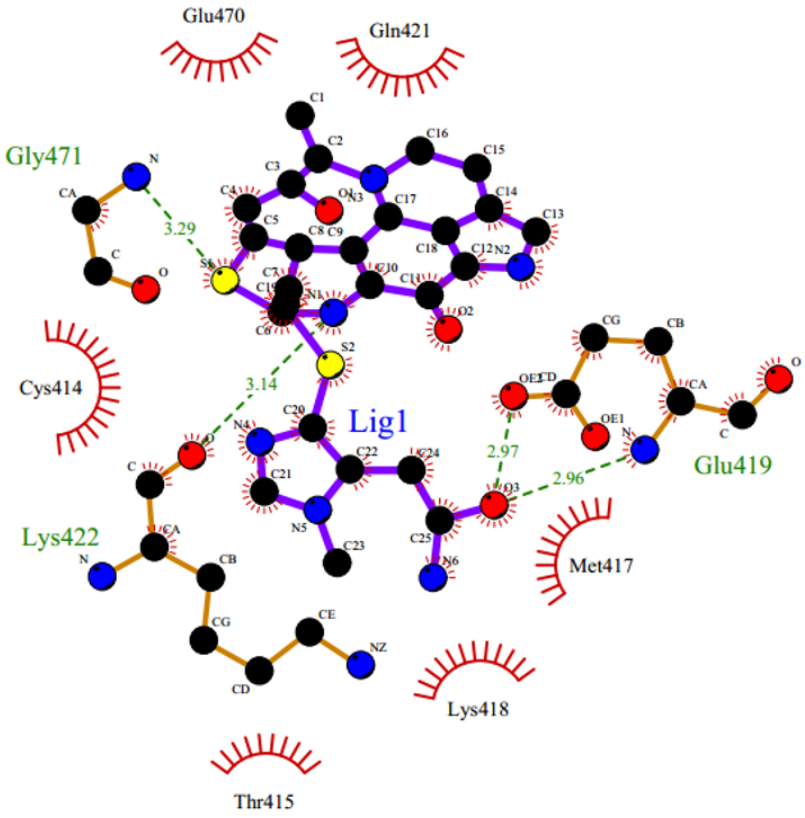
264624-39-7



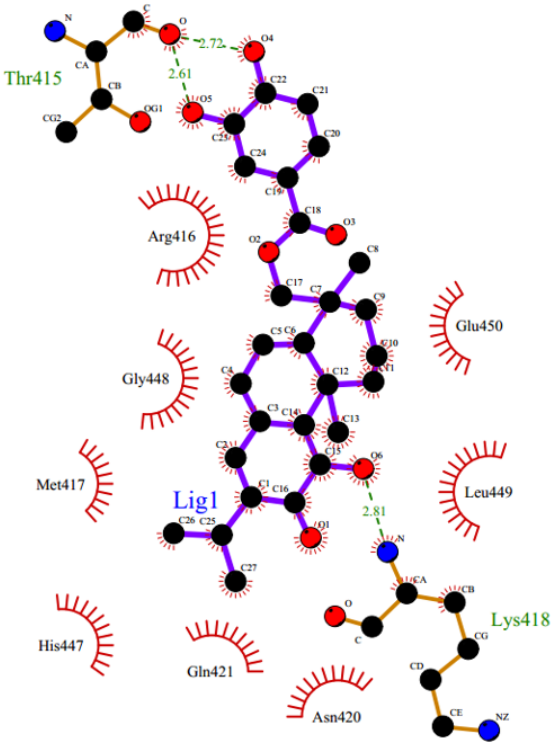
66656-56-2



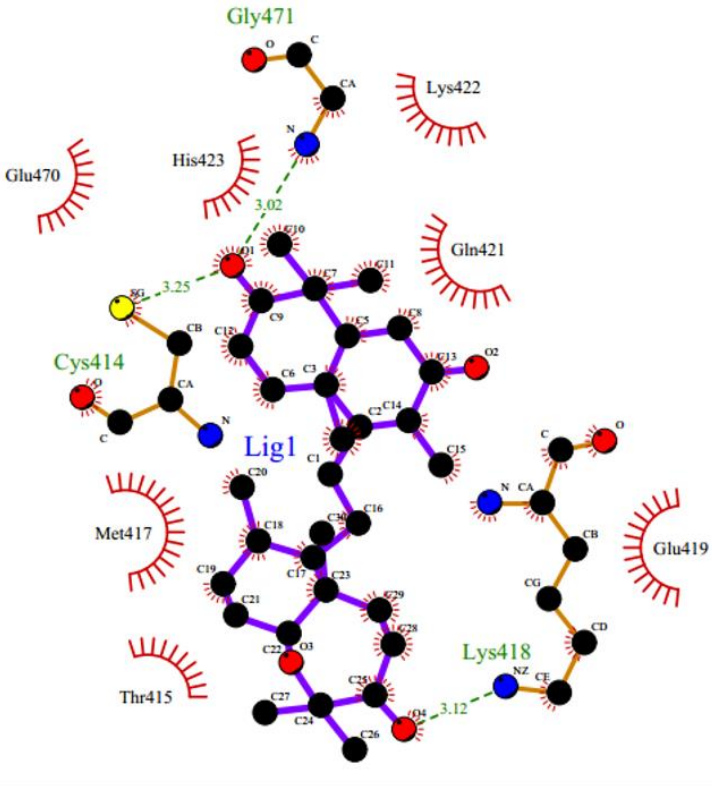
221367-90-4



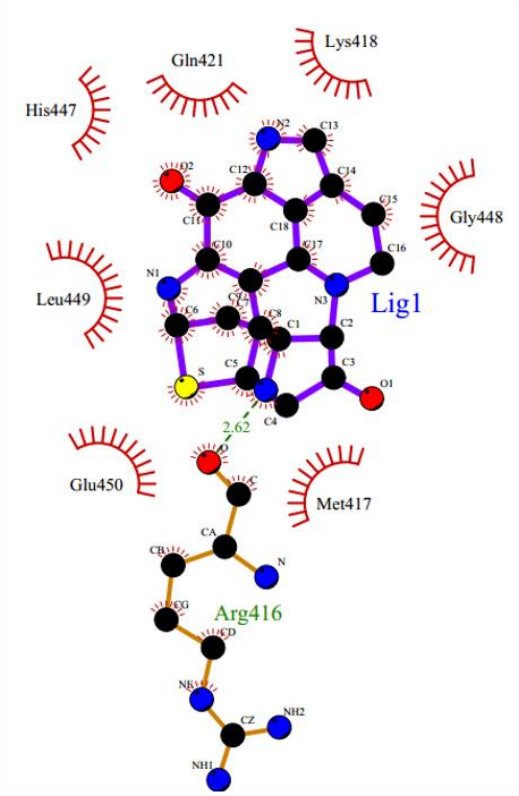
721395-11-5



66700-66-1

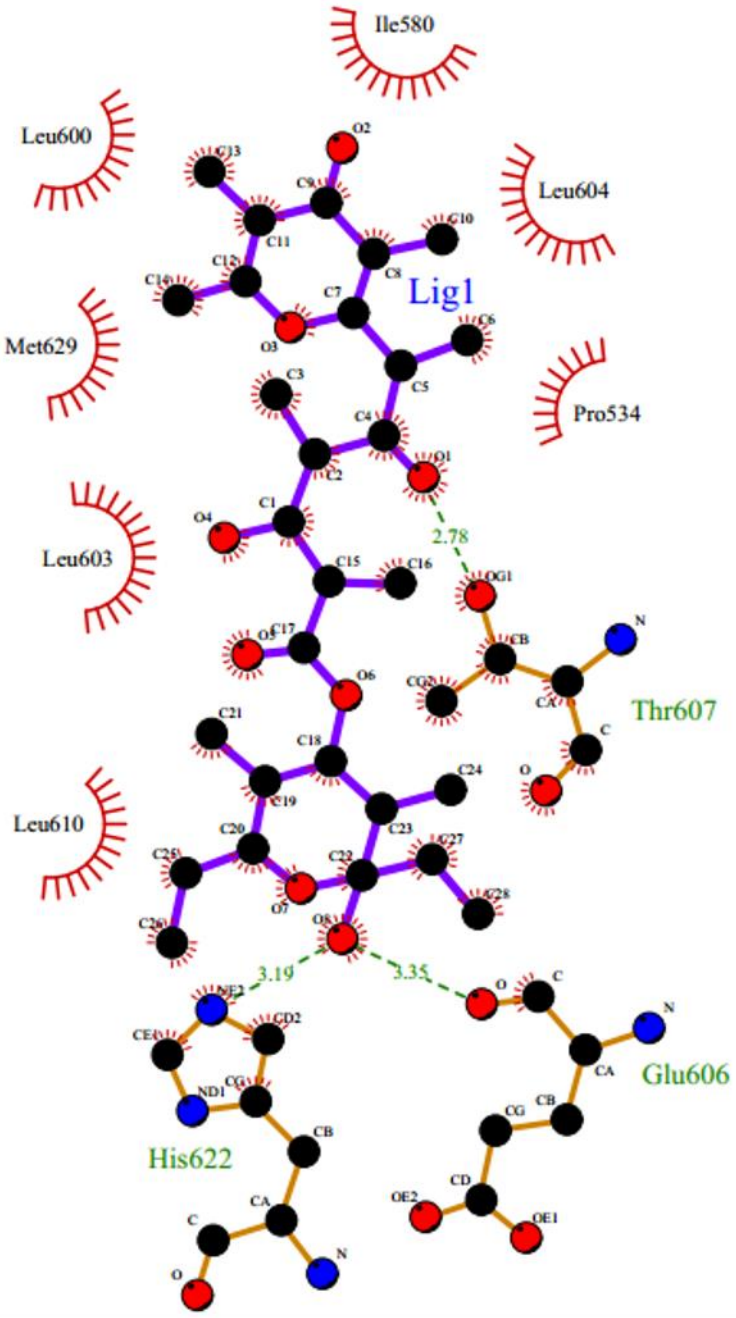


233607-72-2

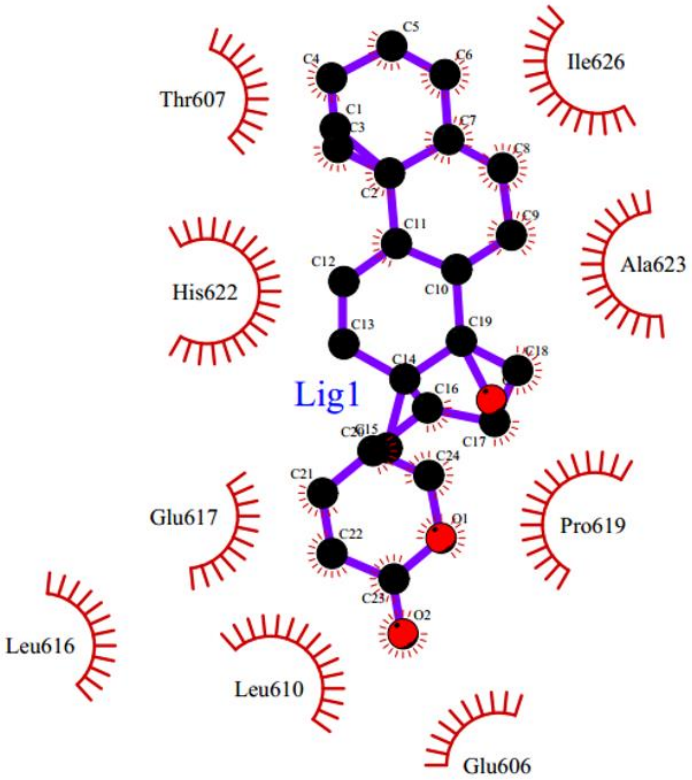


721395-40-0

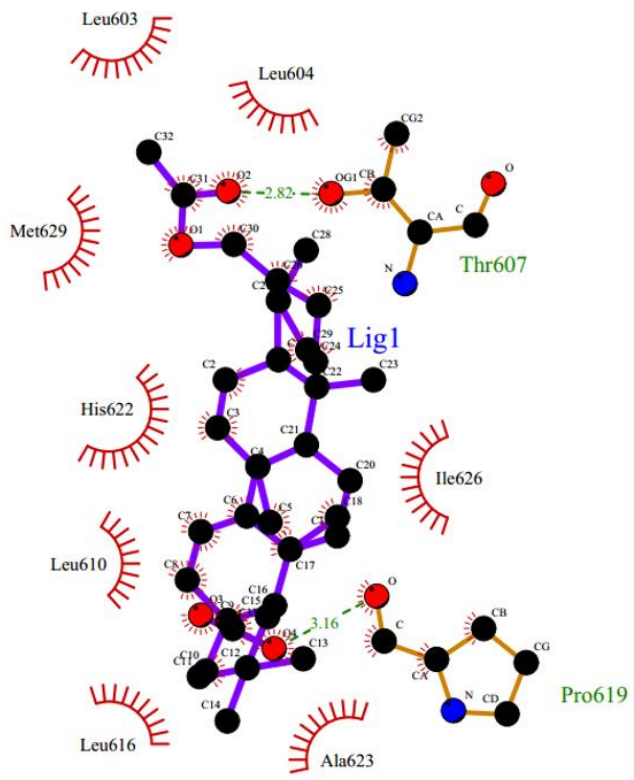
C-3: LigPlot+ 2D interaction maps for all ligands bound at Target site 3



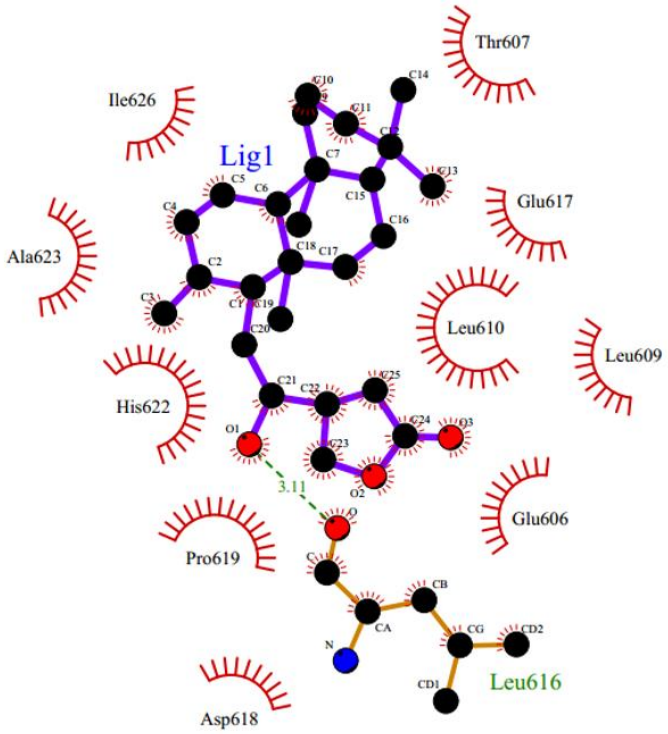
274913-20-1



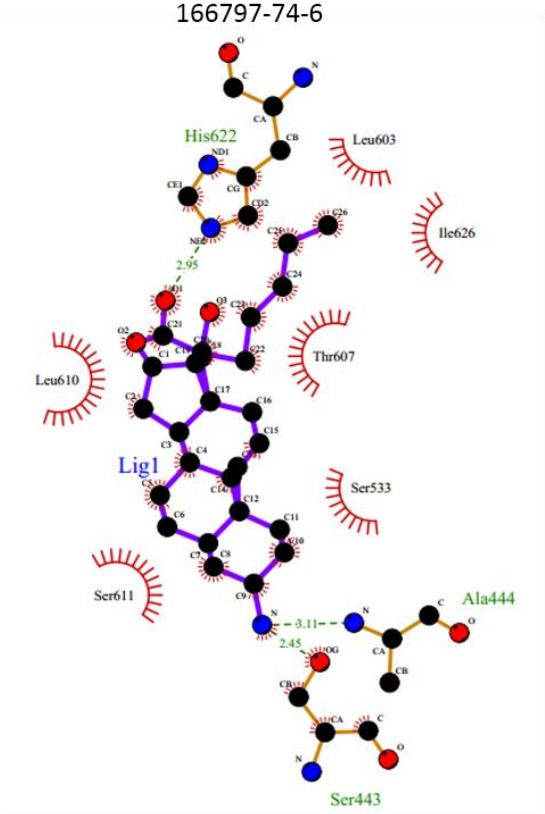
545-51-7



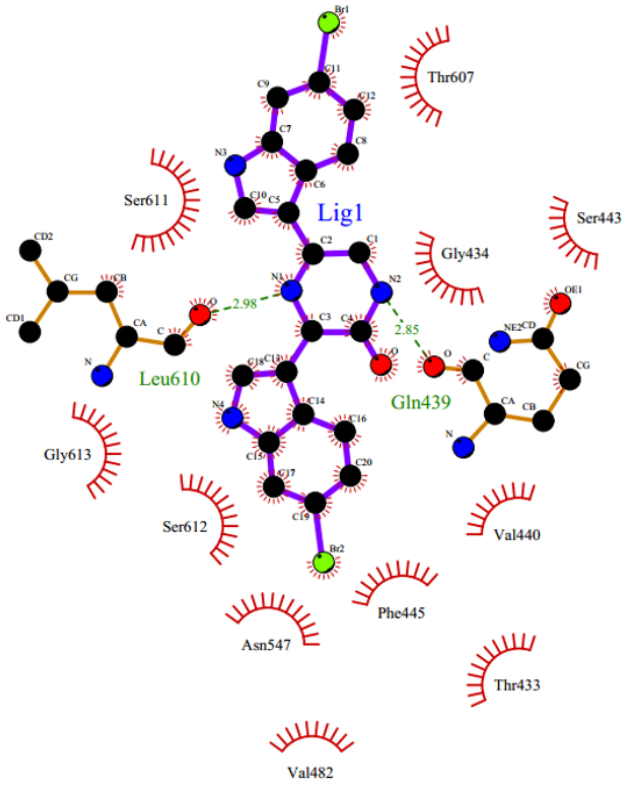
28937-85-1



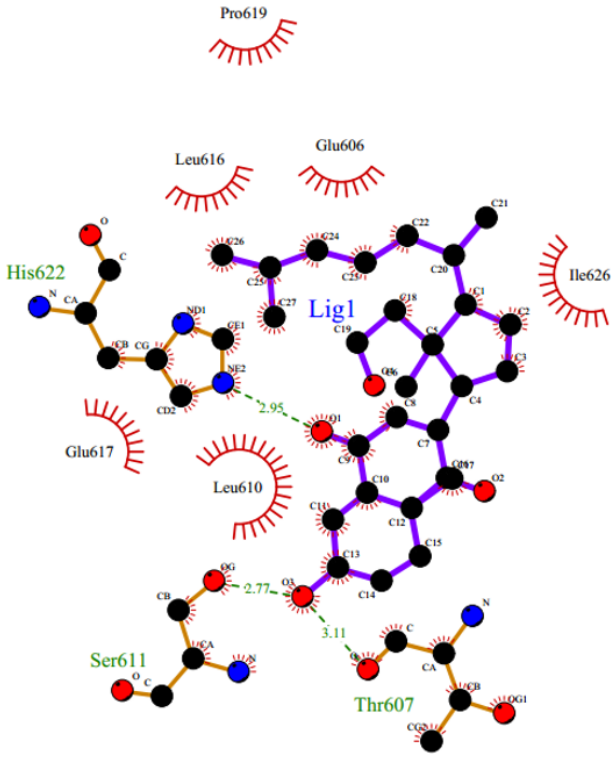
166797-74-6



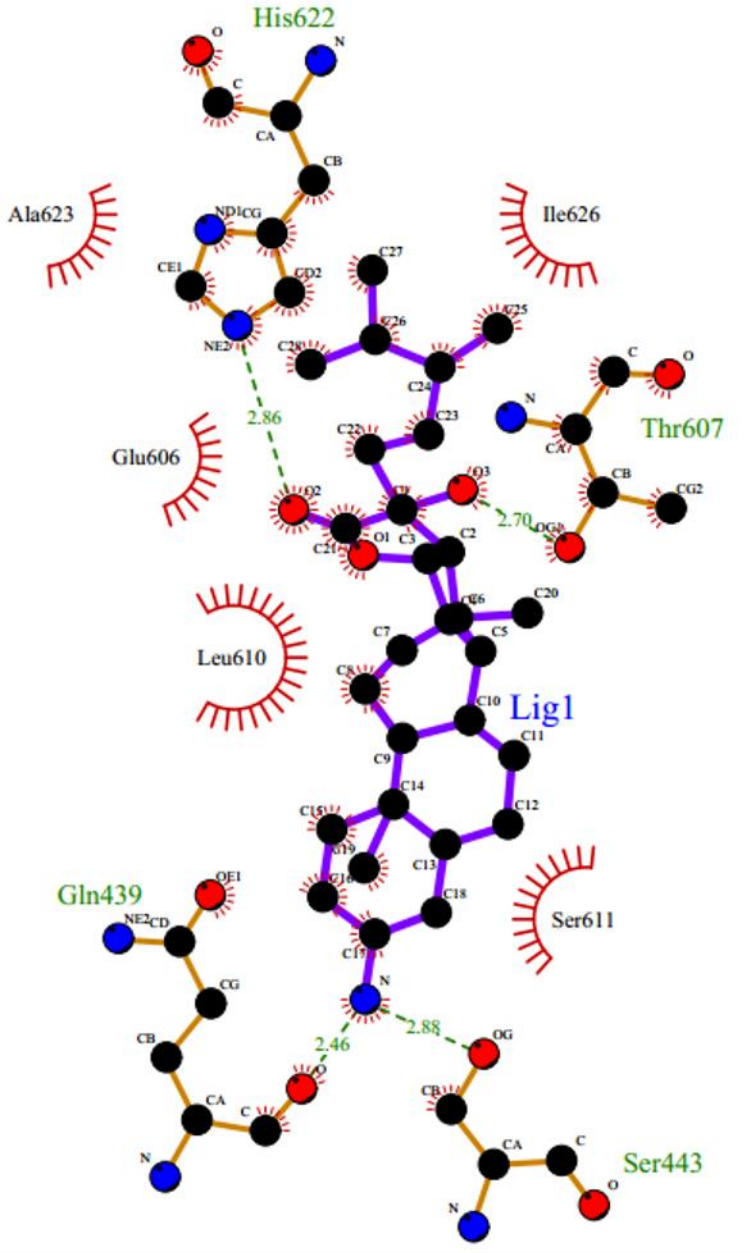
1042138-28-2



264624-39-7



1015763-02-6


His622
Ala623
Ile626
Glu606
Thr607
Leu610
Lig1
Gln439
Ser611
Ser443
2.86
2.70
2.46
2.88

1042138-27-1