

AN IN-SILICO INVESTIGATION OF MORITA-BAYLIS-HILLMAN ACCESSIBLE HETEROCYCLIC ANALOGUES FOR APPLICATIONS AS NOVEL HIV-1 C PROTEASE INHIBITORS

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

Master of Science
(Biochemistry and Computational Molecular Biology)

by

Coursework and Thesis

at

Rhodes University
(Department of
Biochemistry, Microbiology and Biotechnology)

by

Lester Takunda Sigauke

B.Sc. (Hons.) (Rhodes University)

April 2015

Abstract

Cheminformatic approaches have been employed to optimize the *bis*-coumarin scaffold identified by Onywera et al. (2012) as a potential hit against the protease HIV-1 protein. The Open Babel library of commands was used to access functions that were incorporated into a markov chain recursive program that generated 17750 analogues of the *bis*-coumarin scaffold. The Morita-Baylis-Hillman accessible heterocycles were used to introduce structural diversity within the virtual library. *In silico* high through-put virtual screening using *AutoDock Vina* was used to rapidly screen the virtual library ligand set against 61 protease models built by Onywera et al. (2012). CheS-Mapper computed a principle component analysis of the compounds based on 13 selected chemical descriptors. The compounds were plotted against the principle component analysis within a 3 dimensional chemical space in order to inspect the diversity of the virtual library. The physicochemical properties and binding affinities were used to identify the top 3 performing ligands. *ACPYPE* was used to inspect the constitutional properties and eliminated virtual compounds that possessed open valences. Chromene based ligand 805 and ligand 6610 were selected as the lead candidates from the high-throughput virtual screening procedure we employed. Molecular dynamic simulations of the lead candidates performed for 5 ns allowed the stability of the ligand protein complexes with protease model 305152. The free energy of binding of the leads with protease model 305152 was computed over the first 50 ps of simulation using the molecular mechanics Poisson-Boltzmann method. Analysis structural features and energy profiles from molecular dynamic simulations of the protein–ligand complexes indicated that although ligand 805 had a weaker binding affinity in terms of docking, it outperformed ligand 6610 in terms of complex stability and free energy of binding. Medicinal chemistry approaches will be used to optimize the lead candidates before their analogues will be synthesized and assayed for *in vivo* protease activity.

Table of Contents

Section	Page No.
ABSTRACT	ii
TABLE OF CONTENTS	iii
LIST OF FIGURES	vi
LIST OF TABLES	ix
LIST OF ABBREVIATIONS	x
ACKNOWLEDGEMENTS	xiv
Chapter 1: Literature Review	1
<i>1.1 OVERVIEW OF HIV-1</i>	1
1.1.1 Origins and Epidemiology	1
1.1.2 Life-cycle	4
1.1.3 Treatment strategies	6
<i>1.2 HIV-1 PROTEASE AS A DRUG TARGET</i>	8
1.2.1 Protease	8
1.2.2 Drug Resistance and Molecular Dynamics	11
1.2.3 Novel sub-type specific PI scaffold	12
<i>1.3 MORITA-BAYLIS-HILLMAN HETEROCYCLES</i>	14
<i>1.4 PROJECT MOTIVATION</i>	16

1.4.1 Knowledge gap	16
1.4.2 Problem statement	16
1.4.3 Objectives	17
Chapter 2: Construction of a Virtual Library based on the bis-Coumarin scaffold	18
<i>2.1 INTRODUCTION</i>	18
<i>2.2 METHODOLOGY</i>	19
<i>2.3 RESULTS AND DISCUSSION</i>	22
<i>2.4 CONCLUSION</i>	29
Chapter 3: High-throughput Virtual Screening	30
<i>3.1 INTRODUCTION</i>	30
<i>3.2 METHODOLOGY</i>	31
3.2.1 Database Screen	31
3.2.2 Exhaustive Screen	32
3.2.3 Pharmacokinetic and physicochemical screening	32
<i>3.3 RESULTS AND DISCUSSION</i>	33
<i>3.4 CONCLUSION</i>	48
Chapter 4: Molecular Dynamic Simulations	49
<i>4.1 INTRODUCTION</i>	49
<i>4.2 METHODOLOGY</i>	50

<i>4.3 RESULTS AND DISCUSSION</i>	53
<i>4.4 CONCLUSION</i>	63
Chapter 5: Conclusion and Summary	64
References	66
Appendix	75
<i>APPENDIX A - SCRIPTS</i>	75
<i>APPENDIX B – RAW DATA</i>	88

List of Figures

Label	Description	Page no.
<i>Figure 1-1:</i>	Global distribution of HIV-1 sub-types and circulating recombinant forms.	3
<i>Figure 1-2:</i>	HIV-1 life-cycle and key stages that are manipulated during treatment strategies (Abbas & Herbein, 2012).	5
<i>Figure 1-3</i>	Ribbon representation of mature protease model.	9
<i>Figure 1-4</i>	Proposed catalytic mechanism of the HIV-1 protease's acid-base catalyzed concerted proteolysis.	10
<i>Figure 1-5</i>	Proposed mechanism of inhibitor binding showing the influence of hydration on the interaction of the inhibitor with catalytic residues. (Brik & Wong, 2003).	11
<i>Figure 1-6</i>	Pre-exposure docking Energy maps of 1357 docking experiments of Homology Models (59) and ligand test sets (23).	13
<i>Figure 1-7</i>	Bis-coumarin based novel scaffold identified as having PI potential.	13
<i>Figure 1-8</i>	The Morita-Baylis-Hillman reaction and formation of the MBH adduct. EWG = electron withdrawing group.	14
<i>Figure 1-9</i>	The versatility of the MBH reaction and the opportunities of transformation of the MBH adducts.	15
<i>Figure 2-1</i>	Bis-coumarin scaffold with fragments for derivatization illustrated as highlighted portions.	19
<i>Figure 2-2</i>	Bis-coumarin analogue construction using recursive procedure.	20
<i>Figure 2-3</i>	Pseudo-code for the generation of a virtual library of compounds from linker, heterocycle and small substituent fragments by recursive sequential hydrogen substitution.	20
<i>Figure 2-4</i>	Molecular masses distribution of 17750 ligands present in virtual dataset.	22

<i>Figure 2-5</i>	Distribution of the Hydrogen bond donors/acceptors in the virtual library data set.	23
<i>Figure 2-6</i>	Sample virtual library (3512) displayed in 3D Chemical Space embedded according to CDK chemical descriptors.	24
<i>Figure 2-7</i>	Sample virtual library (3512) coupled with approved drugs (#1555) from the DrugBank database.	26
<i>Figure 2-8</i>	Sample snapshot of the Virtual library of <i>bis</i>-coumarin analogues generated by recursive methodology.	28
<i>Figure 3-1</i>	Population Distribution of the Binding Energies of the Virtual library against the Consensus C Homology model.	33
<i>Figure 3-2</i>	Predicted <i>XLogP</i> of compounds in virtual library sub-set, DrugBank dataset and the top 200 Consensus C binders.	34
<i>Figure 3-3</i>	Computed number of Hydrogen Bond acceptors of compounds in virtual library sub-set, Drugbank dataset and the top 200 Consensus C binders.	36
<i>Figure 3-4</i>	Pharmacokinetic profile plot of the 200 best performing ligands selected based on binding with the Consensus C homology model.	37
<i>Figure 3-5</i>	Population distribution showing the interactions of the top200 binders (Figure 3-1) across the protease model dataset.	38
<i>Figure 3-6</i>	Heatmap of the Top 20 performing ligands and their protease model interaction energies.	39
<i>Figure 3-7</i>	Statistical analysis of interaction energies for Top 11 ligands from the 200 binders' dataset across the model data set.	40
<i>Figure 3-8</i>	Binding energies of ligand 6610 with the protease model data set.	42
<i>Figure 3-9</i>	Predicted binding pose of protease model 5086 in complex with ligand 6610 within the active site.	43
<i>Figure 3-10</i>	Ligand receptor interactions of predicted binding poses of ligand 6610 within the 5086 and the 508612 model cavity.	44
<i>Figure 3-11</i>	Computed Atlas of Surface Topography of proteins (CASTp) plot for the protease models. <i>Onywera et al. (2012)</i>.	45

<i>Figure 3-12</i>	Ligplot analyses of ligand 6610 with expansion mutation model 3051 and contraction mutation model 5261.	46
<i>Figure 3-13</i>	Ligplot analyses of ligand 805 with expansion mutation model 3051 and contraction mutation model 5261.	47
<i>Figure 4-1</i>	Block diagram of the implementation GROMACS MD simulation protocol for protein-ligand interaction studies.	51
<i>Figure 4-2</i>	Analysis of GROMACS energy terms during the production dynamics, for ligand 6610 and ligand 805 within protease model 305152	53
<i>Figure 4-3</i>	RMSF of protein atoms (A) and ligand atoms (B)	54
<i>Figure 4-4</i>	Average ligand conformations coloured by isotropic displacement obtained from B-Factors.	55
<i>Figure 4-5</i>	RMSD plot of the protein behaviour and the ligand interactions during 5 ns of MD simulation.	56
<i>Figure 4-6</i>	Convergence of the radius of gyration observed from the protein ligand complexes during 5 ns of MD simulation.	57
<i>Figure 4-7</i>	Protein ligand hydrogen bonding network maintained within the protein ligand complexes during the 5 ns of simulation.	58
<i>Figure 4-8</i>	Ligplot diagrams of the binding modes obtained from Molecular Dynamics (average complex) and docking (start complex).	62
<i>Figure 5-1</i>	Ligands identified as potential lead candidates for future development as protease inhibitors.	65

List of Tables

Label	Description	Page no.
<i>Table 3-1</i>	Pharmacokinetic and constitutional analysis of top 3 lead candidates.	41
<i>Table 4-1</i>	Hydrogen bonding network during 5 ns of simulation for the ligand 805 and ligand 6610 protease complex	59
<i>Table 4-2</i>	Binding energies of ligand 805 and ligand 6610 protease complexes from docking and MD.	60

List of Abbreviations

Abbreviation	Expansion
3D	3 dimensional
ABC	Abacavir
ACPYPE	Antechamber python parser interface
ADME	Adsorption, distribution, metabolism and excretion
AIDS	Acquired immune deficiency syndrome
API	Application program interface
apol	Atomic polarizabilities
ART	Anti-retroviral therapy
Asp	Aspartic acid
ATSm	Moreau-Broto autocorrelation descriptor based on atomic weight
Ave.	Average
AZT	Zidovine
bpol	Bond polarisabilities
cART	Combination ART
CASTp	Computed atlas of surface topography
CCR	Chemokine receptor 5
CDK	The chemistry development kit
CPU	Central processing unit
CXCR4	Chemokine receptor 4
Da	Daltons
DABCO	1,4-diazabicyclooctane

DNA	Deoxyribonucleic acid
DRM	drug resistance mutations
DRV	darunavir
DSV	Discovery Studio Visualizer
EWG	Electron withdrawing group
FDA	Food and Drug Administration
GAFF	Generalized amber forcefield
GBD-17	Chemical universe database 17
Gly	Glycine
GROMACS	Groningen machine for chemical simulations
HIV	Human Immunodeficiency Virus
HIVDR	HIV Drug resistance
HTS	High through-put screen
HTVS	High through-put virtual screen
INSTI	Integrase strand transfer inhibitors
INT	Intermediate
Kcal mol⁻¹	kilo Calories per mole
MBH	Morita-Baylis-Hillman
MCS	Maximum common subgraph
MD	Molecular dynamics
MM2	Molecular mechanics forcefield
MM-PBSA	Molecular mechanics Poisson -Boltzmann surface area
nAromBond	Number of aromatic bonds of molecule
n_{ef}	Negativity factor
nHB_{Acc}	Number of hydrogen bond acceptors

nHBDon	Number of hydrogen bond donors
nm	Nanometer
NNRTI	Non-nucleoside reverse transcriptase inhibitors
NRTI	Nucleoside reverse transcriptase inhibitors
ns	Nano seconds
NVP	Nevirapine
PCA	Principle component analysis
PI	Protease inhibitors
PR	Protease
PR3	Phosphine
ps	Pico seconds
RAL	Raltegravir
rev	Regulator of expression of virion proteins
Rg	Radius of gyration
RMSD	Root mean squared deviation
RMSF	Root mean squared fluctuation
RNA	Ribonucleic acid
RT	Reverse transcriptase
RTV	Ritonavir
RU	Rhodes University
SMILES	Simplified molecular-input line entry system
tat	Trans-activator of transcription
Thr	Threonine
tRNA	Transfer RNA
TS	Transition state

USD	United States Dollars
vDNA	Viral DNA
vif	Viral infectivity factors
vpr	Viral protein r
vpu	Viral protein u
WEKA	Waikato environment for knowledge analysis
WHO	World health organisation
ZDV	Zidovine

Acknowledgements

The following individuals deserve specific mention and recognition for their support and encouragement of my achievements.

Dr K. A. Lobb,

No amount of words of appreciation will be sufficient to communicate my sincere gratitude of your influence over my development as a student and a scientist. I am humbled by the journey that I have walked with you and I am truly honoured to have been given the opportunity to be your student over the last couple of years. Without your vision and direction this project would not have been feasible. It is my hope that this work will contribute significantly to the Medicinal Chemistry and Bioinformatics research groups that you are an integral part of.

Thomas Musyoka,

Thank you so much for challenging my horizons and asking critical questions of my methodology. Thank you so much for assisting me to get GROMACS to run on the Yoda cluster. I appreciate your encouragement to keep going on, even when things seemed like they had hit a stand still.

David Brown,

Thank you for managing the clusters and being willing to work extra hours to help me find the right version of GROMACS that would run my jobs sufficiently to completion. Thanks for being available to answer the odd question that you knew you would have to repeat 5 minutes later. Your patience is an asset.

Prof. Ozlem Tastan-Bishop,

Thank you for pushing me and for challenging me not to give up. I appreciate the firm voice of reason and wisdom. Thank you for the opportunity to participate in this course and I wish you all the best with all that you do.

Last but not least,

To Mom and PK, thank you for always being available to talk and listen to my whining and complaining. Thanks for all the laughs and the tears the anger and the tears. They were a welcome distraction from the menacing code that never seemed to resolve itself. Thank you for being willing to share this journey with me.

Chapter 1: Literature Review

1.1. OVERVIEW OF HIV-1

1.1.1. Origins and Epidemiology

Human immunodeficiency virus (HIV) is described as the causative agent of an acquired immune deficiency syndrome (AIDS) (Sharp & Hahn, 2010). This arises as a result of an untreated HIV infection that leaves the victim susceptible to opportunistic infections and cancers caused by bacteria, viruses, fungi and parasites (Sepkowitz, 2001; Teni, 2014). If the HIV infection is left untreated these infections can prove to be lethal resulting in death due to the health complications associated with the compromised immunity (Fenner et al., 2013; Lawn, Török, & Wood, 2011). According to the World Health Organisation (WHO), approximately 1.4 – 1.9 million people are said to have died due to AIDS-related illnesses in 2012, of which 75% of these deaths occurred in sub – Saharan Africa which experienced a 50% decline in AIDS-related deaths since 2004 (Teni, 2014; WHO, 2014). The WHO attributed the drop in AIDS-related mortality to the increased availability of antiretroviral drugs and the reduction of infection rates since their peak in 1997.

Despite significant advancements in the development and distribution of anti-retroviral drugs together with social interventions aimed at reducing infection rates, a growing concern in the fight against AIDS is the emergence of drug resistant viral strains (Aitken et al., 2013). HIV drug resistance (HIVDR) has been attributed to HIV's high mutation rate which affects drug-specificity and limits the effectiveness of anti-HIV drugs on their drug targets (Kozal, 2009; A. M. Wensing et al., 2014). The existence of drug resistant variants within the infected individual can lead to eventual treatment failure in the presence of anti-HIV drugs that create a selection pressure on the virus (Cambiano et al., 2014). A multi-site study of drug resistance viruses within anti-retroviral naive adults estimated that of the 3.9 million infected people in sub – Saharan Africa that are said to be on anti-HIV treatment approximately 5.6 % prevalence of resistance was believed to persist prior to treatment (Hamers et al., 2011). Hamers' study focused on individuals in Kenya, Nigeria, South Africa, Uganda and Zimbabwe and showed that roll-out of antiretroviral therapy (ART) in Africa was the principal agent on the emergence of primary drug resistance within the population.

HIV, the causative viral agent of AIDS, was first identified in 1986 by Françoise Barré-Sinoussi and Luc Montagnier who were awarded the Nobel Prize in Physiology or Medicine in 2008 for their work (Lever & Berkhout, 2008). There are 2 known types of HIV virus, type 1 and type 2 that are known to infect humans (Hemelaar, 2012). It is thought that HIV-2, the first to be discovered, is less virulent than the more transmissible type 1 virus (De Cock, Jaffe, & Curran, 2012; Sharp & Hahn, 2010). Phylogenetic analyses suggest that this virus version originated in sooty mangabey, *Cercocebus atys*, a West African primate and thereafter crossed into human populations. Major transmission points gave rise to the HIV-2 A and B groups that are widespread in human populations compared to the other 6 HIV-2 lineage subtypes C, D, E, F, G and H (Damond et al., 2004). Type 1 strains appear to have emerged in human populations after crossover events with viruses that infect the *Pan troglodytes* chimpanzees (Gao et al., 1999; Sharp & Hahn, 2010). The majority of the prevalent HIV-1 strains belong to the M group while the N and O subgroups have remained restricted within Cameroon and west-central Africa respectively, with low levels of exponential growth since their transmission (Lemey et al., 2004; Louis, Aniana, Weber, & Sayer, 2011).

The independent transmission event that gave rise to the M subgroup of HIV type 1 has been shown to account for more than 90% of HIV/AIDS cases (Kumar & Herbein, 2014). Genetic variability and rapid evolution due to high mutation and recombination rates of HIV-1 have resulted in a rich diversity of genetic variability. Within the major M group there are 9 genetically distinct clades of HIV-1 namely A, B, C, D, F, G, H, J and K, which share up to 92% sequence similarity and up to 42% sequence variation between themselves depending on the subtype genome (Hemelaar, 2012; Tatem, Hemelaar, Gray, & Salemi, 2012). Recombination of different viral subgroups gives rise to temporary hybrid viruses described as “circulating recombinant forms” (Hemelaar, 2012; Korber et al., 2001). There were at least 48 forms identified as circulating in the global human population in 2011 and an increase of the participation of resistant strains in recombination will contribute significantly in the persistence of these forms increasing viral diversity (Tebit & Arts, 2011).

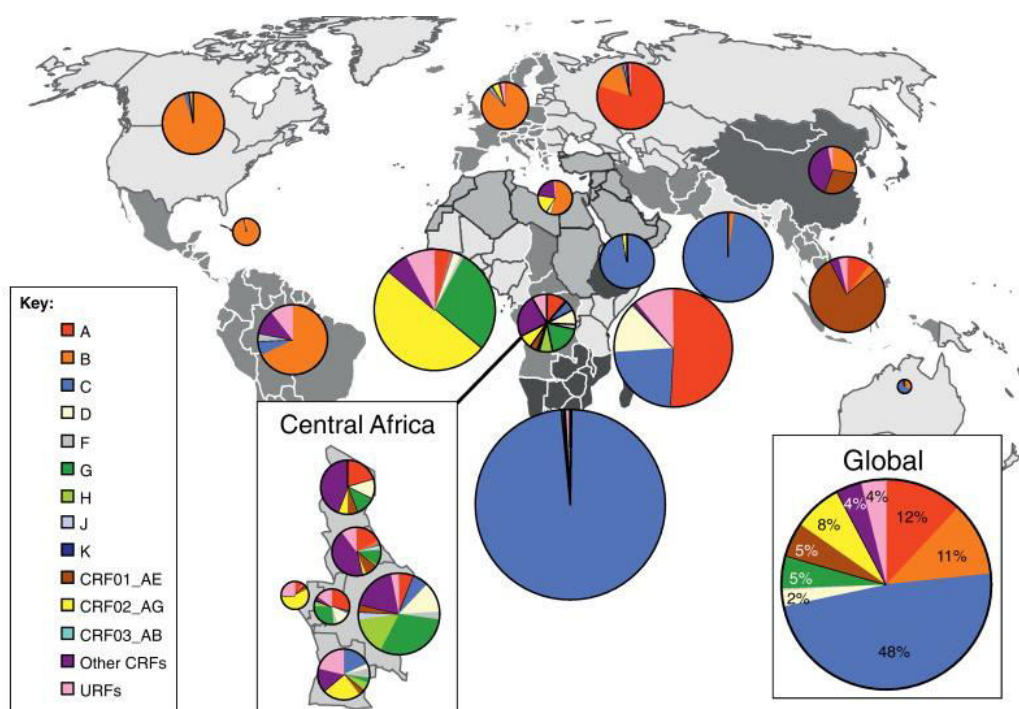


Figure 1-1: Global distribution of HIV-1 sub-types and circulating recombinant forms. Pie-charts of sub-type distributions within that population between 2004 - 2007 are superimposed over the shaded region of map. The colours indicate the dominant subtypes (Hemelaar, 2012).

The global distribution map produced by Hemelaar using data obtained from the WHO between 2004 and 2007 shows that HIV-1 subtype C accounted for 48% of all global infections and was shown to predominate in sub-Saharan Africa, East Africa, Australia and India, Figure 1-1. Sub-type B accounted for only 11% of global infections within this time period and was dominant in North, Central and South America; Western Europe and isolated parts of North Africa. Central Africa showed the highest subtype diversity with rare subtypes such as G and H observed in large proportions within the populations. There are significant differences observed between different regions in Africa compared to the differences observed between the whole of the Americas, Asia or sub-Saharan Africa. It can be thought that this diversity can be attributed to the reduced population mixing in these regions due to cultural separations, state boundaries, language barriers and transport infrastructure (Tatem et al., 2012; Tebit & Arts, 2011). Increased travel and mobility will continue to contribute significantly towards the diversification of the HIV epidemic and the spread of drug resistance by increasing the genomic variability (Tatem et al., 2012).

The HIV-1 virus has a high mutability rate due to the lack of proof-reading mechanisms in its viral RNA reverse transcriptase machinery (Kumar & Herbein, 2014). Its mutability is enhanced by a rapid viral turn-over rate ensuring that enough viral copies are present that contain non-synonymous mutations that are functional (Das & Arnold, 2013; Korber et al., 2001). The viral genome is described as containing 9 gene encoding regions (*gag*, *pol*, *env*, *vif*, *vpr*, *vpu*, *rev*, *tat* and *nef*) where each of these regions has different genetic diversity because they vary at different rates (Kosakovsky Pond & Smith, 2009). For example the gene coding for gp120 varies up to 20 times more than the *gag pol* genes (Korber et al., 2001). These genes encode for 15 proteins (matrix, capsid proteins, nucleocapsid proteins, p6, proteases, reverse transcriptases, integrases, gp120s, gp41s, virion infectivity factors (*vif*), viral protein r (*vpr*), viral protein u (*vpu*), trans-activator of transcription (*tat*), regulator of expression of virion proteins (*rev*), and negative factor (*nef*)) that are involved in different stages of the viral life cycle (Sainski, Cummins, & Badley, 2014).

1.1.2. Life-cycle

The HIV-1 viral life-cycle is divided into 2 main phases, an early and a late phase (Ott & Verdin, 2013). The early phase is associated with viral entry and integration while the late phase consists of the processes that occur after integration resulting in the release of mature infectious viral progeny (Murray, Kelleher, & Cooper, 2011). The virus can undergo 2 types of dormancy or latency where replication is absent (Ott & Verdin, 2013; Williams & Greene, 2007). They can undergo the less clinically significant pre-integration latency which results in incomplete reverse transcription of viral RNA. Post-integration HIV latency has been shown to account for the long-term persistence of HIV in actively treated patients (Abbas & Herbein, 2012). Its exact mechanism is not well understood although it is clear that this latency is associated with resting cells (Ott & Verdin, 2013).

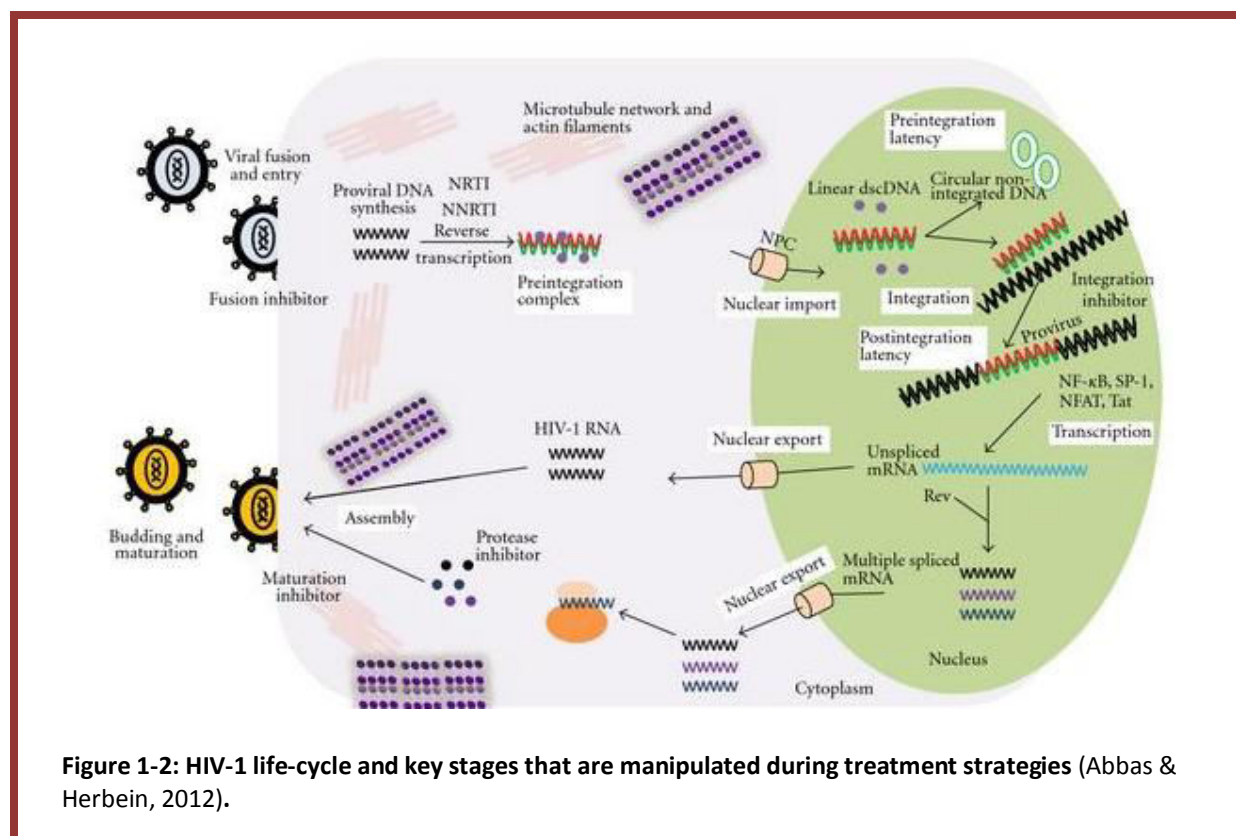


Figure 1-2: HIV-1 life-cycle and key stages that are manipulated during treatment strategies (Abbas & Herbein, 2012).

The main steps necessary for a mature virion to infect and be virulent as illustrated in Figure 1-2 are:

- i) attachment to host CD4 and CXCR4 chemokine receptor type 4 (CXCR4) or CCR5 chemokine receptor type 5 (CCR5) receptors by viral gp120 proteins,
- ii) injection of HIV-1 viral core (single strand positive sense RNA, tRNA primers, viral protease, reverse transcriptase, and integrase) into host cell cytoplasm,
- iii) reverse transcription of viral RNA forming viral DNA,
- iv) viral DNA within the pre-integration complex is translocated into the host nucleus through microtubule and dynein mediation before its integration into the translatable portions of dormant host cell's DNA,
- v) the completion of transcription and translation of viral genes to proteins under the direction of host transcription machinery, viral transactivator (tat) and regulatory proteins (rev) after activation of host cells,
- vi) formation of large immature HIV-1 precursor proteins,
- vii) maturation through proteolytic processing of precursor proteins and budding of mature infectious viral particles (Abbas & Herbein, 2012).

1.1.3. Treatment strategies

The majority of treatment strategies for HIV infections aim to intercept the viral life-cycle and thus reduce the number of infectious viral particles within the infected individual (Hemelaar, 2012; Palmer, Alaeus, Albert, & Cox, 1998). The US National Institute of Allergy and Infectious Diseases "*Reference guide for prescription HIV-1 medications*" divides most HIV-1 drugs available into 7 categories depending on their target and mechanism of action. These categories are nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), protease inhibitors (PIs), fusion inhibitors, integrase strand transfer inhibitors (INSTIs), CCR5 antagonists and combination antiretrovirals (NIH, 2014).

Reverse transcriptase (RT) is the primary target for the majority of antiviral drugs and is responsible for most antiviral drug resistance (V. Johnson et al., 2013). This is because RT lacks an efficient proof-reading mechanism which endows it with an error-rate of 0.2-2 mutations per genome per cycle (Kumar & Herbein, 2014). NRTIs such as Zidovine (AZT/ZDV) and Abacavir (ABC) inhibit reverse transcription by the introduction of a modified analogous nucleoside prematurely terminating its DNA synthesis (Das & Arnold, 2013). NRTI's have been associated with increased levels of mitochondrial toxicity resulting in neuromuscular and cardiac effects such as myopathy, peripheral neuropathy and hypertrophic cardiomyopathy (Gerschenson et al., 2009; Medina, Tsai, Hsiung, & Cheng, 1994). Evidence supporting this was observed in the incorporation of these analogues as substrates for mitochondrial DNA polymerase giving rise to mitochondrial dysfunctions (A. A. Johnson et al., 2001). NNRTI's such as nevirapine (NVP) and delavirdine, interact directly with the viral reverse transcriptase inhibiting its activity by binding to allosteric sites averting the toxicities associated with nucleoside analogue NRTIs (Das & Arnold, 2013). Despite improvements in mitochondrial toxicity, NNRTI's have been associated with liver toxicity coupled with depression and alcohol abuse associated with neuro-psychiatric side-effects (Al-Khindi, Zakzanis, & van Gorp, 2011; Nakimuli-Mpungu et al., 2012; Usach, Melis, & Peris, 2013).

First generation PIs such as ritonavir (RTV) and second generation PIs such as darunavir (DRV) prevent viral maturation by inhibiting the cleavage of precursor proteins with protease (Sainski et al., 2014; Walmsley, 2007). Inhibitor activity is achieved predominantly through competitive peptidomimicry of the protease substrate (Wensing, van Maarseveen, & Nijhuis, 2010). Side-effects associated with PI incorporating ART treatment regimens have

been shown to relate to drug metabolism complications due to secondary interactions with cytochrome P450 enzymes (Walmsley, 2007). Fusion inhibitors such as enfuvirtide (T-20) and maraviroc are designed to prevent HIV-1 binding, fusion, and viral entry into host cell CD4 cells (Lieberman-Blum, Fung, & Bandres, 2008).

Maraviroc is known as a CCR5 antagonist that binds to this receptor preventing strong interactions with viral gp120 protein which is necessary to initiate conformational changes required for viral entry (Kumar & Herbein, 2014). In rare cases that resistance arises, the virus utilises a non-CCR5 host receptor to enter the host cell. Side-effects occur if maraviroc interacts with CXCR4 receptors on immune cells of host while others associated with hepatotoxicity have been reported (Kumar & Herbein, 2014; Lieberman-Blum et al., 2008). Enfuvirtide inhibits covalent hairpin formation which catalyzes the fusion of the virus with the host membrane. Resistance is attributed to the rapid mutability of the viral gp41 protein (Bai et al., 2013; Murray et al., 2011).

INSTIs such as raltegravir (RAL) are designed to target the HIV-1 integrase enzyme and prevent the integration of the viral reverse transcribed genetic material into the host genome (Kumar & Herbein, 2014; Thompson et al., 2010). They are described as interfacial inhibitors as they interact with the vDNA, integrase catalytic magnesium cations and residues within the integrase activity site in order to stabilize an intermediate which inhibits completion of the DNA integration (Métifiot, Marchand, & Pommier, 2013). Single point mutations within viral integrases have been shown to be sufficient to introduce resistance in some integrase inhibitors, thus combinatorial ART is essential to maintain their efficacy (Peat, Dolezal, Newman, Mobley, & Deadman, 2014).

1.2. HIV-1 PROTEASE AS A DRUG TARGET

1.2.1. Protease

HIV-1 protease is a viral protein that is produced in the late stage of an HIV-1 viral infection (Sainski et al., 2014). It is described as being an aspartyl mediated protease responsible for the cleavage and processing of viral proteins critical for maturation (Kumar & Herbein, 2014). As a consequence of its absence after the early stage of infection the protease itself is formed through autoprocessing, because HIV-1 protease is translated as part of the gag-pol polyprotein, a protease substrate (Louis et al., 2011). Because polyprotein cleavage is such a vital part of the HIV-1 viral life-cycle and is important for the production of infectious viral particles, PIs are regularly used in ARV therapy and detailed knowledge of its structure aides in the design of novel protease inhibitors (A. Wensing et al., 2010). Due to the highly variable nature of the HIV-1 virus and the rapid mutability of the gag-pol genome, drug resistance rapidly arises as reflected in its plasticity and large number of protease polymorphisms (Louis et al., 2011; Rhee et al., 2006).

HIV-1 protease has been identified as a non-specific protease which recognises its substrates based on their shape and their ability to optimize their conformation and contacts with catalytic residues in its active site through diffusion and orbital steering (C. Chang, Trylska, Tozzini, & McCammon, 2007). Protease identifies viral poly-protein substrates yielding MA, CA, NC and p6 products from the *gag* substrate and PR, RT and IN enzymes from the *gag-pol* substrate (Trylska, Tozzini, Chang, & McCammon, 2007; A. Wensing et al., 2010). It has been characterized as existing in a homodimer conformation with C_2 symmetry enclosing a conserved Asp-Thr-Gly triad within its active site, Figure 1-3 (C. Chang et al., 2007). Coarse-grained MD simulations of HIV-1 protease showed that the flap regions that cover the active site are dynamic and their opening can allow the entry of substrates into the active site. Entry of the ligands into active site was shown to be dependent on an interaction with the protein-substrate interaction which was shown to make fluctuations of flaps more frequent and more stable (Trylska et al., 2007). It is known that sub-type C PR possesses 8 polymorphisms which are known to indirectly affect flap mobility by increasing its flexibility compared to its subtype B homologue (Coman et al., 2008).

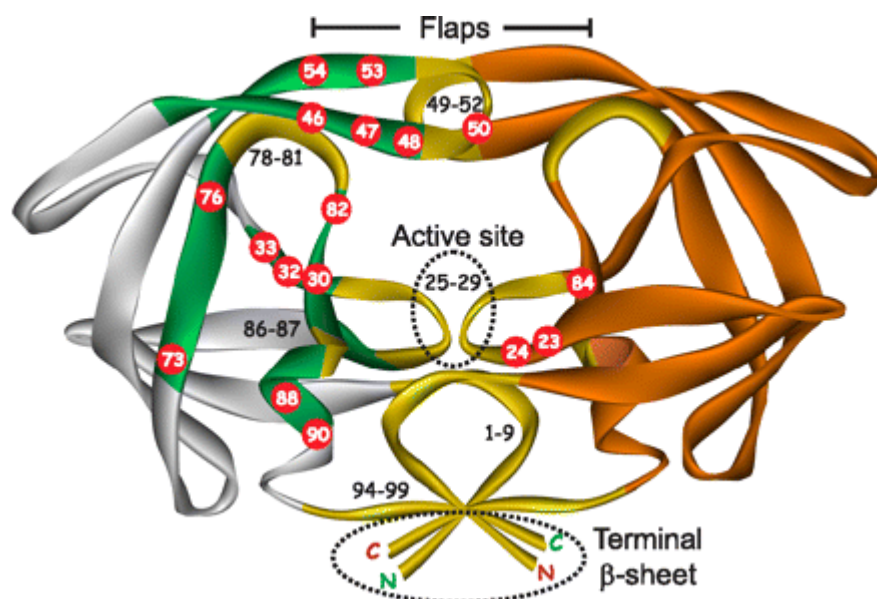


Figure 1-3: Ribbon representation of mature protease model showing the location of highly conserved regions under drug pressure (gold and black lettering)(25), regions of natural variability in PR among the four groups M, N,O, and P (gray), and naturally conserved regions where major DRMs are selected under drug pressure (green). Numbered red circles indicate the positions of major DRMs, as defined in the Stanford database. (Louis et al., 2011).

Although the HIV-1 PR mechanism has not been unambiguously determined, growing consensus supports a mechanism of action that manipulates the protonation state of the ASP residues within the active site (Kipp, Hirschi, Wakata, Goldstein, & Schramm, 2012). The protease catalytic mechanism is generally accepted as being a concerted acid-base dependant electrophilic attack mediated proteolysis generating 2 intermediates and 3 transition state species, Figure 1-4 (Brik & Wong, 2003; Shen et al., 2012).

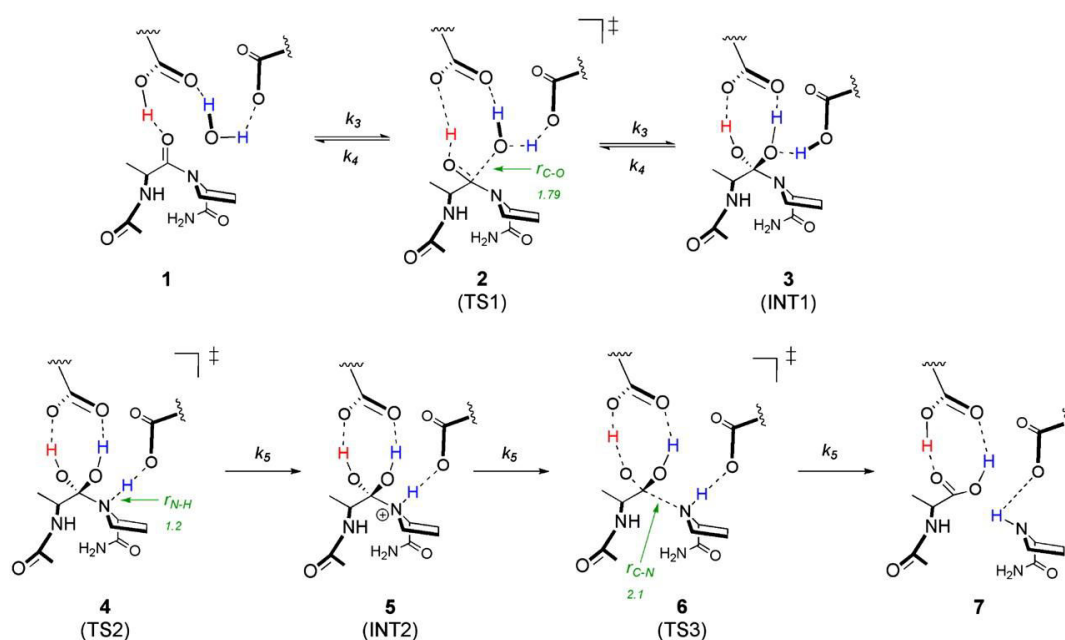


Figure 1-4: Proposed catalytic mechanism of the HIV-1 protease's acid-base catalyzed concerted proteolysis. Structures are labelled as follows 1. PR substrate complex, 2. Transition state (TS) of water attack, 3. Di-ol intermediate (INT), 4. protonation TS, 5. Protonated INT, 6. Cleavage of substrate bond, 7. PR product complex. Green arrows and text refers to bonds and bond lengths. (Kipp et al., 2012).

The protease active site and mechanism has been manipulated in inhibitor design and it is common for a water molecule to play an important role in aspartyl protease inhibitor binding (Babine & Bender, 1997; Schramm, 2013). Saquinavir and indinavir contain a hydroxyethylamine and hydroxyethylene transition state isostere respectively and they are examples of mechanism based inhibitors. When their inhibitor carbonyls within the vicinity of the ASP residues are hydrated they create stable hydrogen bonding networks, analogous to the TS1 (Figure 1-4) transition-state, due to the enhanced electrophilic properties, Figure 1-5 (Brik & Wong, 2003; Kipp et al., 2012).

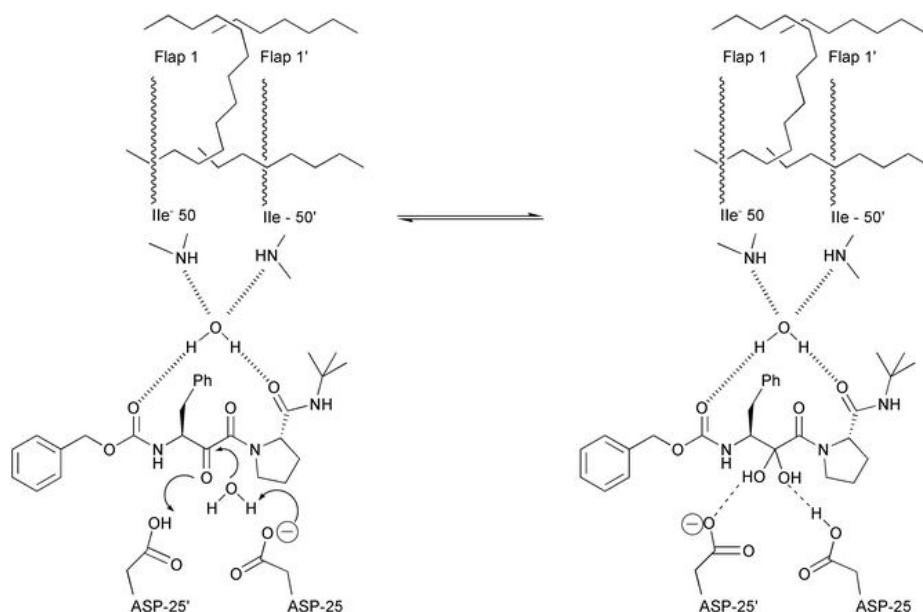


Figure 1-5: Proposed mechanism of inhibitor binding showing the influence of hydration on the interaction of the inhibitor with catalytic residues. (Brik & Wong, 2003).

1.2.2. Drug Resistance and Molecular Dynamics

The inevitability of resistance results in the persistence of a PR polymorphism that is resistant to PI therapy despite incorporating ART in a combinatorial approach reducing the likelihood of survival of this polymorphism (Hemelaar, Gouws, Ghys, & Osmanov, 2006; Shafer, 2006). Detailed structural insights into the topology of the mutant polymorphism allows for the development of novel PI which inevitably undergo a similar phasing out due to resistance (Wensing et al., 2014).

Despite the changes that occur within the PR affecting active site topologies, most PR proteolytic functionality is conserved because of selection pressures (Cambiano et al., 2014; Kipp et al., 2012). In the development of novel inhibitors it is important to understand the mechanism driving resistance of a persistent polymorphism. When an I84V mutation in sub-type B was investigated it was seen from electrostatic potential maps that although the substrate had an identical transition state in wild-type as the mutant, resistance to a previously potent TS mimic was due to changes elsewhere in the protein (Kipp et al., 2012). When a T80V mutation induced saquinavir resistance the DR mutant showed decreased proteolytic activity although the virus maintained infectivity. In silico molecular dynamic simulations suggested that the changes affected the internal dynamics of flaps resulting in

low drug-protein interactions (Foulkes et al., 2006). Coupling NMR relaxation observations with MD simulations allowed MD atomic detail to be validated with experimental observations of overall protein dynamics of DR polymorphs that had mutations in the primary autoproteolysis sites (Cai, Yilmaz, Myint, Ishima, & Schiffer, 2012).

B sub-type developed ART are known to achieve sufficient viral suppression in non-subtype B circulating HIV-1 strains (Kosakovsky Pond & Smith, 2009). However a study in 1998 showed that non-subtype B infected individuals that had previously been exposed to a potent ARV, ZDV, developed resistance to this drug while no resistance was observed with patients that had the B-subtype (Palmer et al., 1998). Based on inhibition studies it has been shown that certain active site mutations are more influential in the development of resistance in C subtypes than in B subtypes (Mosebi, Morris, Dirr, & Sayed, 2008). A follow-up study using simulated dynamics of these polymorphisms showed that the C-SA subtypes had increased flap movements compared to the B subtypes helping to account for the different responses to DR mutations (Ahmed et al., 2013).

Due to the rapid mutability of the HIV virus and the ability of B sub-type developed PIs inducing DR in non-B subtypes there is thus a growing need for the development of novel PIs that are subtype specific.

1.2.3. Novel sub-type specific PI scaffolds

HIV-1 drug design work at Rhodes University within the synthetic medicinal chemistry research group and the computer aided drug discovery group work aims to generate novel lead compounds targeted toward non-B subtype PR strains. Onywera, 2012, undertook an “*in silico*” investigation into the influence of non-synonymous sequence mutations on the architecture of HIV-1 clade C protease receptor sites”. The binding energies of compounds synthesized in the medicinal chemistry research group with 58 subtype PR homology models were evaluated by using docking.

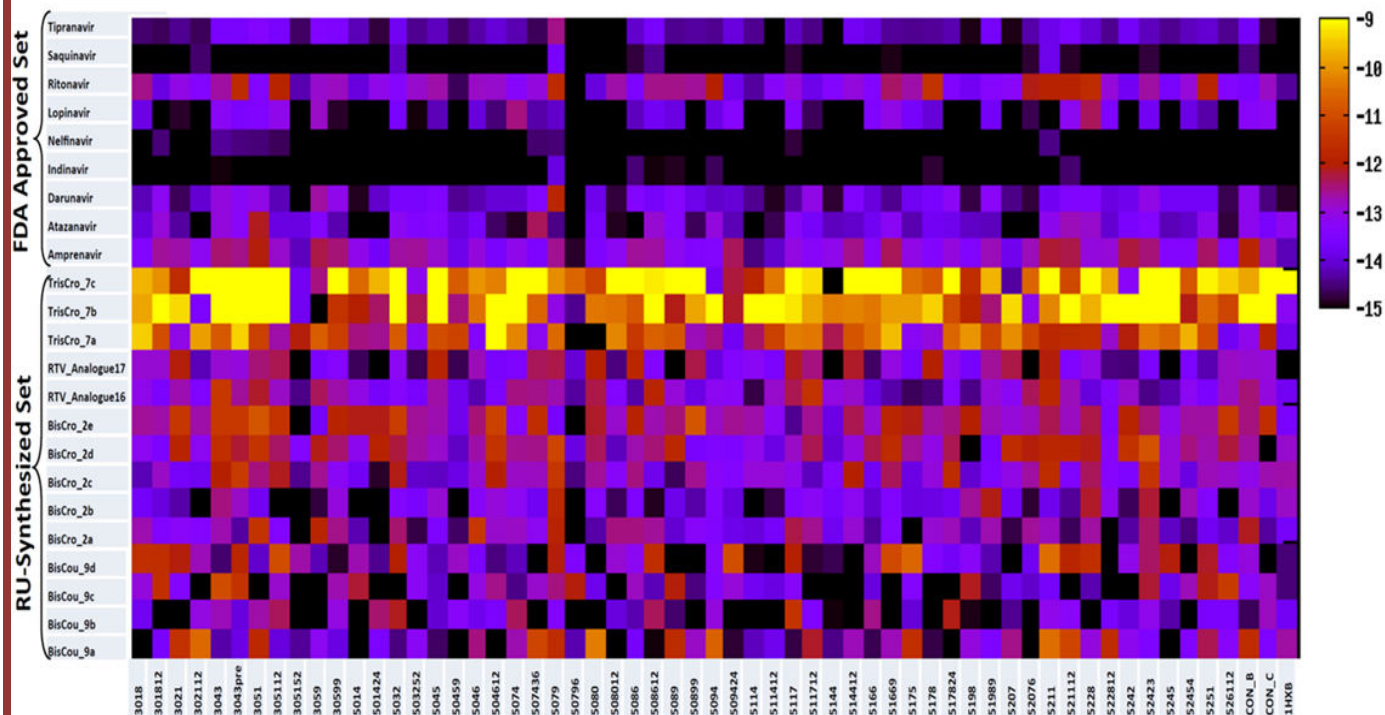


Figure 1-6: Pre-exposure docking Energy maps of 1357 docking experiments of Homology Models (59) and ligand test sets (23). Binding energy is represented by heatmap colour coded in legend. Strong binding (dark, black) and Weak binding (light, Yellow). Models are shown on x-axis while y-axis shows inhibitors. 6 digit values identify the models built from sequences of patients after their exposure to the drug with the 5th and 6th digits describing the duration of exposure to ART treatment. 4 digit values describe the model generated from the sequence obtained from the patient before ART treatment. (Onywera et al., 2012).

Onywera identified the bis-coumarin moiety as a novel scaffold for further investigation based on its high affinity for the PR active sites across the 59 models, Figure 1-6 (Onywera, Lobb, & Tastan Bishop, 2012). The coumarin fragments that make up this scaffold are examples of benzannulated heterocycles that are accessible by applications of the Morita-Baylis-Hillman reactions extensively studied in the RU medicinal chemistry research group.

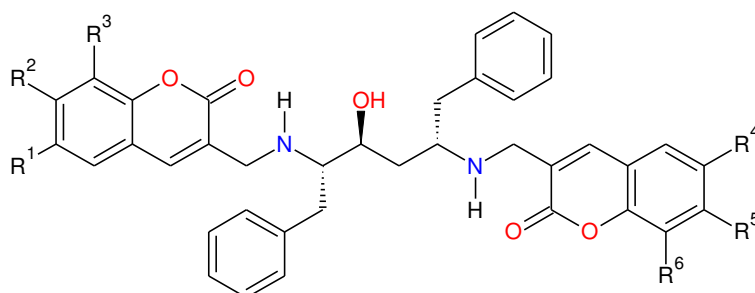
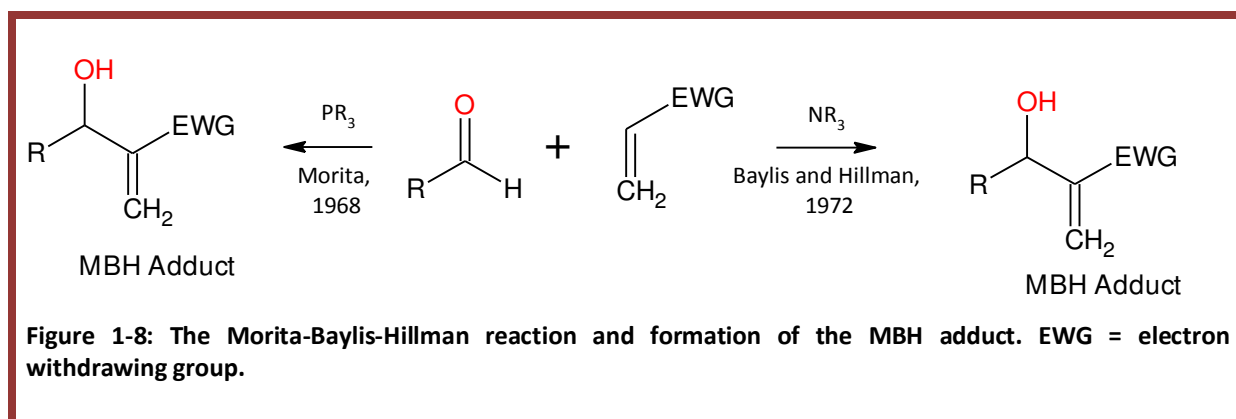


Figure 1-7: Bis-coumarin based novel scaffold identified as having PI potential.

1.3. MORITA-BAYLIS-HILLMAN HETEROCYCLES

The MBH reaction is a versatile condensation reaction that involves the coupling of an sp^2 hybridised carbon electrophilic centre, such as carbonyls and aldimines, with an electron deficient centre such as an alkene Figure 1-8. In 1968 Morita catalysed this reaction with a tertiary phosphine (PR_3) catalyst while Baylis and Hillman achieved a similar condensation using 1,4-diazabicyclooctane, DABCO (NR_3) catalyst (Baylis & Hillman, 1972; Morita, Suzuki, & Hirose, 1968).



The MBH reaction is an important carbon-carbon bond-forming reaction because it has a high atom efficiency, requires mild reaction conditions, avoids heavy-metal pollution and has flexible and multi-functional MBH adducts (Zhao, Wei, & Shi, 2011). It suffers however from having poor reaction rates restricting its practicability in general synthetic routes (Wei & Shi, 2013). Despite these setbacks the versatility of the MBH reaction affords various opportunities for transformation of the adduct giving rise to the construction of numerous benzannulated heterocyclic systems with various applications (Nyoni, Lobb, & Kaye, 2013; Peng, Huang, Jiang, Cui, & Chen, 2011). Its practicality can be increased by improving the reaction rates by carefully considering the nature of the substrate, the catalyst used, the reaction temperature, pressure and solvent (Zhao et al., 2011). Coumarins, quinolones and chromenes are examples of some of the benzannulated heterocycles accessed by the MBH reaction in the Rhodes University medicinal chemistry research group, Figure 1-9.

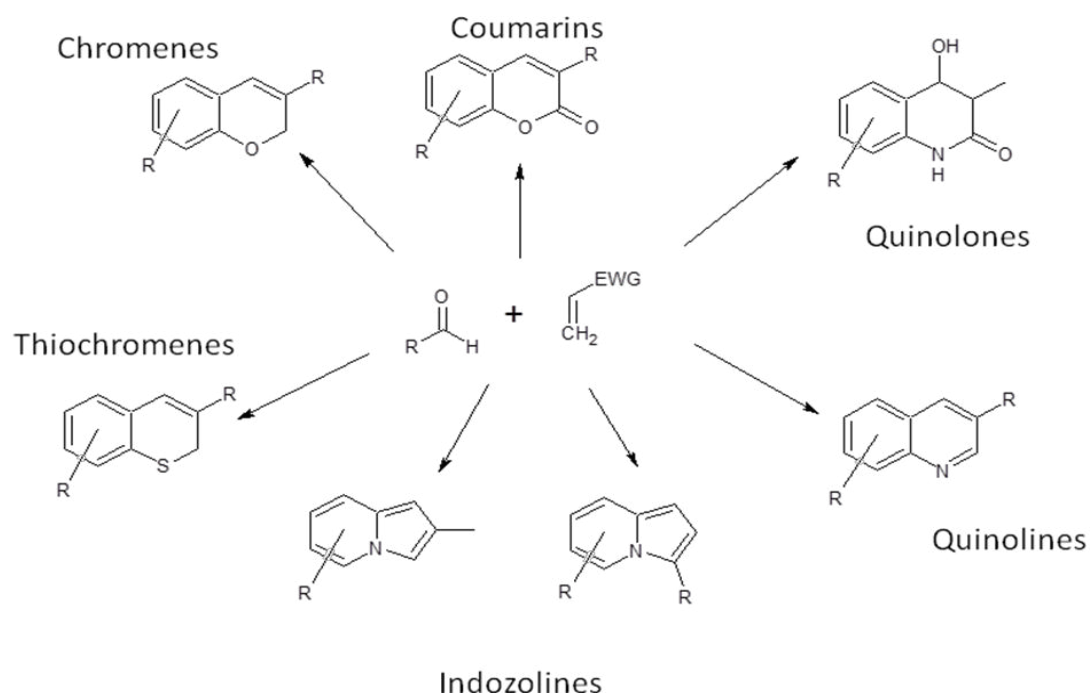


Figure 1-9: The versatility of the MBH reaction and the opportunities of transformation of the MBH adducts.

1.4. PROJECT MOTIVATION

1.4.1. Knowledge gap

Despite the topological active site conservation, mechanistic similarities and drug susceptibilities C subtype and B subtype HIV-1 viral strains, they display subtle differences in the drug responses and development of resistance in their proteases (Ahmed et al., 2013; Kosakovsky Pond & Smith, 2009; Mosebi et al., 2008; Palmer et al., 1998). Because of the inevitability of the resurgence of drug resistance there is a growing need for the development and incorporation of novel subtype specific PIs in cART (Tatem et al., 2012).

The time it takes to develop a new drug has on average been shown to take between 10 to 15 years, with the average cost of development being higher than \$1.2 billion USD. These facts are startling if we consider that only 2 out of 10 marketed drugs return revenues that match or exceed Research and Development costs (PhRMA, 2013). Computer aided drug discovery aims to speed up the screening process, reduce costs associated with lead identification and optimisation by employing targeted de novo ligand design with improved specificity coupled with *in silico* virtual screening and efficient structure based techniques (Nicolaou, 2014; Zhang, 2011).

Onywera, 2012, investigated the architecture of HIV-1 clade C protease receptor sites of PR sequences obtained from infants who were experiencing treatment failure to FDA approved PIs. By employing *in silico* approaches to investigate ligand binding PR model susceptibility he examined the influence of key protease residues that were useful for targeting resilient HIV-1 assemblages in cART and identified a novel scaffold (Onywera et al., 2012).

1.4.2. Problem statement

Identifying a novel scaffold allows the identification of a ligand backbone that interacts with the target active site architecture in a manner that optimizes binding reducing its activity (Nicolaou, 2014; Sinko, Lindert, & Mccammon, 2013). This scaffold undergoes various optimizations in order to identify a satisfactory hit with the lead potential showing improved specificity and binding consistent with quantitative structure activity relationships (G. Sliwoski, Kothiwale, Meiler, & Lowe, 2014).

Our research is concerned with whether we can efficiently, accurately and precisely employ *in silico* approaches in order to optimize the scaffold into a hit or lead candidate that can be

synthesised using application of the Morita-Baylis-Hillman reaction. Although we are limited by time our biggest limitation in this regard will be our computational resources. It is our hope that our modest 94 processor computer cluster will allow for an appropriate pilot test of our methodology and design.

1.4.3. Objectives

The aims and objectives of the research problem which will allow us to conduct a pilot study to access protease subtype specific lead candidate are outlined below:

1. Generate a synthetic library

Based on the bis-coumarin scaffold that utilises MBH benzannulated heterocycles we aim to synthesize over 5000 unique compounds that allow us to exhaustively search a narrow chemical space.

2. Perform high-throughput virtual screening

In order to identify a series of hits from the synthetic library that show optimised binding interactions across all the 61 Onywera protease models, rapid and efficient *in silico* screening methods will be utilised.

3. Perform molecular-dynamics simulations

In order to investigate receptor-ligand solution interactions of the lead candidates across the Onywera models, protease-hit molecular dynamics will be simulated in explicit water solvent.

Chapter 2: Construction of a Virtual Library based on the *bis*-Coumarin scaffold

2.1. INTRODUCTION

The *bis*-Coumarin ligand identified by Onywera et al. (2012), was chosen as a novel scaffold on which to build a virtual library of related analogues that could be searched in order to identify a plausible lead candidate for chemical synthesis.

Since the 1960s various cheminformatic approaches have been utilised in order to estimate the stoichiometric combinations of electrons and atomic nuclei in all possible topological isomers, defined as the chemical space universe (Lederberg, 1965; Reymond & Awale, 2012). This chemical space represents the accessible landscape of small molecules for lead identification in drug-discovery applications. An estimate of organic molecules containing C, N, O, S and the halogens up to 17 atoms yielded an estimate of over 160 billion drug-like molecules available in the chemical universe (Ruddigkeit, Van Deursen, Blum, & Reymond, 2012). As of January 2015, the Chemical Abstracts Service, who build and maintain a collection of molecular substances and their publicly disclosed substance information, had registered more than 91 million unique organic and inorganic chemical substances in its registry (ACS, 2015). These represent less than 0.1% of the searchable chemical space available from the GBD-17 drug-like molecules (Ruddigkeit et al., 2012). There is a growing trend in drug-discovery projects to explore the chemical space made available by chemical universe databases in order to identify novel scaffolds that possess improved selectivity and ADMET profiles (Reymond & Awale, 2012; Gregory Sliwoski, Kothiwale, Meiler, & Lowe, 2014).

By building over 5000 MBH accessible benzannulated systems it was our hope that we would exhaustively search the chemical space available to analogues of the *bis*-Coumarin moiety for applications as novel PIs lead candidates.

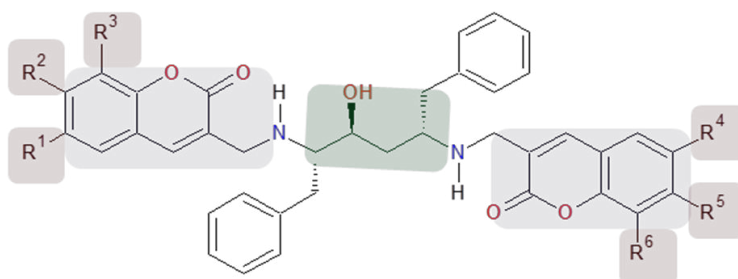


Figure 2-1: *Bis*-coumarin scaffold with fragments for derivatization illustrated as highlighted portions. Linker region (Green), the heterocyclic portion (Purple) and the small substituents (Red).

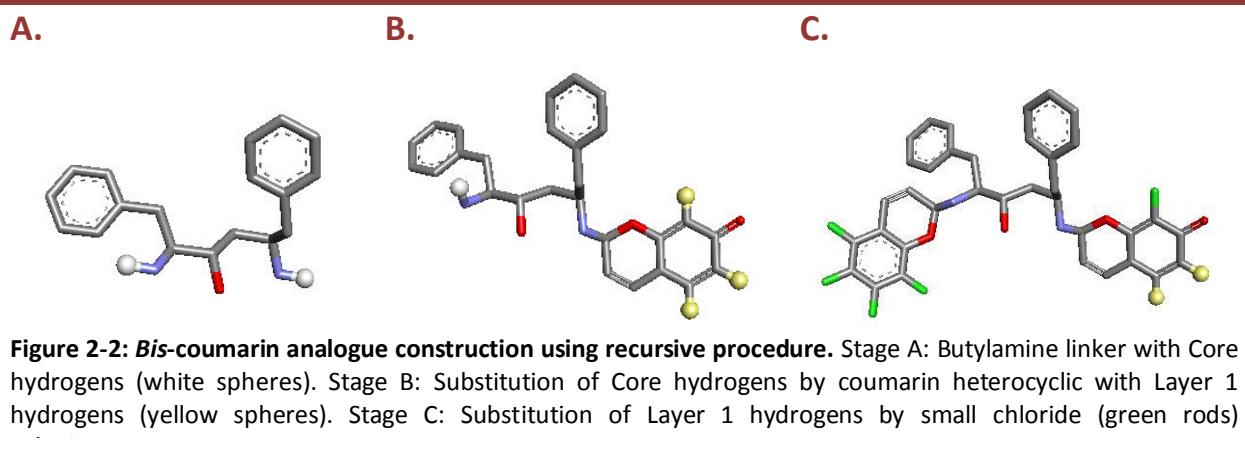
Analogues of the *bis*-Coumarin moiety were to be built by using derivatives of the linker regions, heterocyclic portions and the small substituents. In assembly the linker regions would then either be butyldiamine, pentyldiamine or a hexyldiamine while the heterocyclic portion chromenes, thiochromens, coumarins, quinolines quinolones and indozoline MBH accessible heterocycles. The small substituents were to be CH₃, Cl, F, H, NH₂, and OH.

2.2. METHODOLOGY

Open Babel is a cheminformatics software system that was initially developed to facilitate interconversion of multiple chemical file formats but has become a useful program library for organic chemistry, drug discovery, materials science and computational chemistry (O'Boyle et al., 2011).

A C++ program that incorporates Open Babel commands was written in order to construct *in silico* analogues. The different portions were saved as fragments with the hydrogens acting as level identifiers. Each fragment was prepared using Discovery Studio Visualizer and saved in .xyz format. For each fragment all the hydrogen atoms were deleted after the aromaticity is corrected and only the hydrogens to be substituted were drawn before the fragment is saved.

For each molecule a single procedure was called as a recursive procedure until all core hydrogen atoms were replaced. The layer 1 hydrogen atoms present on the core substituents were replaced in a similar manner by the small substituents, Figure 2-2.



This procedure was implemented in a discrete-time Markov chain process that looped through all the linkers, heterocycles and substituents. All possible combinations of the fragments were prepared and the resulting molecules constituted the members of the virtual library.

One issue that had to be resolved was the placement of the fragments at levels 1, 2 and 3. In order to do this the bond vector representing the H to be replaced was established,

```

procedure elaborate(moleculetype molecule)
  if(corehydrogens)
    foreach corehydrogen
      foreach heterocycle (chromene,thiochromene,coumarin,...)
        newmolecule = molecule + heterocycle
        elaborate (newmolecule)
  else if (layer1hydrogen)
    foreach layer1hydrogen
      foreach smallsubstituent
        (hydroxide,chloride,methide,...)
        newmolecule = molecule + smallsubstituent
        elaborate (newmolecule)
  else
    # molecule is complete
    savemolecule()

if(linker)
  foreach linker (butyldiamine, pentyldiamine, ...)
    molecule = linker
    elaborate (molecule)
  
```

Figure 2-3: Pseudo-code for the generation of a virtual library of compounds from linker, heterocycle and small substituent fragments by recursive sequential hydrogen substitution.

together with the bond vector for the incoming fragment. These two vectors were aligned and the new bond distance appropriately set before deletion of the two hydrogen atoms completed the substitution.

In silico approaches to analyse the chemical space of the virtual library were performed using CheS-Mapper 2.0 (Gütlein, Karwath, & Kramer, 2014). The Java Web Start application was downloaded and installed before the 3D viewer of small molecule data sets was accessed locally through the user interface. The compound libraries were compressed into multiple compound **.sdf** file formats using babel, prior to manipulation in CheS-Mapper. The wizard was used to access different steps necessary for dataset preprocessing before visualization in the 3D viewer. After selecting the dataset, in large dataset format, 3D structures within the dataset are preserved before CDK features for extraction such as XLogP and Molecular weight were selected for clustering and embedding in 3D within CheS-Mapper. The dataset was clustered and embedded without cluster alignment before preprocessing and displaying the dataset in 3D space. The features chosen for rendering and embedding were apol (sum of atomic polarizabilities), bpol (absolute value of apol for bonded atoms), nHBAcc (hydrogen bond acceptors), nHBDOn (hydrogen bond donors), nAromBond (aromatic bonds of molecule), XLogP (atom type specific prediction of partition coefficient, log P), Rule of 5 compliance (failures of Lipinski's rule of 5) and ATSm1-5 (Moreau-Broto autocorrelation descriptors based on atomic weight) (Steinbeck et al., 2003). The models generated using the recursive methodologies were curated within Open Babel in order to eliminate any imperfections and inconsistencies associated with hybridization, substitution, bonding and optimized aromaticity. Duplicates were removed before models were saved using simplified molecular-input line entry system (SMILES) identifiers.

Images of the models were generated by the Discovery Studio Visualizer application program interface (API) which automated appropriate rotations and transformations of the models within the dataset before they were saved. The *convert* function of ImageMagick Studio was used in order to format the images to change their size and include labels while *montage* program generated a tiled composite image of the dataset for publication (ImageMagickStudio, 2015).

2.3. RESULTS AND DISCUSSION

The procedure successfully produced 17732 separate ligand structures with the intended diversity of structure and substitutions for Morita-Baylis-Hillman analogues. The diversity in molecular mass (due to variation in linker region, masses of the separate ring systems and the substituents) is shown in Figure 2-4.

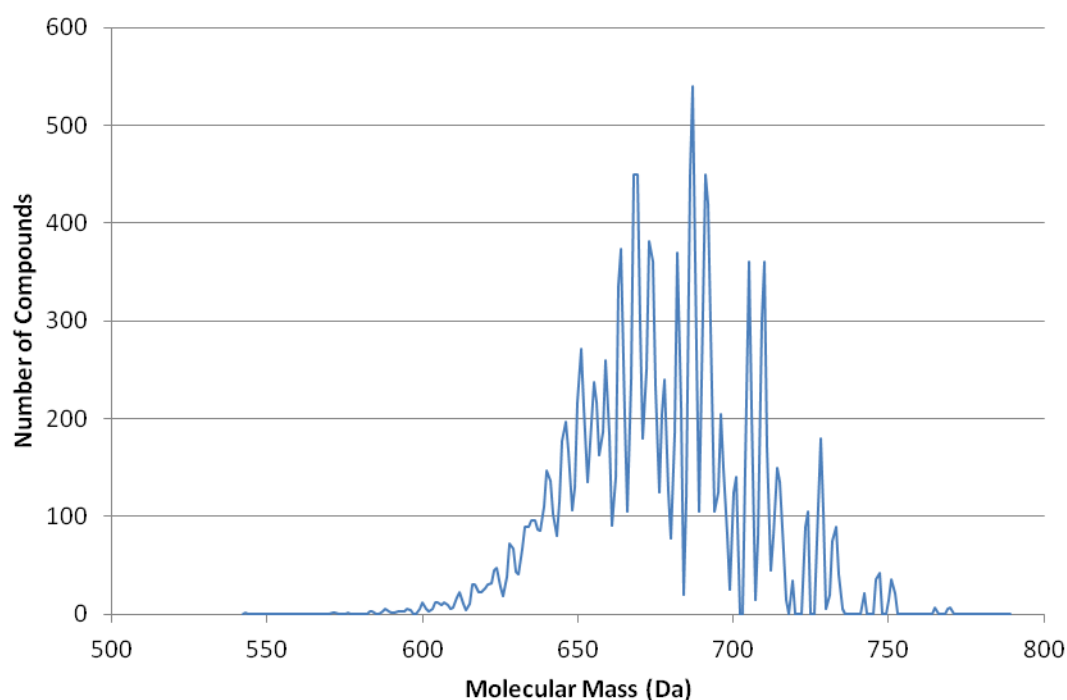


Figure 2-4: Molecular masses distribution of 17750 ligands present in virtual dataset.

The ligand with the largest molecular mass had a mass of 788 Da while the smallest mass was 543 Da. The most popular mass was 687 Da from 540 ligands, while the mean mass of our virtual library set was 675.7 Da. The diversity in the number of hydrogen bond donors/acceptors is shown in Figure 2-5.

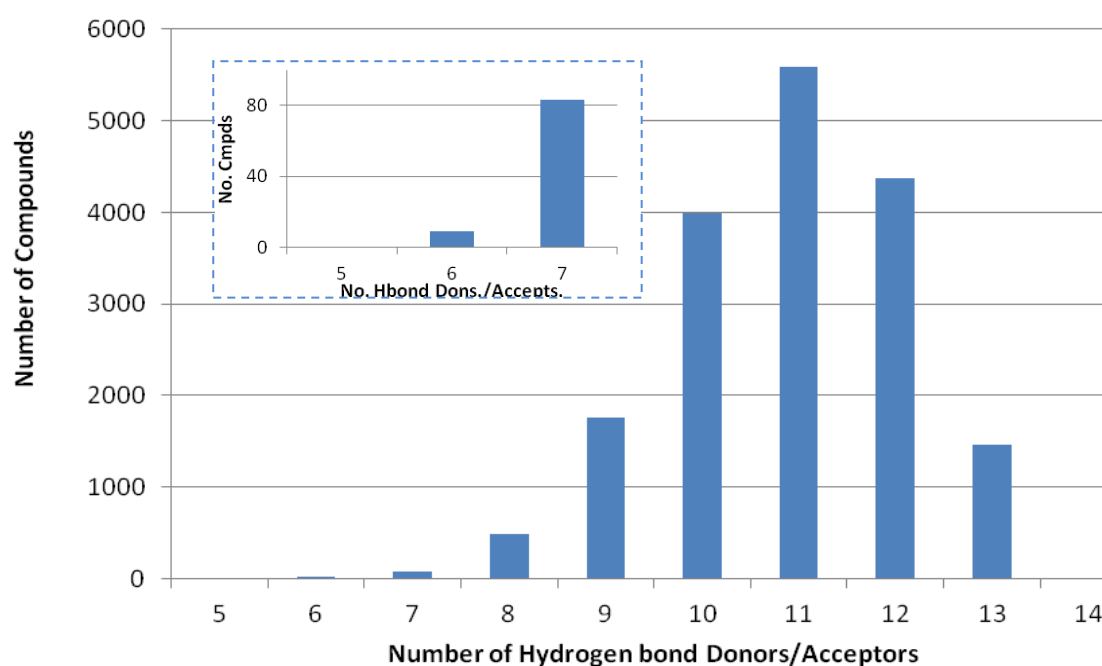


Figure 2-5: Distribution of the Hydrogen bond donors/acceptors in the virtual library data set. Insert shows ligands with 5, 6 and 7 hydrogen bond donors or acceptors.

From Figure 2-5 9 compounds had 6 (lowest number) hydrogen bond donors and acceptors while 5589 compounds 11 had about 11 donors and acceptors. Compounds in the virtual data set had 10.8 hydrogen bond donors and acceptors on average with 1458 compounds having as many as 13.

A random sample of 3512 compounds was extracted from the virtual dataset and analysed using CheS-Mapper. Figure 2-6 shows a 3D chemical space plot of this random dataset with clustering based on a principle component analysis considering 13 chemical descriptors from the CDK toolkit.

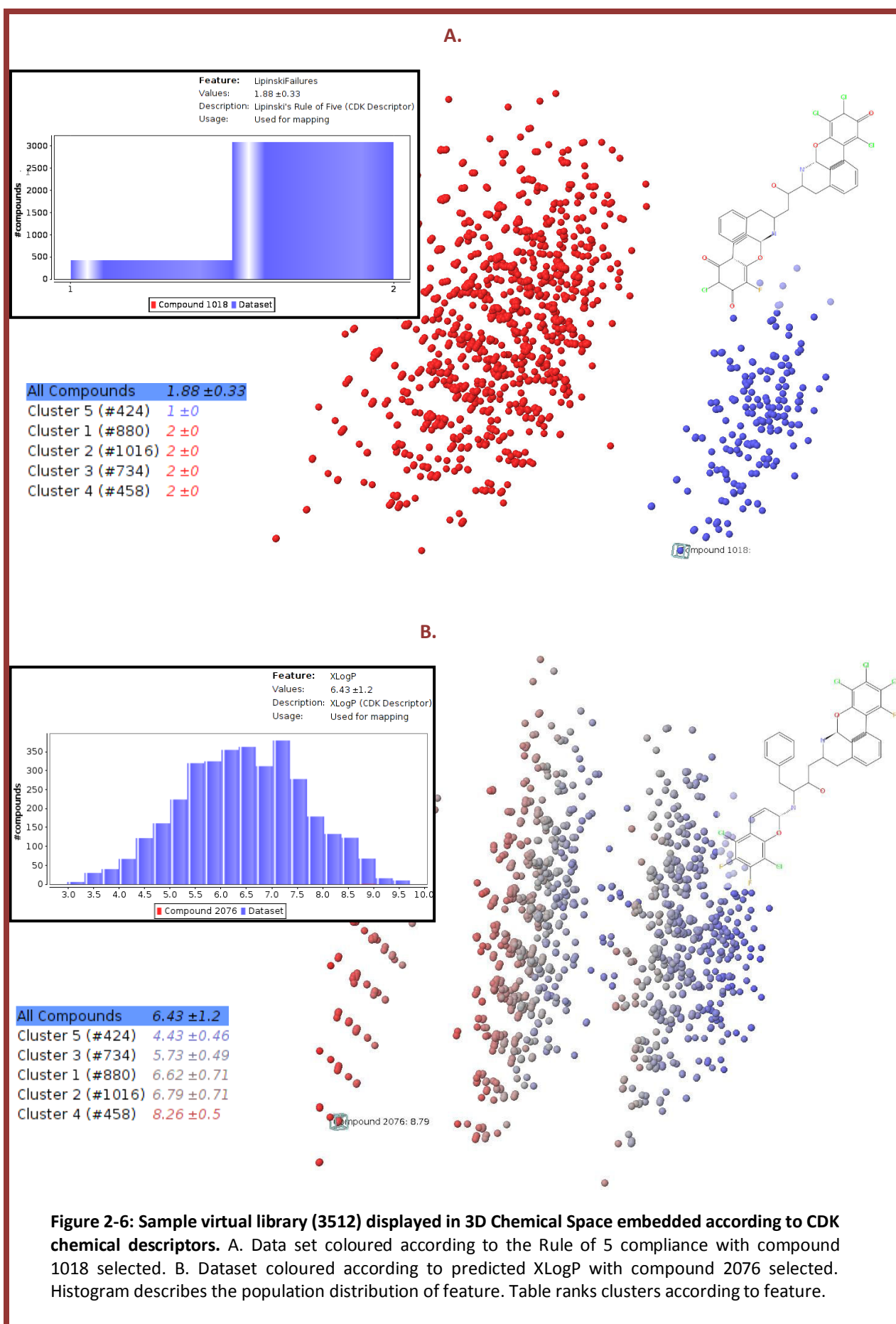
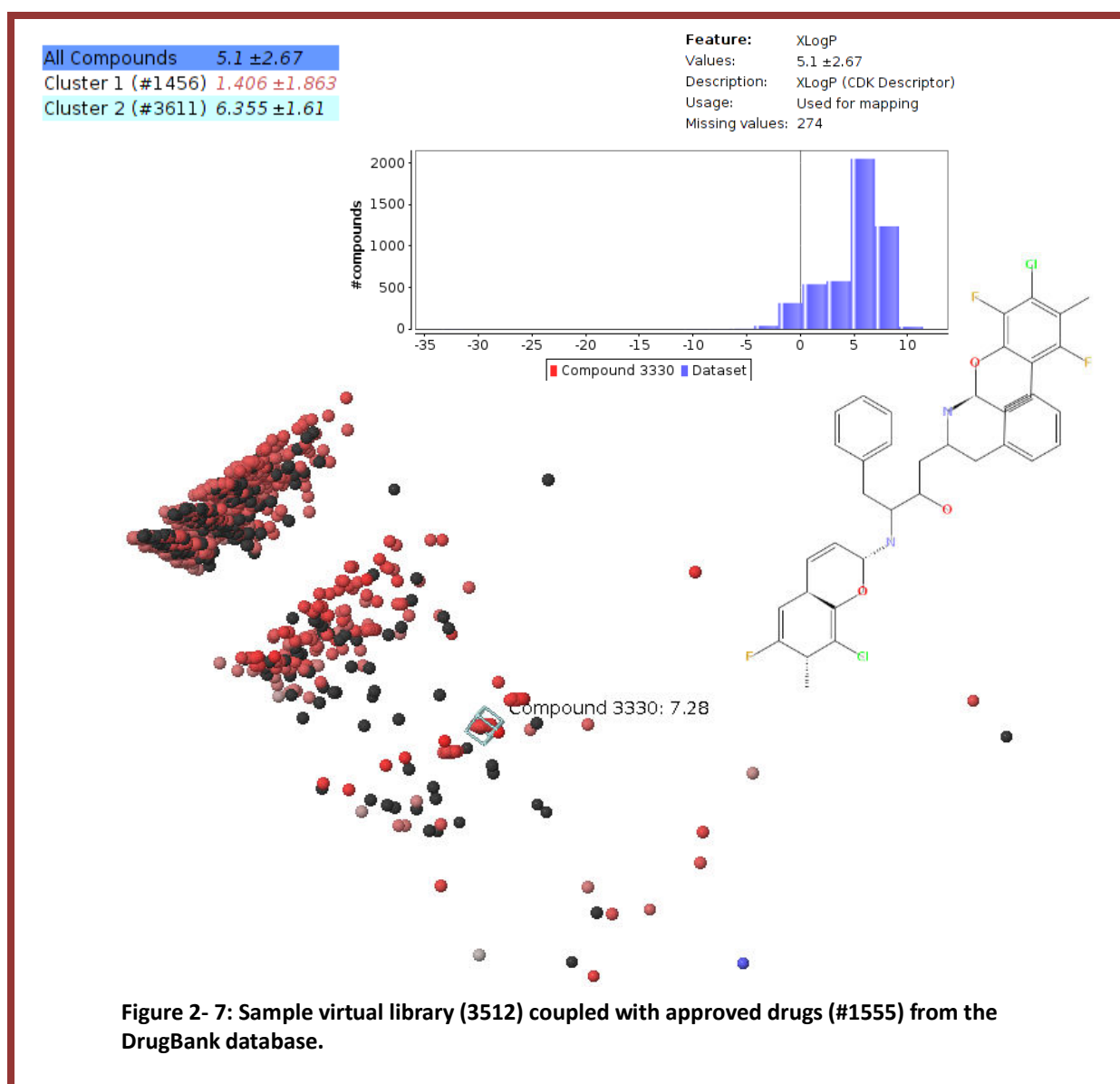


Figure 2-6 A illustrates the distribution of ligands within the dataset that have failed the rule of 5 compliance. When the colouring is based on the Lipinski tests it can be seen that there are 2 major classes within our sample data set. Of the 5 major Lipinski factors the majority of the compounds, 3088, in this sample have failed at most 2 factors with all of them having failed at least once. Rule of 5 compliance with respect to molecular weights defines that drug-likeness is best achieved by masses between 180 and 500 Da (Hann & Oprea, 2004). Considering the distribution shown from Figure 2-4, it is unsurprising that all compounds fail at least once. Rule of 5 compliance with regards to solubility defines the partition coefficient range of drug likeness to be between -0.4 and 5.6 (Hann & Oprea, 2004). The majority of the compounds within our database possess poor predicted XLogP profiles. Ligand 1018 for example had a predicted Lipinski value of 1 and an XLogP value of 4.23 while ligand 2076, however, had a failure value of 2 and a predicted XLogP value of 8.79. Should the lead compounds identified from the screening possess poor drug likeness with regards to the rule of 5 compliance by considering their adsorption, distribution, metabolism and elimination profiles medicinal chemistry approaches will be implemented to optimize them. Figure 2-7 compares the diversity of the sample data set with 1555 compounds from the approved set of drugs extracted from the DrugBank database. In order to remove any bias the 2 datasets were coupled together and loaded into the CheS-Mapper tool.



Coupling the virtual library into a dataset that contains a list of the approved drugs obtained from the DrugBank database allows us to monitor the exhaustiveness of our sample database and thus virtual library. When considering the predicted solubility term it is evident that the majority of the DrugBank compounds possess an XLogP value of 1.4 characteristic of optimum ADME pharmacokinetic properties of orally administered drugs (Hann & Oprea, 2004). Indications of thorough exhaustiveness can be seen from the distribution of our sample dataset in space. It is important to note that computation of chemical space coordinates are performed considering the influence of 13 descriptors selected. Clustering is indicative of the degree of similarity while dispersion indicates dissimilarity. By manipulating the CheS-Mapper viewer we observe that the MBH analogues

aggregate together within the clusters that are present in the coupled dataset, as indicated by green box. This clustering in the presence of the drug-bank dataset indicates to us that we are exhaustively searching for a lead candidate within a finite chemical space. This approach has benefits in its ability to find elusive candidates by searching within a well defined region of chemical space, but however it can be limiting if the ideal ligand possesses properties dissimilar to those within the dataset. By searching within a smaller region of space one is likely to miss this individual. We would recommend using this approach to generate a library of analogues only if the scaffold has been shown to possess antiviral potential as in our case.

A random sample snapshot of the virtual library that is based on the *bis*-coumarin scaffold identified by Onywera *et al.* (2012) as being a likely hit is illustrated in Figure 2-8.

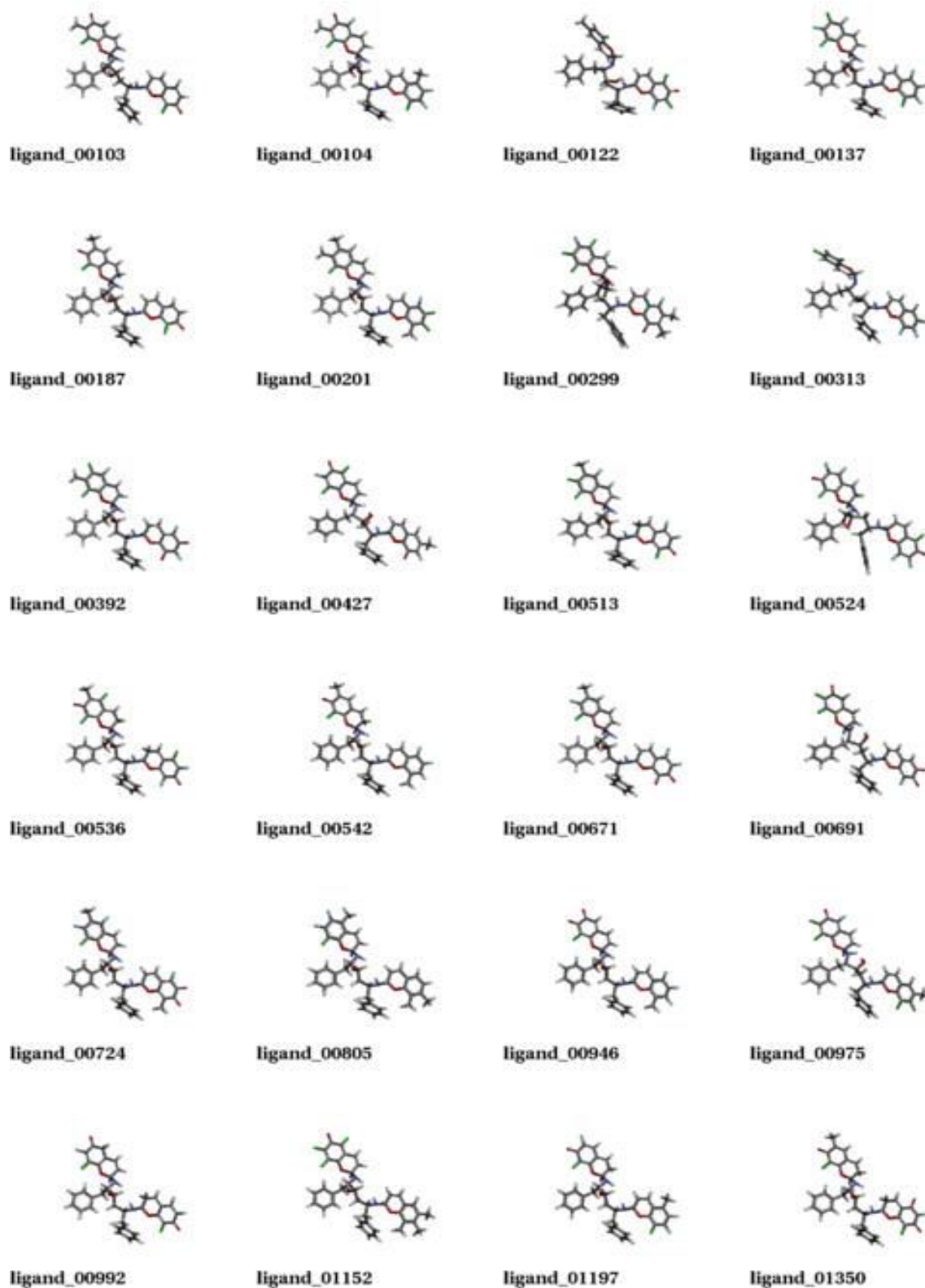


Figure 2-8: Sample snapshot of the Virtual library of *bis*-coumarin analogues generated by recursive methodology.

2.4. CONCLUSION

A virtual library of 17732 analogues was created by utilizing an algorithm that systematically substitutes specific hydrogen atoms in the linker region, heterocyclic region or the substituent region by calling a substitution function in a recursive manner. Each of the 17732 compounds was rapidly curated *in silico* by considering incomplete bonding, unsatisfactory hybridization and optimized aromaticity before being stored in the virtual library. Execution of the program to generate the compounds was time inefficient. It is suspected that exponential duplication occurred during substitutions despite using SMILES format as a storage identifier to prevent similar compounds being present in dataset. Optimisation of this program is necessary before it can be incorporated in the construction of larger and more diverse datasets.

Due to memory restraints on the workstation available, 20 % of the virtual dataset was extracted at random for analysis. The DSV API, ImageMagick and Montage were used to display the sample dataset into a collage for inspection. CheS-Mapper 2.0 was used to inspect the drug-likeness and likelihood of the identification of a lead candidate that had near optimal pharmacokinetic drug like properties. The majority of the compounds in our database show poor solubility features which may need to be optimized for the lead candidate in order to improve its ADME profile.

The use of cheminformatic approaches allowed the rapid and systematic generation of a suitably large virtual library for applications in the identification of a potent lead candidate using high-throughput screening methodologies. *In silico* approaches to investigate pharmacokinetic properties are promising for applications in large datasets.

Chapter 3: High-throughput Virtual Screening

3.1. INTRODUCTION

Incorporation of high-throughput screening (HTS) in the lead identification process is a rapid method of identifying a lead compound from a library consisting of a large number of compounds that show pharmacological potential. The incorporation of miniaturization and robotic handling in HTS allows for the efficient and rapid screening of chemical compounds but has limitations in its applicability (Stanley et al., 2012). Limitations in HTS that occur in the early lead discovery stages are those due to the high-rate of false positives, the tendency of molecules to aggregate when purified as well as the limited scope of the compound library to identify all possible leads (Hann & Oprea, 2004). The process dependant limitations can be mitigated by practical enhancements in order to minimize errors and improve specificity (Stanley et al., 2012). The incorporation of computer-aided drug discovery and high-throughput virtual screening has been shown to be able to give access to Virtual collections which allow the expansion of the lead-like compound space for identification of a suitable hit (Lipinski, Lombardo, Dominy, & Feeney, 2012; Medina-Franco, Giulianotti, Welmaker, & Houghten, 2013). The target specific limitation of empirical screening associated with the existence of an unknown target and the dramatic increase in pharmaceutically relevant targets has allowed the growth in the incorporation of *in silico* approaches, such as high-throughput virtual screening (HTVS), in the drug discovery lead identification pipeline (Nicolaou, 2014; Gregory Sliwoski et al., 2014).

Automated docking with reliable scoring functions in structure-based screening approaches can expand the searchable chemical space in drug discovery protocols (Park, Lee, & Lee, 2006). *Autodock* is a docking program that estimates the binding energy of a ligand in a receptor and predicts its likely binding mode (Goodsell, Morris, & Olson, 1996). By integrating the *Autogrid* program it pre-calculates grid points of interaction within the receptor active sites estimating the atom specific interaction energies incorporated by the AMBER forcefield during docking calculations. The docking simulation is performed by the *Autodock* program which extracts the binding positions from the grid affinity potentials by using a Lamarckian genetic algorithm (Morris et al., 2009).

Despite having a superior scoring function over other docking programs the *Autodock* program is not ideal for applications in HTVS with large datasets. This is due to the need of calculation of interaction energy maps for each unique ligand set (Park et al., 2006). Scripting approaches allow the extraction of atom types within the dataset which can then be used to rapidly calculate the maps, improving the computation costs.

For rapid screening of large ligand datasets against multiple targets we decided to use *Autodock Vina*. *Vina* is a better choice because it achieves faster computation and improved accuracy of binding mode prediction by incorporating a “machine-learning” based scoring function which combines knowledge-based potentials and empirical information in the optimization algorithm for conformation-dependant terms of the function (Trott & Olson, 2010). Further optimizations which allow the implementation of parallelism of processes over multiple central processing units (CPU) and CPU cores help to speed up the run-time. By limiting artificial restrictions such as torsions and the size of the search space available *Vina* improves its accuracy and is thus ideal for applications in High-throughput virtual screening applications over multiple CPU clusters (M. W. Chang, Ayeni, Breuer, & Torbett, 2010).

3.2. METHODOLOGY

The database of 17750 compounds generated using the *in silico* approaches in Chapter 2 were screened using *Autodock Vina* in order to select highly active lead candidates against the 59 HIV-1 protease homology models generated by Onywera et al. (2012). The 59 homology models were built from HIV-1 sequences obtained from infants that were experiencing treatment failure to FDA approved drugs.

3.2.1. Database Screen (*Vina*)

In order to identify a lead candidate that has high activity across the entire model series in an efficient manner it was decided that the bulk dataset of 17750 compounds needed to be trimmed. In order to trim the dataset a Consensus C homology model built by Onywera et al. (2012) from an HIV-1 protease subtype C consensus sequence obtained from Stanford University’s HIV Drug Resistance Database was selected as a target for database screening. From the 17,750 docking experiments evaluated with a *Vina* exhaustiveness of 1, 200 ligands were selected as candidates to be explored thoroughly across the model set.

3.2.2. Exhaustive Screen

The top 200 performing ligands that had affinity for the Consensus C homology model were screened against the entire model data set in order to extract the top performers. By increasing the exhaustiveness to 4 the probability of not finding a global minimum that is far from the native conformation was decreased while the time for each experiment increased. Lead identification incorporated the statistical analysis of the predicted binding affinities across the protease model set. Binding interactions of the top ligands with specific models were analysed by using the Ligplot and Discovery Studio Visualizer.

3.2.3. Pharmacokinetic and physicochemical screening

The *in-silico* pharmacokinetic interrogation was performed in CheS-Mapper 2.0. The 200 ligand dataset obtained from the Consensus C screen was loaded with their binding modes preserved. The 3D structure algorithm selected was the CDK structure generator that used the molecular mechanics forcefield (MM2) in order to optimize the 3D structures from the dataset. This was followed by the selection of 13 molecular descriptors that were calculated by OpenBabel and CDK feature calculators. Clustering was applied by utilising the Cascade k-Means algorithm with machine-learning through the Waikato Environment for Knowledge Analysis (WEKA). The minimum number of clusters chosen was 2 while the maximum was set at 5. The WEKA clustering was incorporated during the principle component analysis (PCA) 3D embedding stage after all the descriptors are calculated. Default maximum attributes of 5 and variance of 0.95 were unchanged. Compound alignment calculations were permitted before embedding. The maximum common subgraph (MCS) of each cluster was calculated before each compound within the cluster's orientation was aligned with the common substructure.

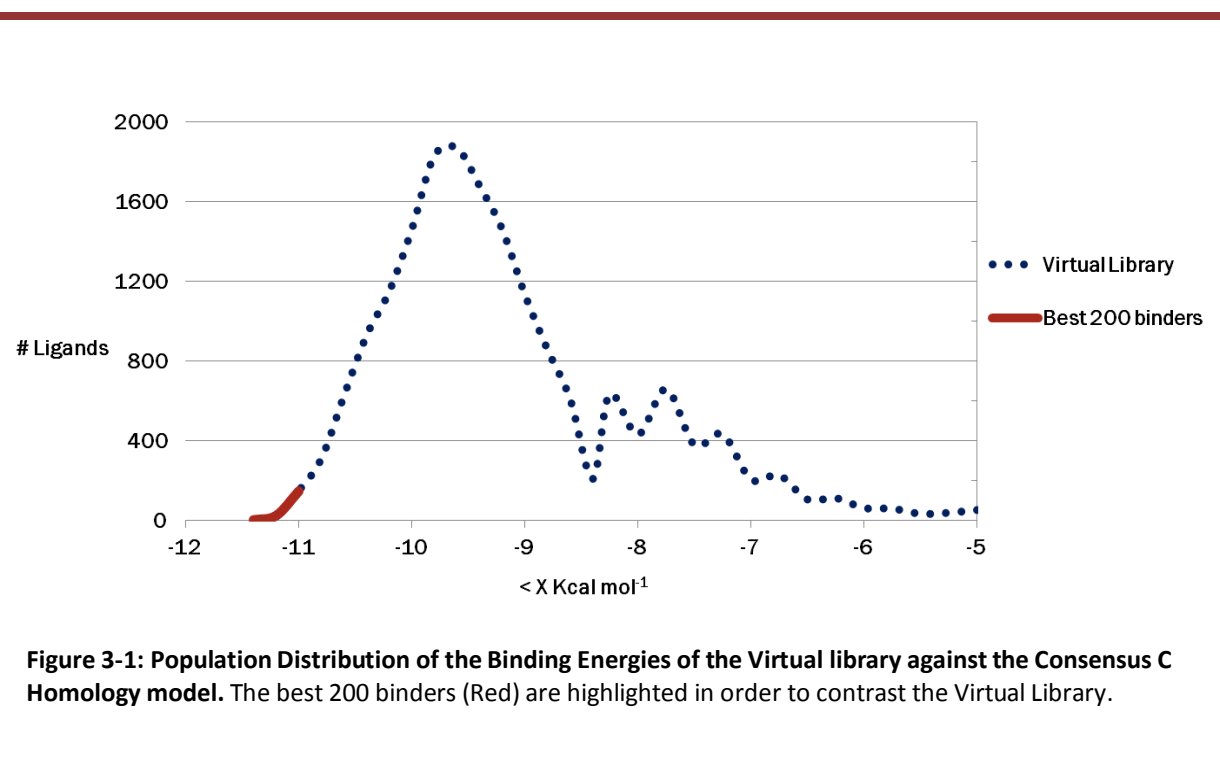
In order to evaluate the exhaustiveness and diversity of the chemical space of the 200 selected ligands, in a second calculation the 200 binders' dataset was coupled with the Drugbank dataset of approved ligands in order to compare its diversity. This hybrid dataset was exported in big data mode and visualized in CheS-Mapper. In this case, the original structures were preserved while no attempt was made to perform alignment calculations of the compounds.

The Antechamber Python parser interface, *ACPYPE*, was used to validate the constitutional integrity of the selected models. *ACPYPE* generates topologies and their topology parameters of small non-nucleic acid, organic molecules, by considering the charge, net

charge, multiplicity and atom types according to a suitable quantum mechanics forcefield (Silva, Alan, & Vranken, 2012). For atoms with open valences, due to inaccurate connectivity, *ACPYPE* fails to generate a suitable topology and thus can be used to verify constitutional integrity.

3.3. RESULTS AND DISCUSSION

The binding energies of individual ligands from the virtual library were obtained from the *Vina* docking experiment against the Consensus C homology model, Figure 3-1.



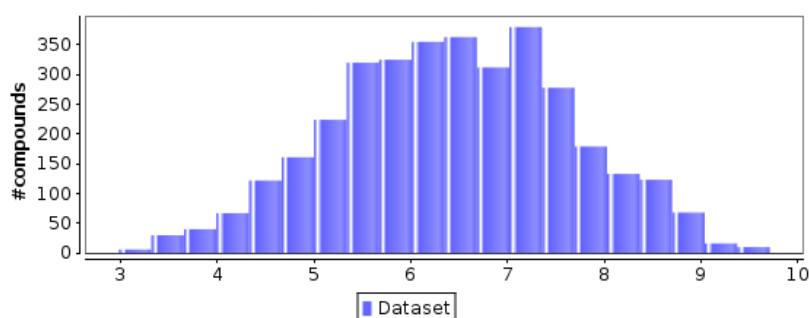
The best 200 binders have a binding of less than $-11 \text{ Kcal mol}^{-1}$ against the Consensus C model. The majority of the library had binding greater than -9 Kcal mol^{-1} , with some interactions as poor as -4 Kcal mol^{-1} . By using the CheS-Mapper tool, it was possible to have a glance at the predicted pharmacokinetic properties of the 200 best performing ligands against the Consensus C model.

The population distribution of the refined dataset of 200 lead candidates was investigated for their predicted *XLogP*, Hydrogen bond acceptors and Hydrogen donor chemical properties. The results obtained from the CheS-Mapper dataset are described below.

Figure 3-2 displays the predicted *XLogP* partition coefficient of each compound. The computed value uses an atom-additive calculation of the octanol/water partition coefficient and estimates its solubility (Wang, Fu, & Lai, 1997).

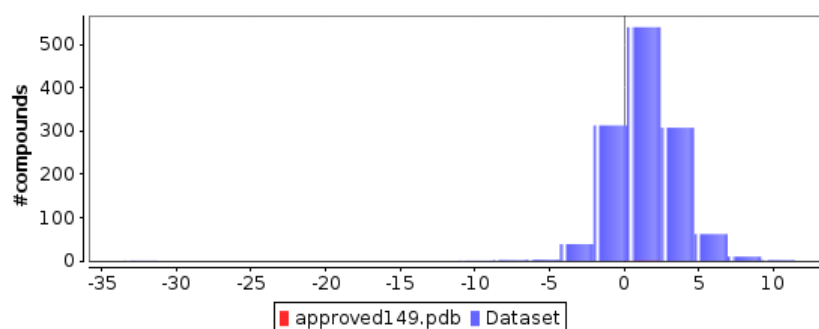
Virtual library sub-set:

Distinct groups = 1068	Min = 2.98	Median = 6.43	Std. Dev. = 1.20	Max = 9.71
------------------------	------------	---------------	------------------	------------



Drugbank approved data-set:

Distinct groups = 1241	Min = 33.52	Median = 1.42	Std. Dev. = 2.24	Max = 11.44
------------------------	-------------	---------------	------------------	-------------



Top 200 binders' data-set:

Distinct groups = 195	Min. = 3.86	Median = 6.53	Std. Dev. = 1.16	Max = 9.27
-----------------------	-------------	---------------	------------------	------------

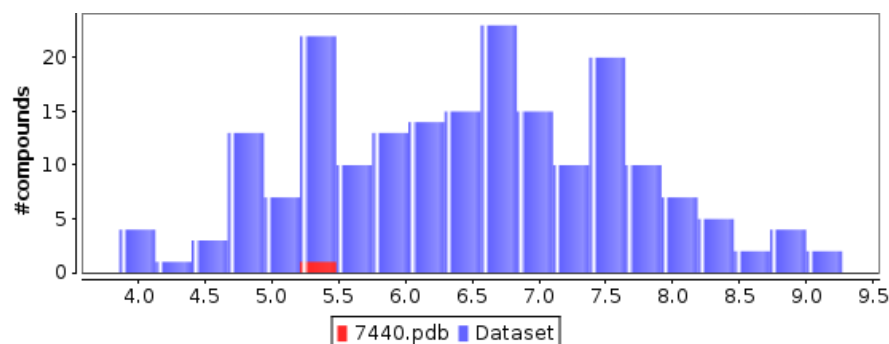


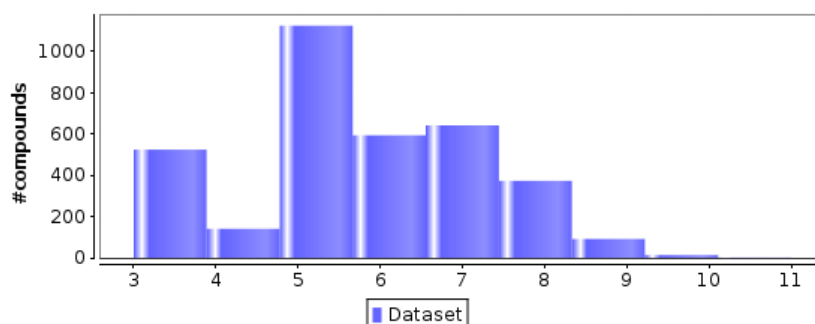
Figure 3-2: Predicted *XLogP* of compounds in virtual library sub-set, DrugBank dataset and the top 200 Consensus C binders.

The Top 200 binders have a bias between 3.8 and 9.3 with the majority having an *XLogP* value greater than 5.6 which is the limit of drug-likeness defined by Lipinski's rule of 5 (Amidon, Lennernäs, Shah, & Crison, 1995). A large *XLogP* is indicative of the tendency of toxic build up in fatty tissues and would have to be mitigated before a viable drug candidate can be selected for (Plika, Testa, & van de Waterbeemd, 1996; Testa, Crivori, Reist, & Carrupt, 2000).

Figure 3-3 shows the computation of the number of Hydrogen bond acceptor groups of compounds present in the virtual library sub-set, the Drugbank data-set and the top 200 binders. Despite having a compound with 191 Hydrogen bond acceptor groups the majority of the compounds in the Drugbank data-set have 4 hydrogen bonds. The top 200 binders by comparison, are spread between 3 and 9 hydrogen bond acceptors with a median of 5, consistent with a median spread in the Drugbank dataset. The top 200 binders set was biased by the inclusion of the linker fragment that ensured the presence of at least 3 Hydrogen bond acceptors. Based on the molecular properties that influence oral bioavailability of drug candidates it is known that increasing permeation and bioavailability of drug candidates is consistent with low hydrogen bond counts (Veber et al., 2002). This preference for compounds that have high permeability needs to be abated with compounds that maintain large numbers of hydrogen bonds which have been shown to stabilise intermolecular interactions which stabilize ligand conformations. Removal of either acceptors or donors have been shown to influence binding affinity and thus attempts at improving permeability will have to exclude the removal of hydrogen bond acceptors (Kuhn, Mohr, & Stahl, 2010).

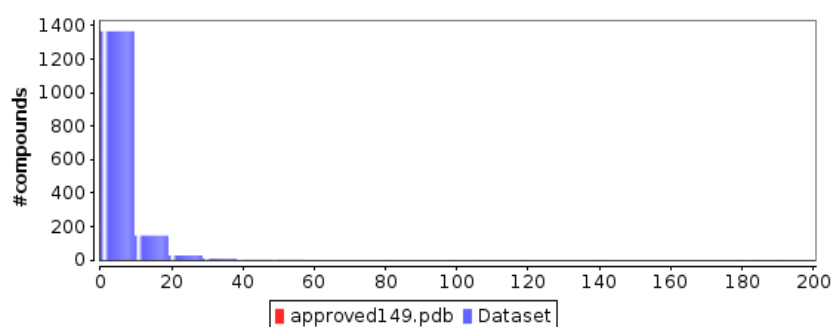
Virtual library sub-set:

Distinct groups = 9	Min. = 3.0	Median = 5.0	Std. Dev. = 1.62	Max = 11.0
---------------------	------------	--------------	------------------	------------



Drugbank approved data-set:

Distinct groups = 41	Min. = 0.0	Median = 4.0	Std. Dev. = 7.37	Max = 191.0
----------------------	------------	--------------	------------------	-------------



Top 200 binders' data-set:

Distinct groups = 7	Min. = 3.0	Median = 5.0	Std. Dev. = 1.58	Max. = 9.0
---------------------	------------	--------------	------------------	------------

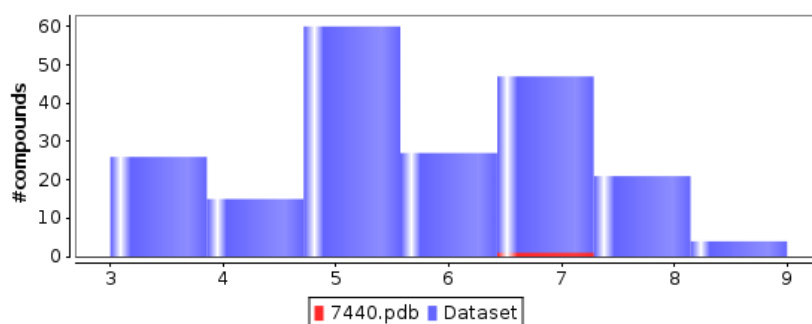


Figure 3-3: Computed number of Hydrogen Bond acceptors of compounds in virtual library sub-set, Drugbank dataset and the top 200 Consensus C binders.

The top 200 binders were embedded on a 3D plot that used co-ordinates obtained from a computation of the principle component analysis of the 13 selected chemical descriptors, Figure 3-4.

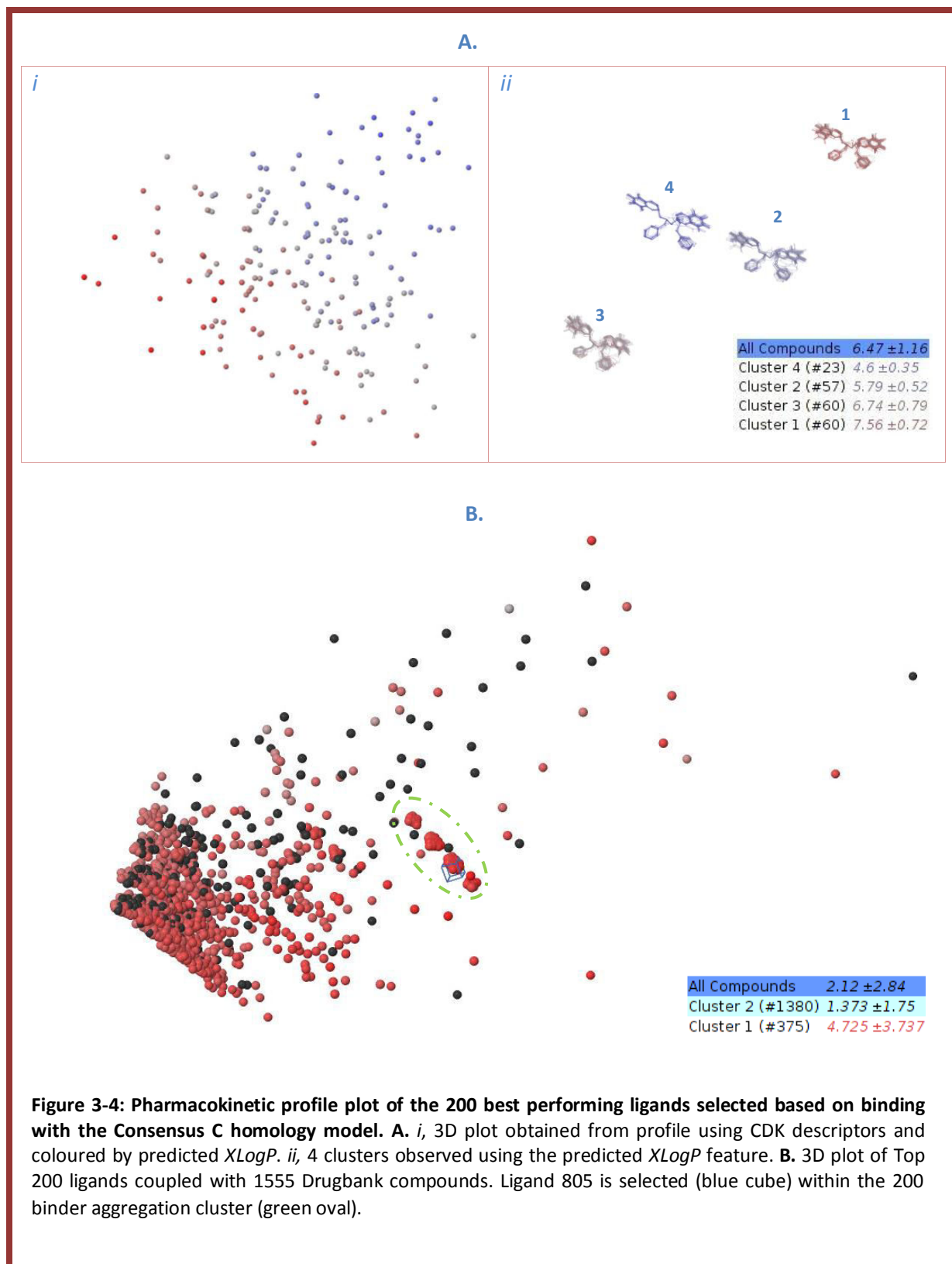
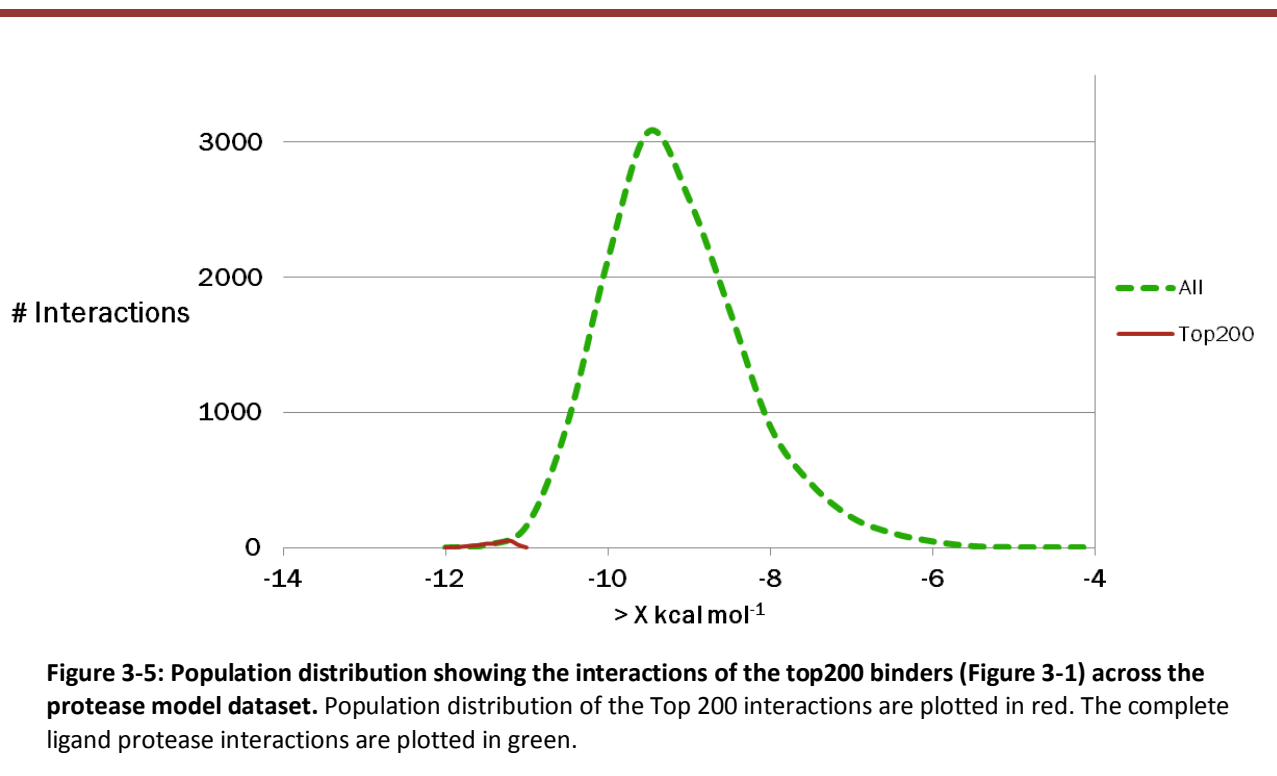
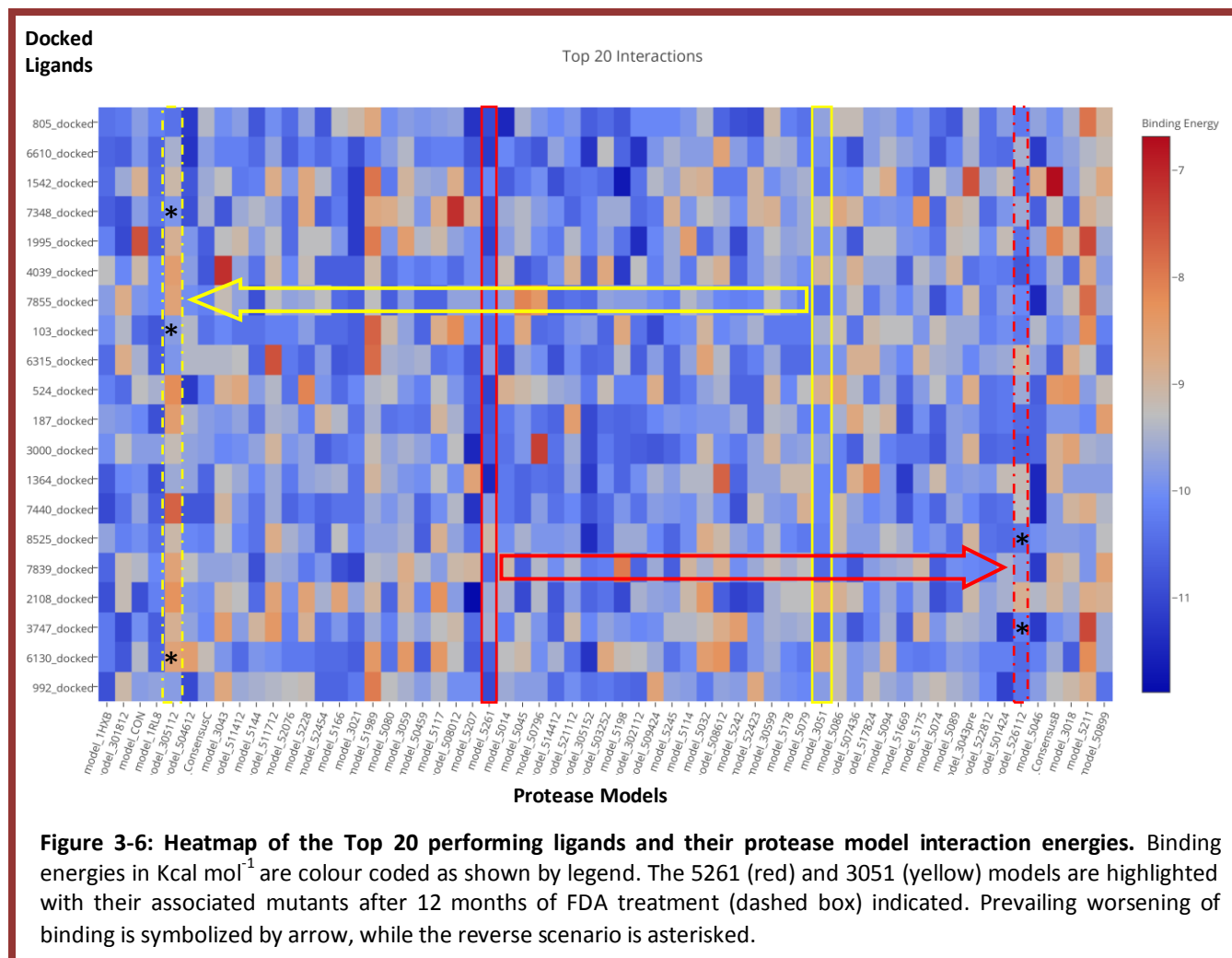


Figure 3-4 A describes the data-set coloured by the $XLogP$ values. The data has 4 clusters defined as having a variance of at least 0.95. Cluster 3 and Cluster 1 appear to dominate the dataset with values greater than 5.94 and less than 8.40. By computing the maximum common subgraph (MCS) of each cluster the compounds within each cluster can be aligned and displayed in Figure 3-4 A, *ii*. This clustering is alternative way of displaying the dataset within the chemical space. Figure 3-4 B, highlights the exhaustiveness of the procedure we are employing in order to identify a synthetic PI candidate. A 3D plot of the hybrid dataset shows that the 200 binders' data-set is localized within a narrow region as opposed to the sparse chemical space exploited by the Drugbank dataset. The influence of anyone feature was limited by exporting data in big-data mode, omitting alignment calculations and using up to 13 features to compute the PCA. Clustering, green oval, thus implies physicochemical similarities between the binders which form a cluster of aggregates.

These top 200 Consensus C model binders were evaluated for their effectiveness against the entire set of 62 protease models and the population distribution of the ligand model interactions are plotted in Figure 3-5. The Top 200 interactions are highlighted in the negatively skewed normal distribution plot.



The mean binding interaction is $-9.5 \text{ Kcal mol}^{-1}$ while the Top 200 interactions are above $-11 \text{ Kcal mol}^{-1}$. Across the model set the top 20 ligands and their ligand vs model binding interactions were plotted in a heat map, Figure 3-6.



In Figure 3-6, desirable binding affinities are reflected by dark colours while light colours are indicative of reduced affinities of the ligands with the receptor in question. It can be seen that the 5261 model (solid red) has the highest susceptibility to the ligand set. When compared to the 3051 model there is a consistent trend across the ligand set. From the CASTp plot (Figure 3-11), the 3051 mutant model showed an increase in surface area and volume while the 5261 mutant had a surface area and volume contraction. There is a possible trend between cavity size and drug susceptibility.

In most instances there was a decrease in binding affinity between the native and the post-treatment mutant for both the contracted and expanded scenarios. Instances when the

reverse scenario of increased binding after mutation was maintained are highlighted with asterisks in Figure 3-6. It is unclear whether the mutations influenced topological changes contributing to the direct influence on binding specificity. Our search aims to identify candidates that maintain significant affinity with the models despite the introduction of drug-resistant mutations within the protease.

A statistical analysis was performed in order to assess the performance of the ligands further equipping us with data to support the choice of the most likely lead candidate. Figure 3-7 shows a histogram of the top 11 ligands selected from the interaction of the top 200 Consensus C binders with the 62 protease model dataset. The ligands were selected based on their mean and median binding energies across the 62 protease models.

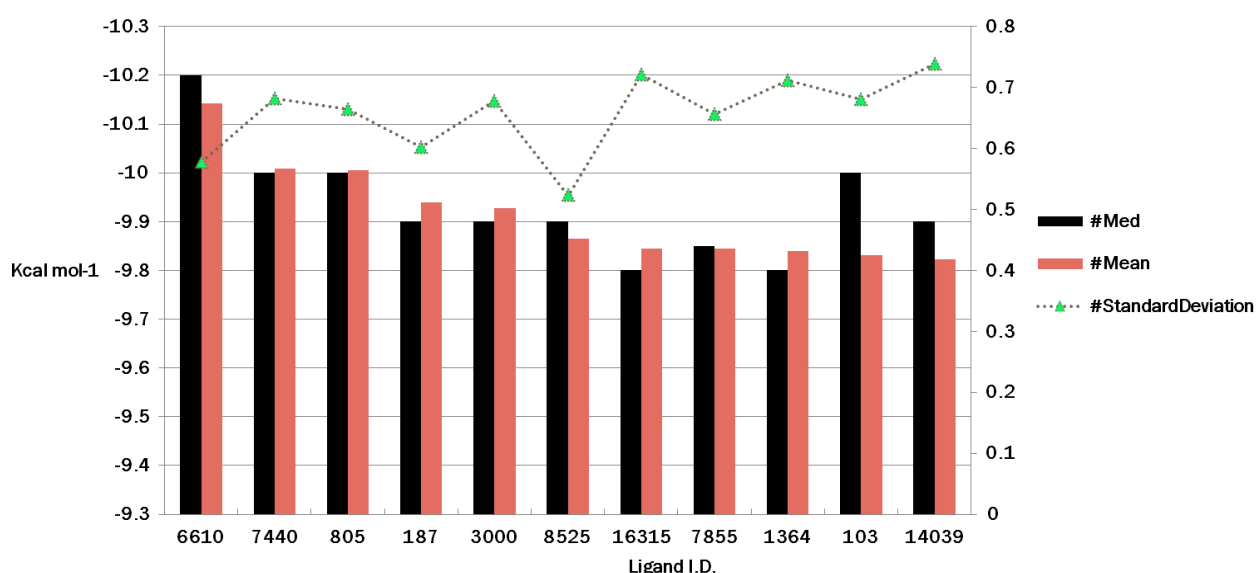
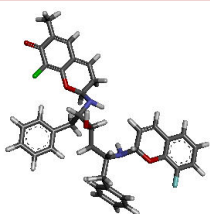
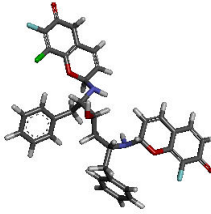
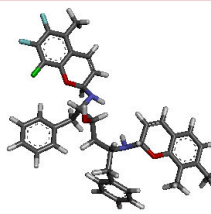


Figure 3-7: Statistical analysis of interaction energies for Top 11 ligands from the 200 binders' dataset across the model data set.

From Figure 3-7 we can see that ligands 6610, 7440 and 805 were the 3 ligands that had the best affinity profiles. Although ligand 103 had a high median binding energy which was comparable to that of ligand 7440 and ligand 805, it had a significantly lower mean binding energy which disqualified it from pharmacokinetic and constitutional screening.

The 3 ligands had their pharmacokinetic and constitutional properties interrogated in detail. Table 3-1 summarizes their properties and results.

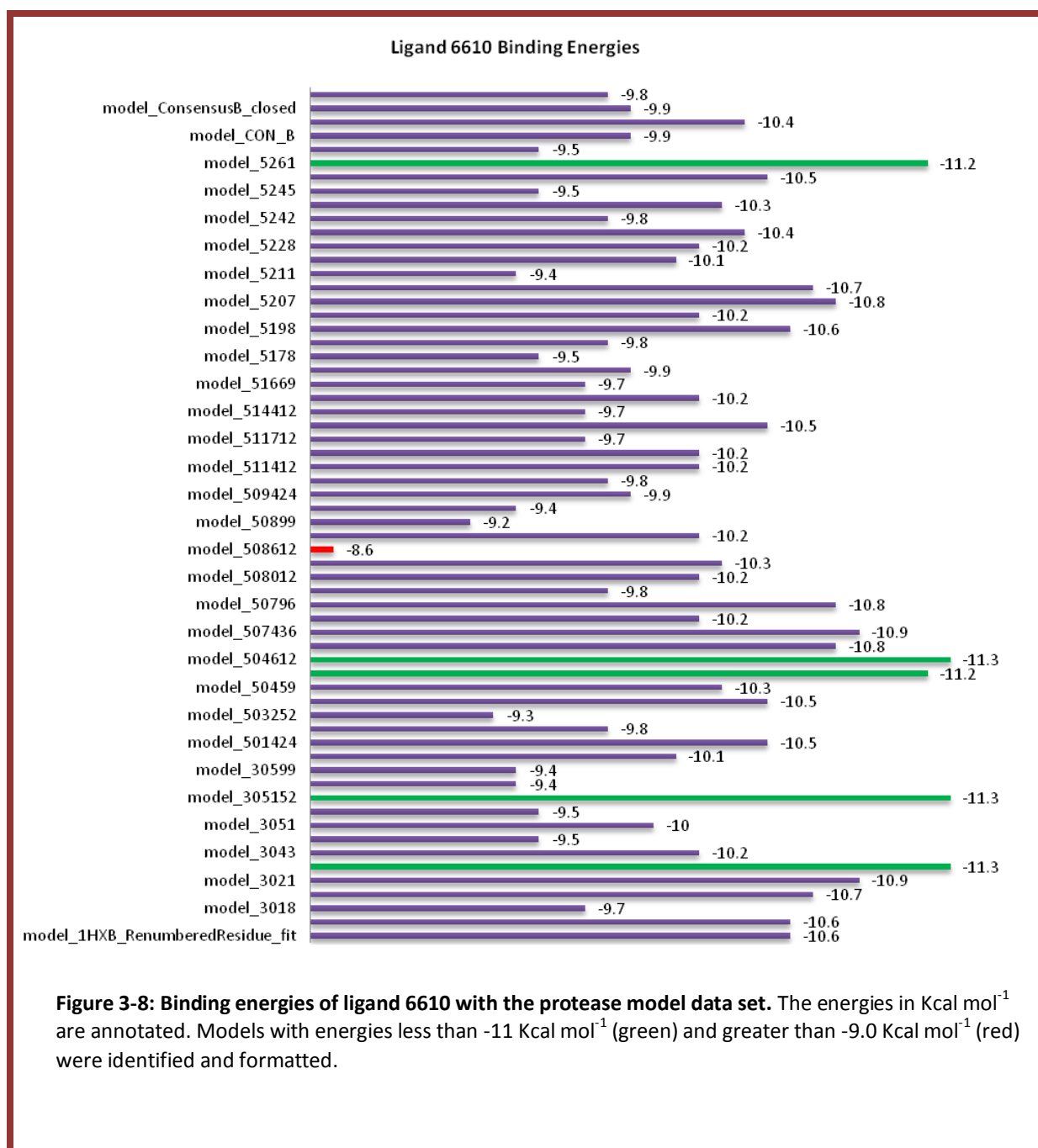
Table 3-1: Pharmacokinetic and constitutional analysis of top 3 lead candidates.

Characteristic	Top 3 Docked Ligands		
Structure	 ligand_06610	 ligand_07440	 ligand_00805
Constituents (substituents; heterocycles; and linkers)	1 x F, 1 x Cl, 1 x CH ₃ , 4 x H, 1 x OH; 2 x Chromenes; Butyldiamine	2 x F, 1 x Cl, 3 x H, 2 x OH; 2 x Chromenes; Butyldiamine	2 x F, 1 x Cl, 3 x CH ₃ , 2 x H; 2 x Chromenes; Butyldiamine
<i>apol</i>	95.94	93.53	101.21
<i>nHBacc</i>	5	7	3
<i>nHBDon</i>	3	3	3
Lipinski	2	2	2
<i>XLogP</i>	6.34	5.38	7.68
<i>ACPYPE</i>	✓ Yes	✗ No	✓ Yes

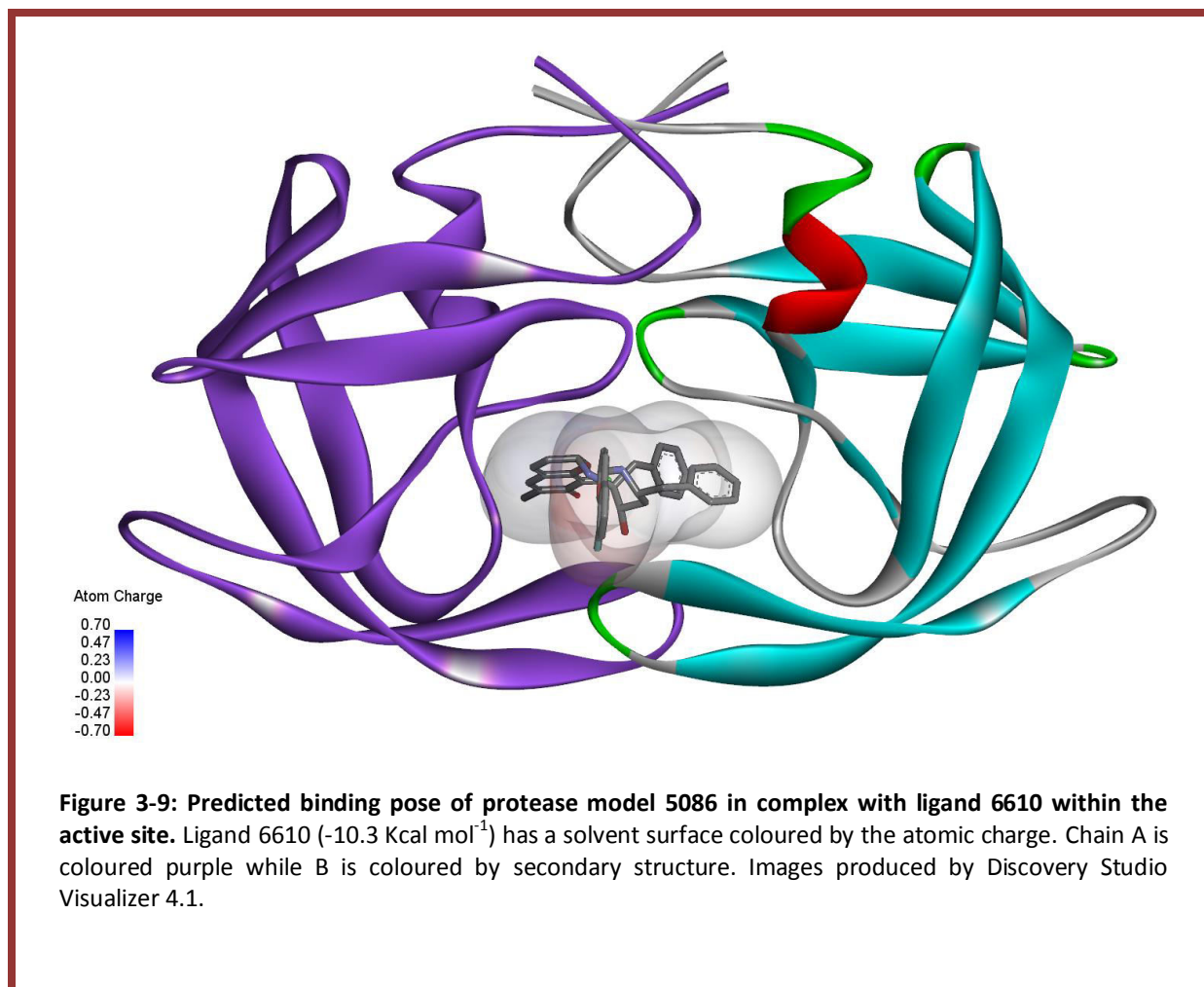
The atomic polarizabilities were calculated within the CDK toolkit of CheS-Mapper and gave an indication of the relative polarity of the molecule which influences its partitioning and stability. The lower the polarity the greater the tissue, brain, affinity of the compound, while polar compounds have a higher tendency to be hydrophilic and thus have less efficient distribution (Hou & Xu, 2003). For our applications ligand 805 has less optimum *apol* properties compared to ligand 7440 and ligand 6610, 93.5 and 95.9 respectively. With regards to hydrogen bond acceptors (*HBacc*) and donors (*HBDon*), ligand 7440 outperforms the rest with 10, 7 and 3 groups respectively, while ligand 6610 has 8 groups and ligand 805 has 6 groups. Due to their large size and their positive *apol*, which influences their predicted *XLogP*, all 3 ligands do not fare well with regards to Lipinski properties that indicate drug likeness. Despite this, ligand 7440 has the best partition coefficient of 5.38 which is consistent with its polarizability. Ligand 6610 with a coefficient of 6.34 and ligand 805 with

7.68, have better constitutional integrity over ligand 7440 with regards to the *ACPYPE* conversion. *ACPYPE* generated topologies for ligand 805 and ligand 6610 satisfactorily, whilst ligand 7440 generated an error associated with valences. Ligand 7440 was thus excluded from further analysis.

Binding data of the best performing, ligand 6610, across the entire protease model data set was extracted and the binding energies were plotted in Figure 3-8.

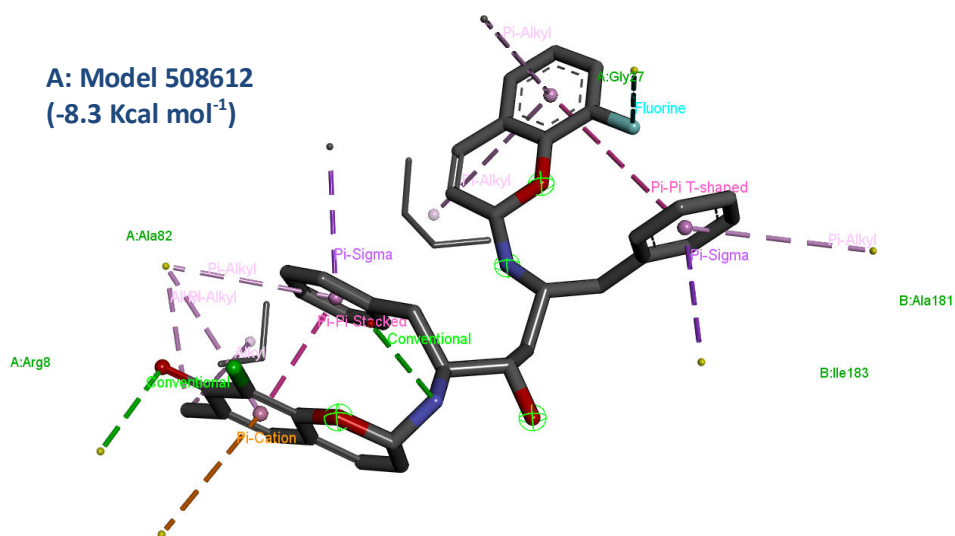


From Figure 3-8 it can be seen that model 5261 ($-11.2 \text{ Kcal mol}^{-1}$), model 305152 ($-11.3 \text{ Kcal mol}^{-1}$), model 302112 ($-11.3 \text{ Kcal mol}^{-1}$) and model 5046 ($-11.2 \text{ Kcal mol}^{-1}$) with its post treatment mutant 504612 ($-11.3 \text{ Kcal mol}^{-1}$) had strong affinities for ligand 6610. However the 12 month post-treatment mutant model of 5086 ($-10.3 \text{ Kcal mol}^{-1}$), model 508612 ($-8.3 \text{ Kcal mol}^{-1}$), had the lowest binding affinity for ligand 6610. Figure 3-9 shows the location of ligand 6610 within the active site.

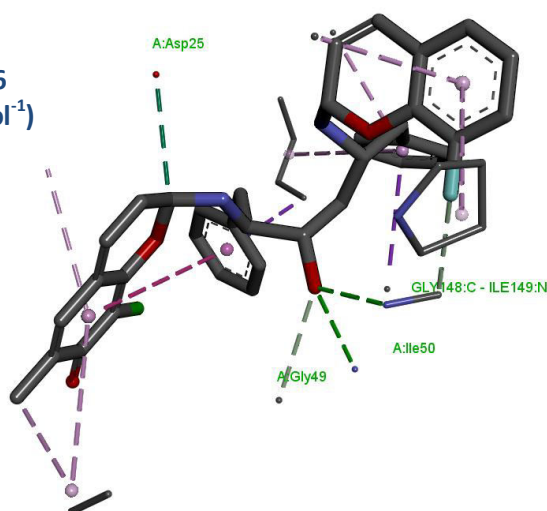


From Figure 3-9 we can see that within this pose, interactions with the symmetric protease model do not impel the conformation of the ligand. The heterocyclic groups interact with Chain A while the phenyl rings are positioned within the Chain B pocket. The ligand and receptor interactions of ligand 6610 with the 5086 and 508612 models are displayed in Figure 3-10.

A: Model 508612
(-8.3 Kcal mol⁻¹)



B: Model 5086
(-10.3 Kcal mol⁻¹)



C: Model 5086 and Model 508612 overlapped

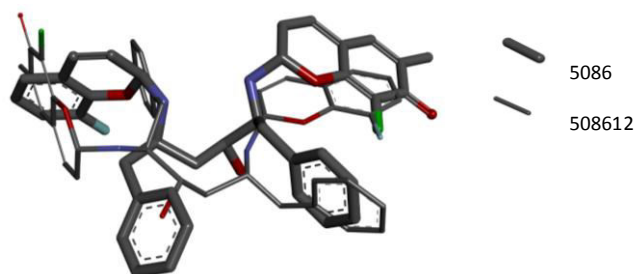


Figure 3-10: Ligand receptor interactions of predicted binding poses of ligand 6610 within the 5086 and the 508612 model cavity. A. Ligand 6610's binding pose with protease model 508612; B. Ligand 6610's binding pose with protease model 5086; C. Overlap of the 5086 binding poses of ligand 6610.

Despite the significant overlap observed from Figure 3-10 C, the pose adopted by ligand 6610 within the vicinity of model 5086 (B) allows the possibility of a non-bonded carbon interaction with the catalytic aspartic acid of Chain A (Qian, Xu, Li, & Frontera, 2003). These electrostatic forces are absent within the vicinity of the pose of model 508612 bound to ligand 6610 and could contribute to the lower binding affinity within this pose. It is known from the CASTp analysis that model 508612 experiences a drop in surface area and volume from the native 5086 model. The drop in the binding affinity experienced from the 5086 -> 508612 mutation can be explained by 2 phenomena, the ligand orientation affecting the ability to interact with the active site aspartyl while an active site contraction contributes to ligand protease interactions experienced.

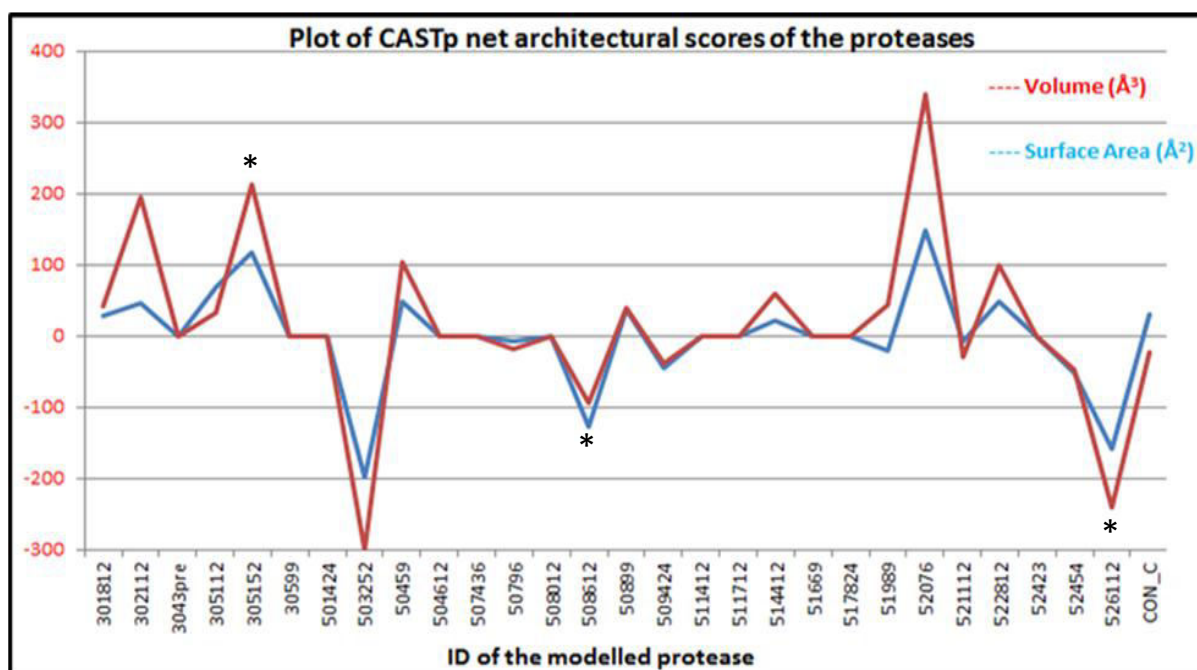


Figure 3-11: Computed Atlas of Surface Topography of proteins (CASTp) plot for the protease models.
Onywera et al. (2012).

Interactions of ligand 6610 and 805 with models of 3051 (expansion) and 5261 (contraction) were visualized using LigPlot. The results are shown in Figure 3-12 and Figure 3-13.

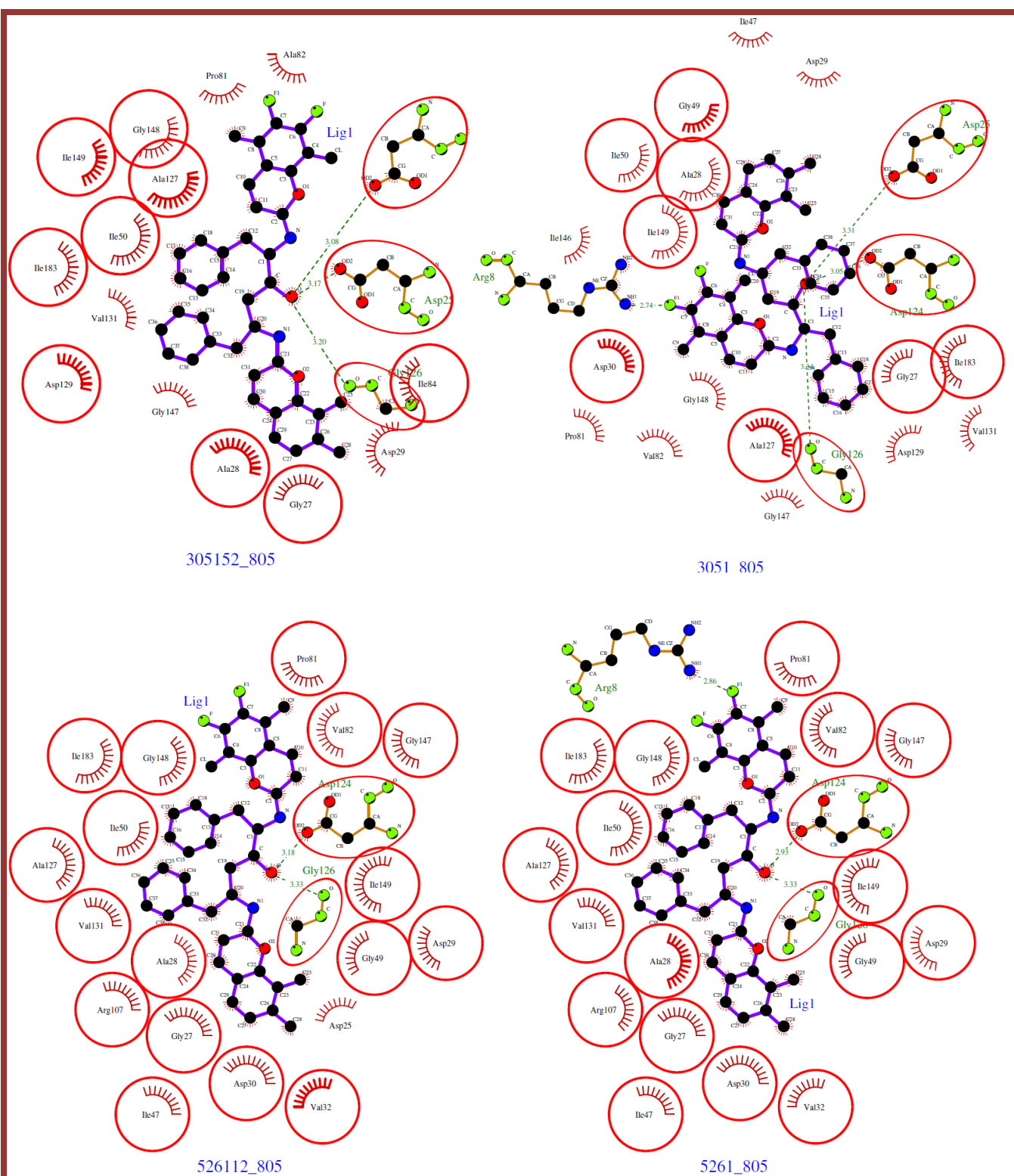


Figure 3- 13: Ligplot analyses of ligand 805 with expansion mutation model 3051 and contraction mutation model 5261.

3.4. CONCLUSION

The virtual data set of 17750 compounds has been screened in order to identify a suitable lead candidate for synthesis and optimization. After high-throughput virtual screening that implemented the rapid *Auto Dock Vina* scoring function the best 200 binders to a Consensus C homology model were chosen from the initial *in-silico* dataset. An exhaustive screen across the entire protease model data set allowed the identification of 11 hits that showed potential to be lead candidates based on their mean and median binding energies. Ligand 6610, 7440 and 805 were selected as having the most optimal binding energies. *In silico* pharmaco-profiles obtained from the PCA of chemical descriptors which were computed in CheS-Mapper 2.0 were used to generate the 3D plot of chemical space exploited by our library.

Although ligand 7440 had the best pharmacy-profile in terms of polarisation, hydrogen bonding, *XLogP* and aromatic bonding, it had improper connectivity which resulted in failures during the *ACPYPE* test for constitution and structural integrity. Ligand 6610 was assumed to have closed valences and despite its poor partition coefficient was selected as the lead candidate from our virtual dataset.

Sub-optimal binding of ligand 6610 was observed against the 508612 model and closer inspection revealed that the pose lacked an interaction that integrated the catalytic aspartic acid 25 residue within its complex. Simulating the dynamic solution behaviour of the ligands with the model sets in suitable solvent will give insight into the stability of the ligand-protein complexes. Understanding the forces involved in stabilizing the receptor-ligand complex will help to identify portions that can be modified in order to optimize the oral bioavailability and perhaps improve the potency of their interaction with the active site.

Chapter 4: Molecular Dynamic Simulations

4.1. INTRODUCTION

Originating in physics, molecular dynamics is the application of computational approaches to the simulation of molecular perturbations. By calculating their trajectories based on solving Newton's equations of motions and if resources allow quantum-mechanical laws, simulations of molecular behaviour that approximates reality can be achieved (Feynman, 1985; Rahman, 1964). Developments in the understanding of the influence of microscopic simulations on macroscopic properties through the implementation of statistical mechanics lead to the application of MD in the study of thermodynamic and kinetic behaviour of biological systems (Wereszczynski & McCammon, 2012).

Successes of MD simulations are limited because of the incorporation of assumptions and approximations to the calculation of forces affecting molecular perturbations. In molecular mechanics force-fields, electronic polarization is ignored and thus atoms are given permanent partial charges which neglect quantum effects. Incorporation of quantum mechanic calculations allows transition metal effects and catalytic mechanisms to be modelled at the cost of computational time. Due to their high-computational demands simple molecular mechanic approximations have relatively short simulation times which overlook the effect of conformational shifts. There is a consensus surrounding the dynamic state and druggability of pharmacologically relevant targets. In the absence of ligands these proteins, such as the acetylcholine-binding protein, sample through conformational states which possess numerous binding modes that can be stabilized by inhibitors (Bourne, Talley, Hansen, Taylor, & Marchot, 2005). By not sampling all possible conformations in the simulations relevant binding modes may not be identified and exploited in target specific drug discovery.

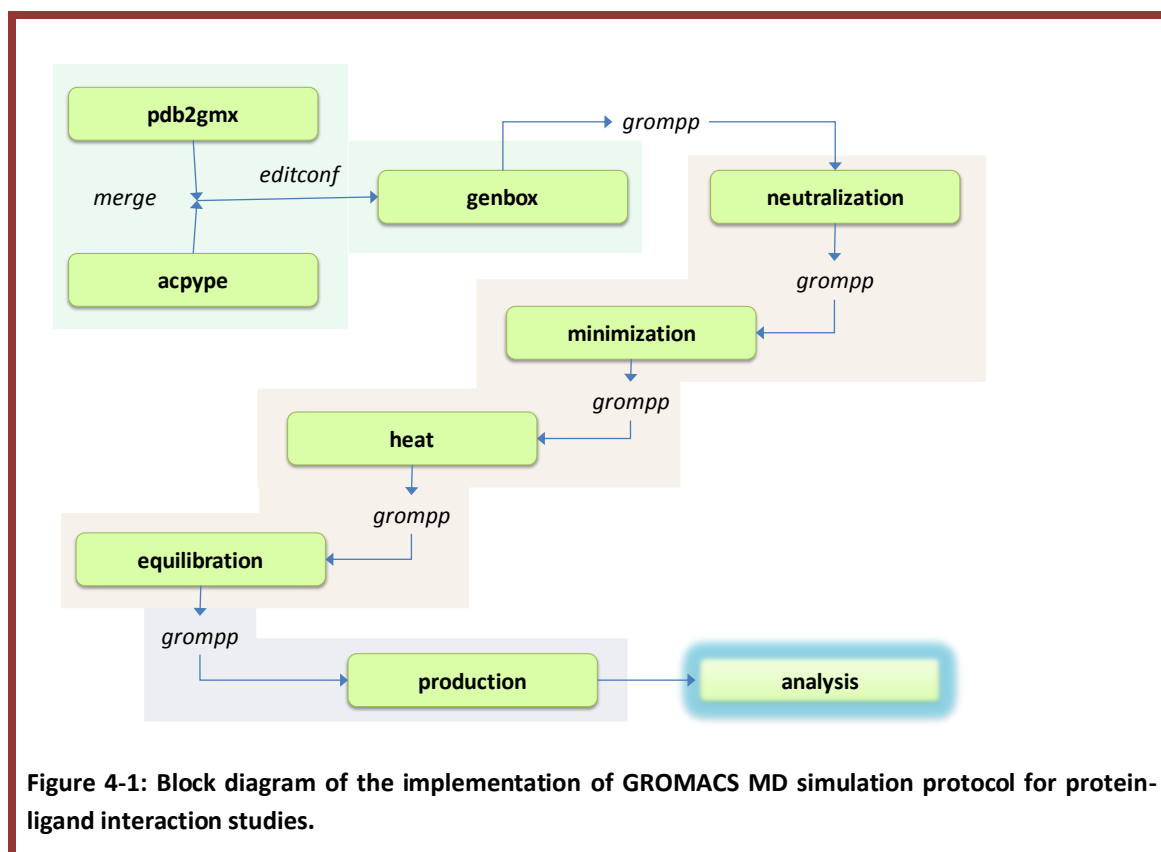
There has been growing frustration from incoherent reduction in the number of approved drugs despite the increase in drug development costs. The escalation of computing power has allowed the incorporation of computational approaches such as MD simulations into drug discovery pipelines (Durrant & McCammon, 2011). The information contained in MD simulations can be exploited for drug discovery by identifying elusive cryptic binding sites as

in the development of the HIV integrase inhibitor raltegravir (Hazuda et al., 2004). Optimisation of conformational sampling is achieved through the application of a relaxed complex scheme. In this approach multiple conformations obtained from simulations are used to obtain a multi-spectrum docking score for a ligand with a receptor thus improving the approximation of the ligand binding energy and the identification of true binders (Schames et al., 2004). MD simulation based methods of calculating the binding of a ligand to its receptor such as thermodynamic integration, single-step as well as free energy perturbation approaches have been devised in order to obtain accurate estimates of ligand potency (Adcock & McCammon, 2006; Kim et al., 2006; Schwab & van Gunsteren, W. F. Zagrovic, 2008).

The Groningen Machine for Chemical Simulations (GROMACS) was used to perform MD simulations of models of the protease enzymes with lead candidates that showed pharmacological potential. Analysis of the simulations will give us a more accurate prediction of the binding energy and stability of the protein ligand complexes. By implementing a Message Passing Interface, GROMACS is able to efficiently utilise multiple processors by splitting its tasks to speed up longer simulations (Berendsen, van der Spoel, & van Drunen, 1995).

4.2. METHODOLOGY

We used GROMACS version 4.5.7 installed on a Centos 5.10 Linux server. Ligand-Protein complexes of ligand 6610 and ligand 805 across the model data set were solvated and allowed to proceed for 5ns of simulations. A block diagram that illustrates the implementation of the GROMACS programmes that produced the trajectories is described below. Python scripts were written in order to automate the execution of the simulations. The script is outlined in Appendix A.3.



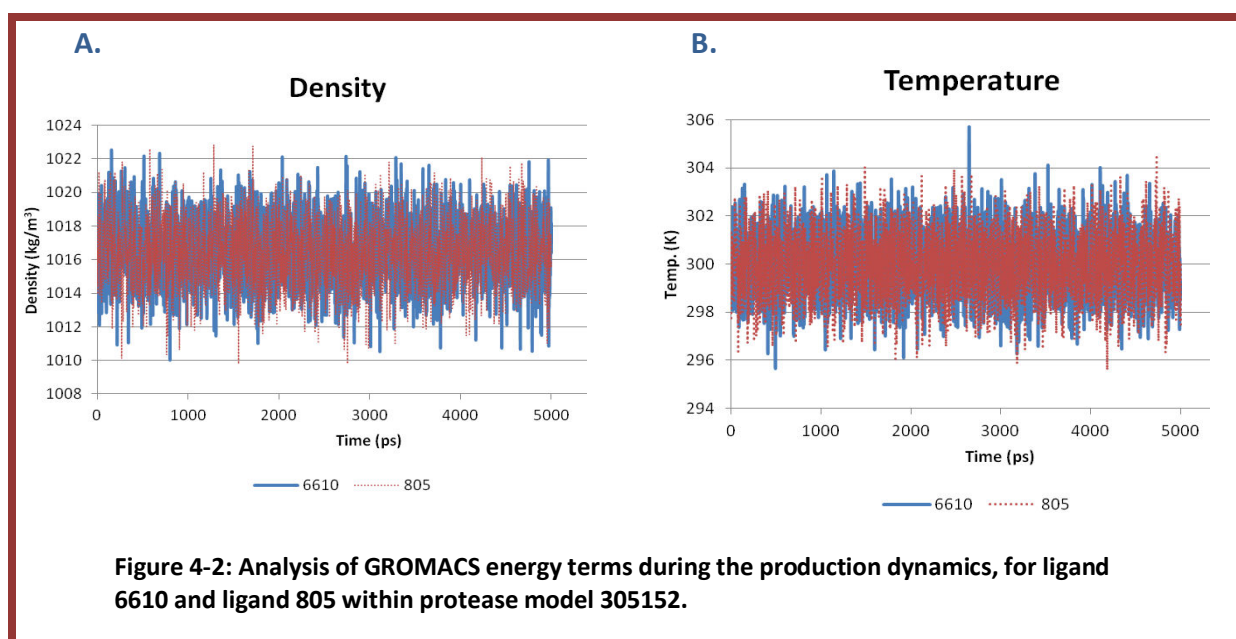
The antechamber python parser interface (ACPYPE) was used to generate topologies for the lead candidates by using the generalized amber force field (GAFF) to assign the net charges and force field parameters of atoms within a ligand that did not have any open valences. Protein topologies were created by the *pdb2gmx* program before the ligand *pdb* and the protein *pdb* files were merged using UNIX commands. The protein topology file was appended with the ligand name before a simulation box was prepared by *editconf*. The protein was placed in the box before it was soaked with solvent waters by running the *genbox* script. The solvation GROMACS structure file (.gro) and its corresponding topology output file (.top) specifying the ligand, protein and water system was neutralized by replacing 6 water molecules with 6 chloride ions during the *genion* dynamic run after a neutralization preprocess step was initiated. Relaxation of the neutralized system to remove steric clashes and inappropriate geometries is prompted by a 2000 step steepest descent energy minimization simulation after preprocessing. The minimized structure files (em.gro) and topology files are prepared for heating in order to couple them to the heat bath. Over 5000 steps of simulation, temperature coupling and relaxation to the new conditions were executed.

After heating, the pressure of the system is increased during the equilibration simulation step by pressure coupling. This 5000 step equilibration stage utilizes position restraining forces in order to restrain all non-hydrogen atoms whilst optimising the water solvent with the solutes in order to achieve an accurate density. The equilibrated system is prepared for production simulation by a preprocessing step. Due to restricted computational resources the production simulations were allowed to proceed for 5ns in order to interrogate structural fluctuations within the complexes. The extraction of binding free energies based on the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method was computationally expensive (Houa, Wangb, Lia, & Wang, 2011; Miller et al., 2012). We thus repeated the production dynamics over 50 ps in order to extract initial binding free energy approximations.

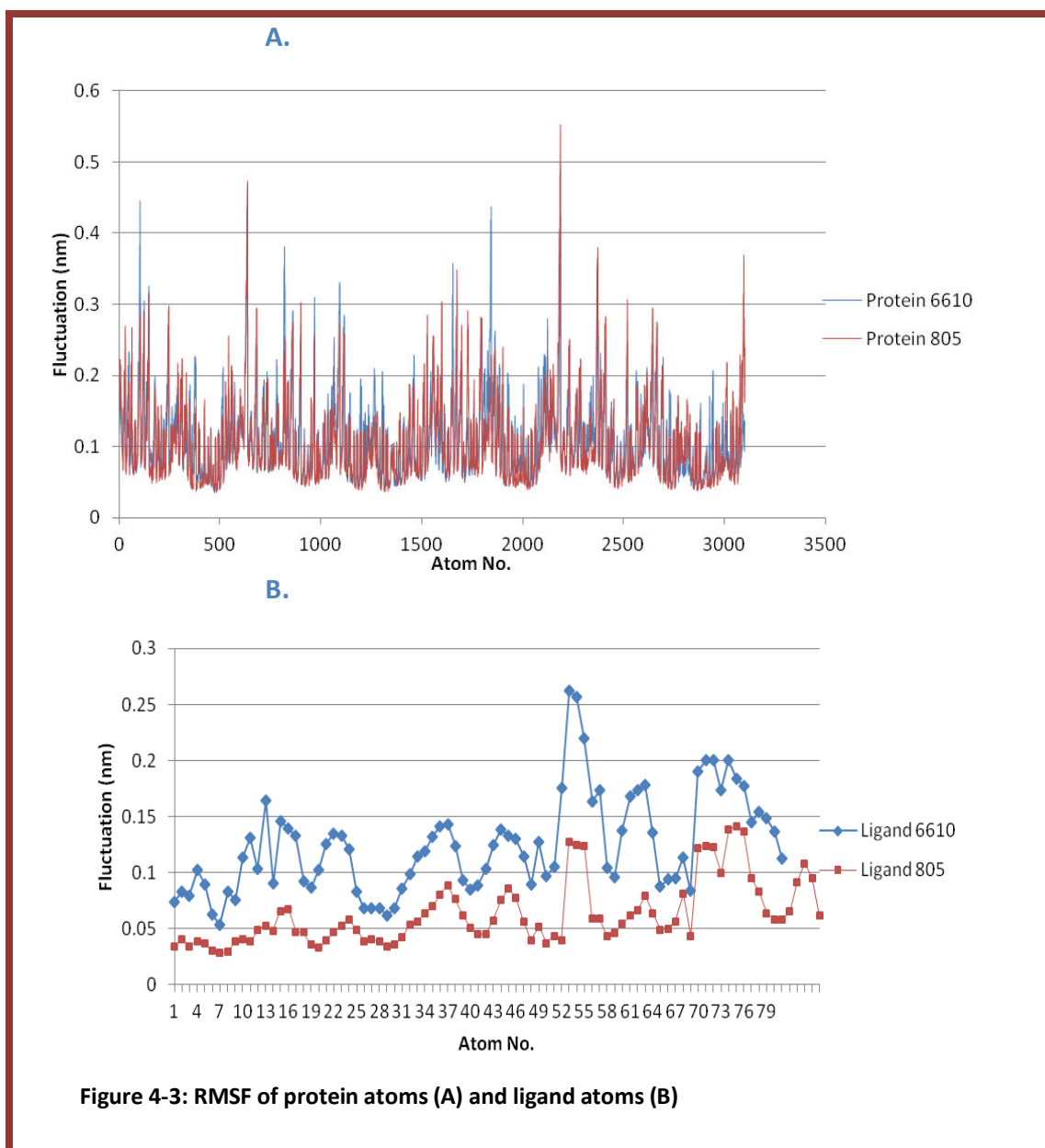
The results of the MD investigations are described and discussed below.

4.3. RESULTS AND DISCUSSION

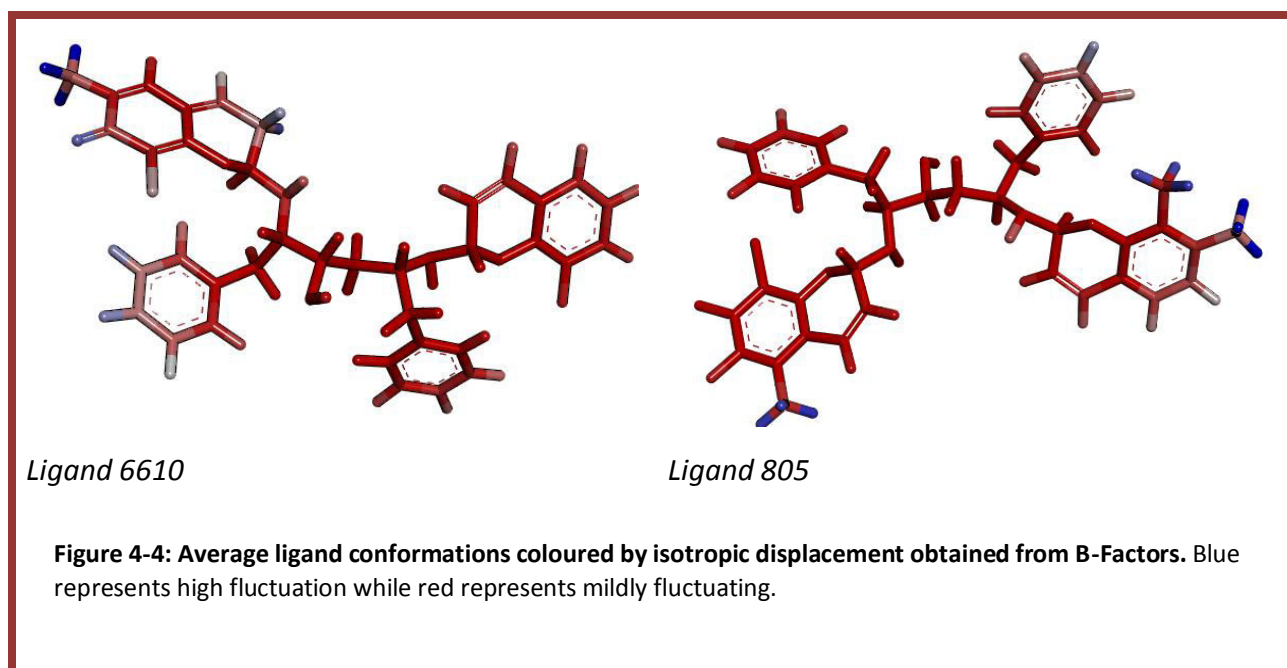
MD simulations of the ligand 805 and ligand 6610 lead candidates were analysed in order to ascertain the quality of the simulations and the structural behaviours that are perturbed as a result of the simulation. In order to reinforce our selection of ligand 6610, an investigation that interrogated the ligand stability within 305152 protease model was followed. Figure 4-2 is an illustration of the convergence and stability of the density and temperature during simulation as a result of the pressure and temperature coupling respectively.



From Figure 4-2 we can see that during the 5 ns of simulation the systems pressure and temperature were stable as a result of the efficient preparation steps, that allowed the energy terms to converge and stabilize prior to production. During the simulations the ligand and protein atoms undergo fluctuations associated with the stability and flexibility of the protein domains and the strength of the interactions with the ligands. A plot representing the stability of a complex can be traced by calculating the root-mean square fluctuation, RMSF, of the protein and the ligand atoms about an average position throughout the simulation, Figure 4-3.



There are minimal distinguishing features between the protein model when it is in complex with either ligand 805 or ligand 6610 in terms of its RMSF fluctuation during the simulation. Analysis of the ligand fluctuations in more detail showed that atoms within ligand 6610 fluctuate marginally more than atoms within ligand 805, Figure 4-3.



Perhaps due to the increased volume due to methyl groups the ligand 805 heterocyclic systems are more stable compared to the largely less substituted ligand 6610 systems with the hydroxyl group as main entities.

When the deviation between the starting conformation and a conformation extracted during the simulation is calculated via the root mean squared deviation, RMSD, Figure 4-5 is obtained. The RMSD is expected to increase and eventually stabilize during the simulation.

From Figure 4-5 A, we can see that the protein RMSD expands to 0.2 after 1.3 ns of simulation in protein in complex with ligand 6610 but continues to show large fluctuation, while protease model in complex with 805 expands to 0.18 in the same time frame remains relatively unchanged over the next 2 ns.

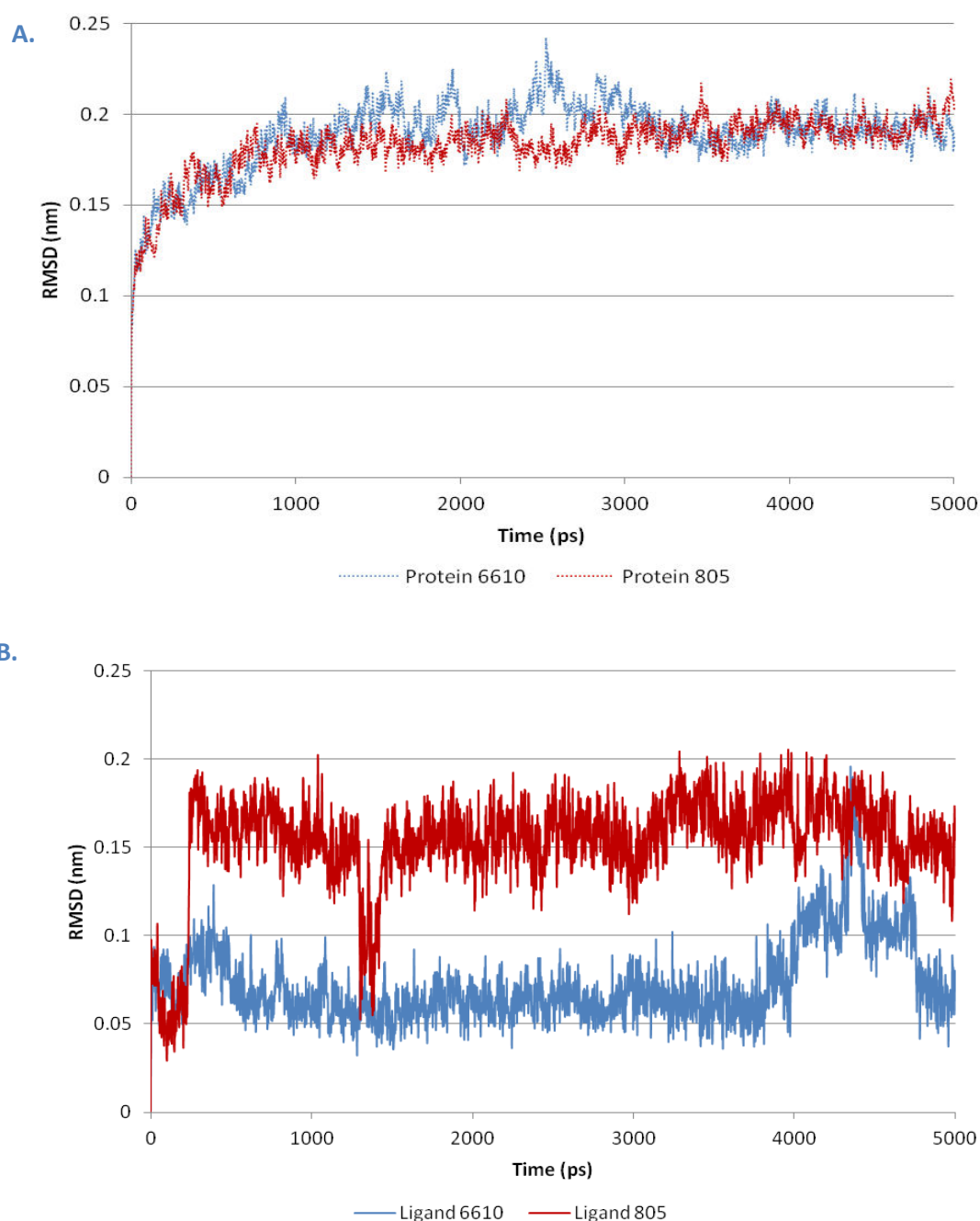
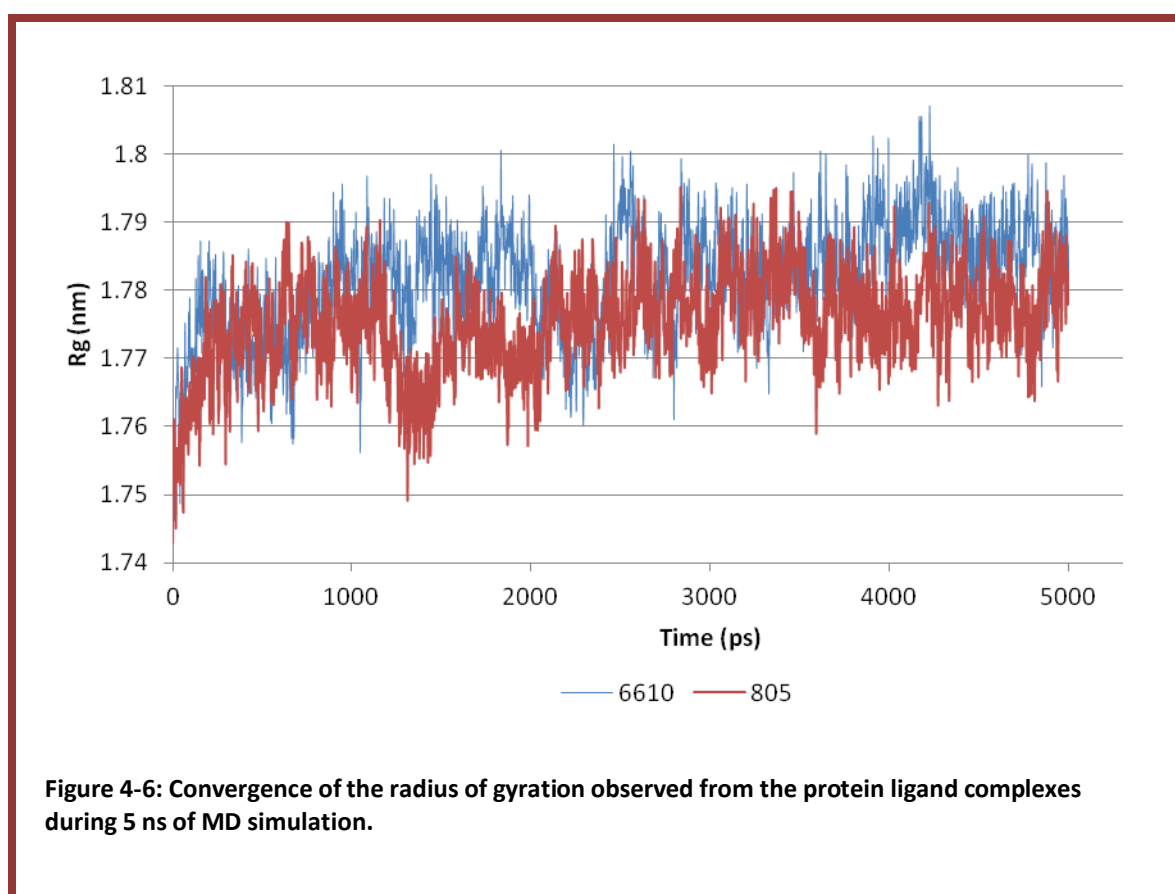


Figure 4-5: RMSD plot of the protein behaviour and the ligand interactions during 5 ns of MD simulation.

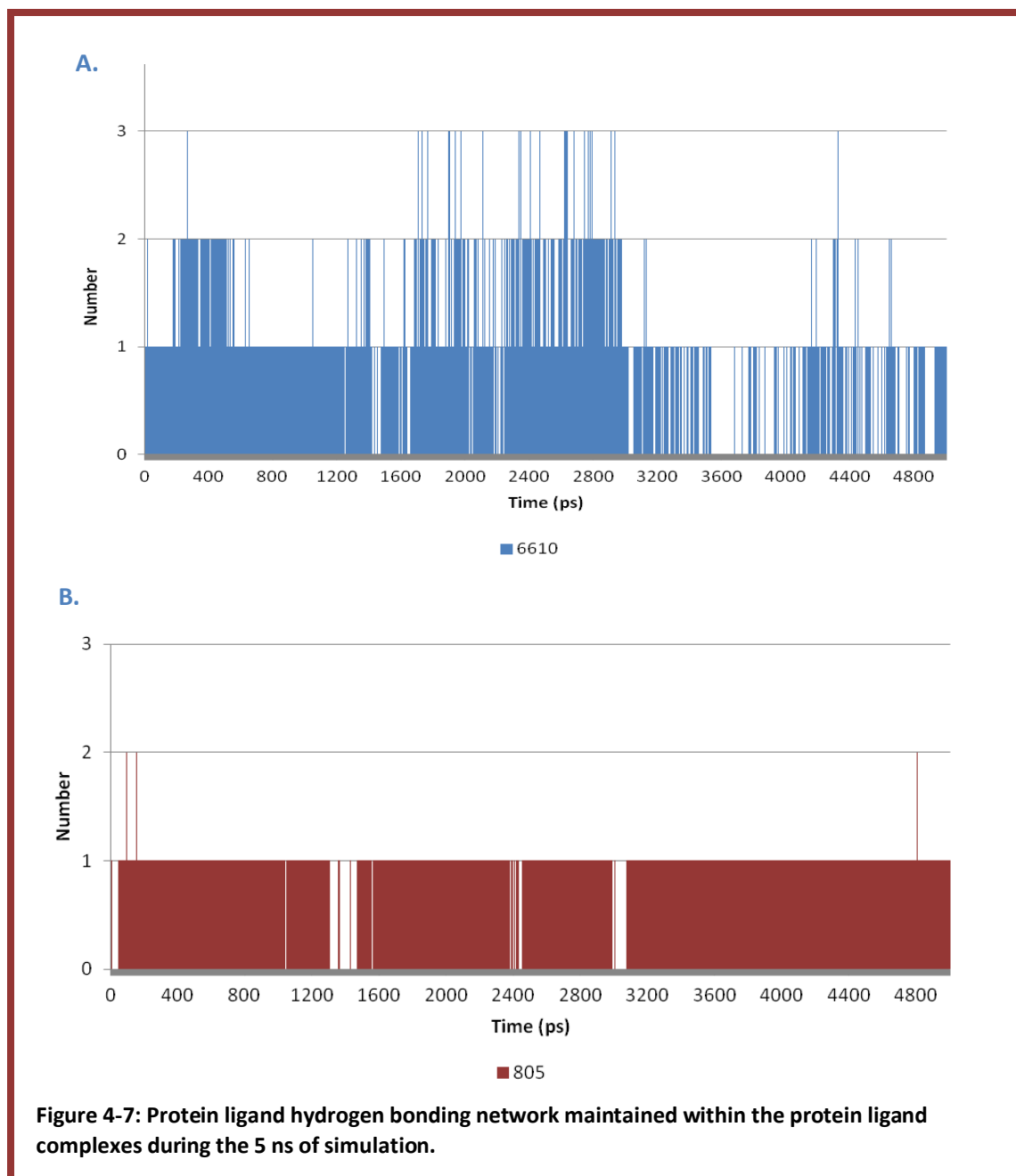
From Figure 4-5 protein in complex with ligand 6610 experiences an instability after 2.5 ns possibly associated with the relaxation of steric clashes that were introduced by the simulations. From Figure 4-5 B, ligand 805 has a larger deviation from its starting structure than ligand 6610. This deviation occurs sharply after 200 ps and is as a result of the ligand traversing an energy barrier allowing it to obtain a conformation with a lower energy that

was more stable than the binding mode identified by docking simulations. It is important to consider that docking simulations only allow the ligand to be flexible while the protein is maintained in a rigid manner. Thus the conformation maintained as the optimum binding mode applies only to the protein in that conformation. Flexibility and thus fluctuations of the protein as shown by the protein RMSD convergence, Figure 4-5 A, may introduce energetically unfavourable interactions with the ligand which will result in the abrupt change in conformation observed after 200 ps of simulation, Figure 4-5 B. A change in conformation is observed for the ligand 6610 and can be seen at 4.4 ns in the simulation. Based on the observations of the ligand 6610 change in RMSD after 4.4 ns, it would be advisable for production dynamics to proceed for longer than the 5 ns in order to ensure that the structure has accessed its equilibrium conformation within the protein active site. Evolution of the radius of gyration, Figure 4-6, is less drastic compared with the changes associated with atomic distances.



From Figure 4-6 we can see that there is a change of 0.04 nm in the predicted hydrodynamic radius for the protease in complex with ligand 805. The complex with ligand 6610 has a similar change although it appears to have a higher radial convergence of 1.79 nm while ligand 805 is converging at 1.78 nm.

Figure 4-7 highlights the difference between ligand 6610 and ligand 805 in the protein-ligand hydrogen bond network in terms of the number of bonds at specific time intervals.



From Figure 4-7 A we can see that a strong hydrogen bond network was established between ligand 6610 and the protease model between 1.2 ns and 3.2 ns. This network allowed 2 or 3 hydrogen bonds to persist between the ligand and the protein model. However, this apparent stability was not permanent and it was followed by a long term instability in terms of the hydrogen bonds forming between the ligand and the model. It is our hope that we would modify the lead candidate in such a way that these intermediate states can be selected for and maintained in order to improve the duration of potent binding and while minimizing unstable binding.

From 4-7 B, the ligand 805 complex rarely accesses intermediates that possess more than 1 hydrogen bond with the protein model. A consequence of this would be the observed lower binding energy over the duration of simulation and thus decreased binding affinity and potency of the lead candidate. It appears as though this single hydrogen bond is consistently maintained throughout the majority of the simulation, from inspection of the MD trajectory. Closer observations from the trajectory confirms the hydrogen bonding between the ligand 805 and protease model as being the same single hydrogen bond throughout the majority of the 5 ns of simulation. In comparison to the intermittent bond network observed in the 6610 model the periods of destabilisation are longer. Ligand 805 appears to be a better lead candidate than ligand 6610 despite having a higher predicted binding energy from docking. The hydrogen bonding data is summarized in Table 4-1.

Table 4-1: Hydrogen bonding network during 5 ns of simulation for the ligand 805 and ligand 6610 protease complex

Number of Hydrogen Bonds	Ligands	
	805	6610
0	285	960
1	2211	1155
2	3	358
3	0	27
Ave.	0.89	0.78

Table 4-1 in addition to the hydrogen bond network, it was found that ligand 805 has 3 binding modes that possess more than 1 hydrogen bond compared to ligand 6610 which

experienced more than 1 hydrogen bond in 385 modes during the 5 ns of simulation. In ligand 6610, 27 of these modes boasted 3 hydrogen bonds which could potentially be modified to stabilise this system. Despite this flattering number of bonds, on average across 5 ns of simulation, the average number of bonds in the ligand 805 complex was 0.89 which exceeded that for the ligand 6610 complex with 0.78 hydrogen bonds on average.

Calculations of binding free energies using molecular dynamics based calculations are computationally expensive. We implemented the MMPBSA approach at calculating the binding free energy for 50 ps of simulations for each of the 62 protease models with ligand 805 and ligand 6610 (Miller et al., 2012). Table 4-2 summarises the binding energies for ligand 805 and ligand 6610 with protease model 305152. This data was only extracted for the 2 cases due to time constraints and available processing resources.

Table 4-2: Binding energies of ligand 805 and ligand 6610 protease complexes from docking and MD.

Ligand	Method	
	<i>Vina</i>	<i>MD – MMPBSA</i>
805	-10.7	-33.94 kcal mol ⁻¹
6610	-11.3	-29.16 kcal mol ⁻¹

When the static binding energy is considered, ligand 6610 appears to be able to access a binding mode with a free energy of -11.3 kcal mol⁻¹ that is more stable than the binding mode accessible to ligand 805 with -10.7 kcal mol⁻¹. Over 50 ps of simulation when the solvent behaviour and receptor flexibility are accounted for, the average binding mode maintained within the ligand 805 complex has a free energy of binding of -34 kcal mol⁻¹. This valuable is more favourable than that obtained by the proposed lead candidate ligand 6610 with a binding energy of -29.1 kcal mol⁻¹.

It would be premature to state that ligand 805 is in actual fact a more potent lead candidate than ligand 6610 based on the evidence obtained. From Figure 4-5 and Figure 4-6 we can certainly see that the conformations obtained during the first 50 ps were not stable conformations. After 5 ns of simulation the complexes had not yet achieved their equilibrium conformations and thus estimating binding energies during this early stage

could be flawed. This calculation however represents the gross approximation of docking approaches to the calculation of binding energies. Although ligand 6610 does appear to possess stronger hydrogen bonding networks within the first 5 ns of simulation it is plausible to assume that its binding free energy might be superior to that of ligand 805.

Future studies that aim to resolve the binding stability of ligand 6610 would benefit from incorporating a polarised protein special charge obtained from utilising quantum mechanics levels of theory to solve for the solvent surface charges which estimates the electrostatic potential (Yang, Jiang, & Jiang, 2013).

The ligand interactions observed during the average conformations of the complexes are displayed in Figure 4-8. The average complexes show more ligand bonding interactions with the proteases complexes than the starting structures, which are conformations obtained from binding modes predicted by docking with *Vina*. The docking binding modes have more hydrophobic contacts than the average complexes.

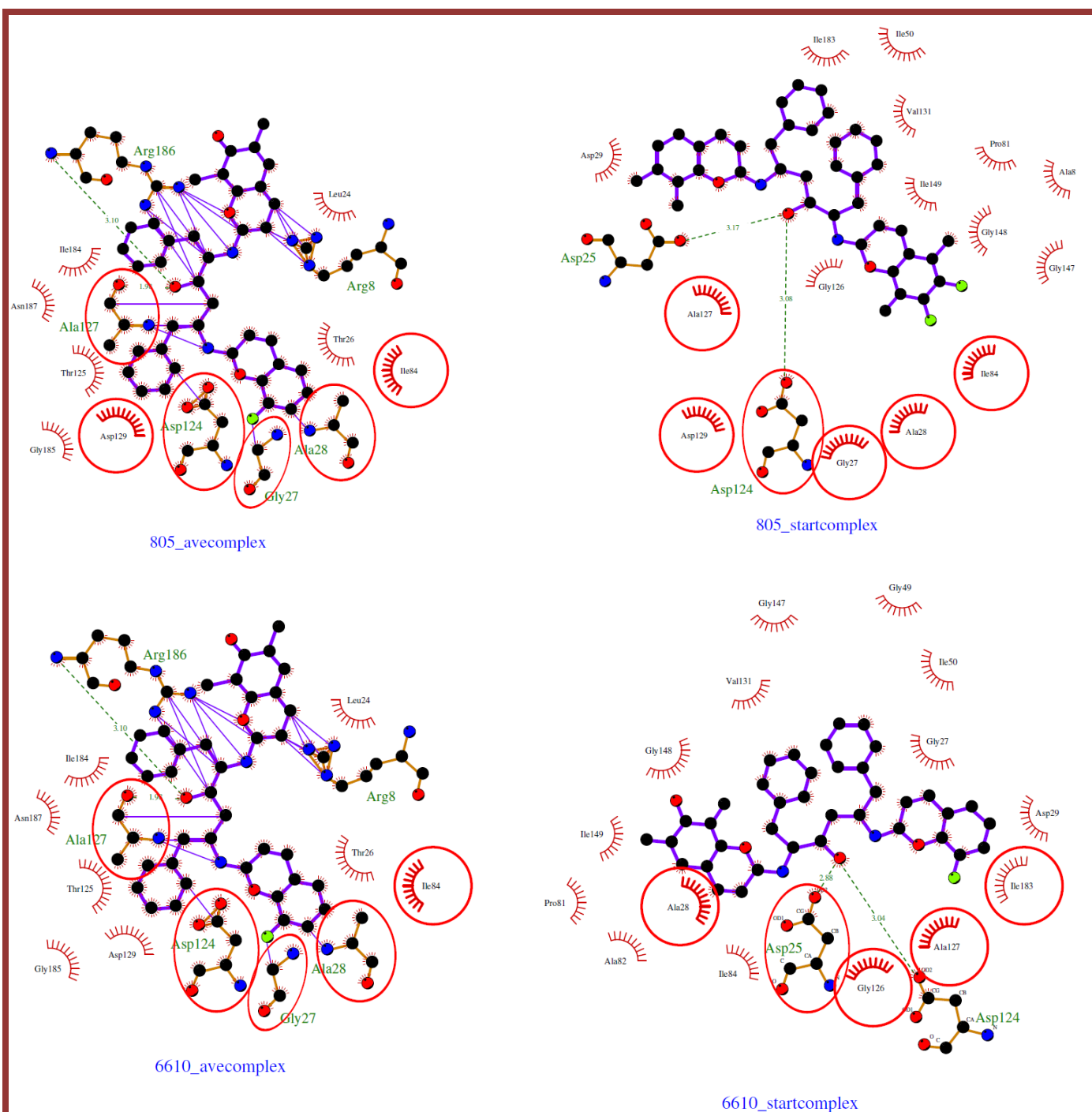


Figure 4-8: Ligplot diagrams of the binding modes obtained from Molecular Dynamics (average complex) and docking (start complex).

4.4. CONCLUSION

The binding stability of the lead candidate ligand 6610 and ligand 805 were compared and interrogated using MD simulations and MD based approaches to calculating the binding energies. GROMACS was used to produce 5 ns of protein-ligand complex simulations in the presence of water solvent. The system was prepared appropriately by neutralizing the charges and coupling the water and pressure to a water bath using equilibration and minimization steps.

Analysis of the simulations revealed that atoms within ligand 6610 experienced larger magnitudes of fluctuations than the ligand atoms within the complex with ligand 805 across the 5 ns simulation. An investigation into the hydrogen bonding networks maintained during the simulations revealed that ligand 6610 frequently accessed modes that possessed up to 3 hydrogen bonds. This many bonds were absent in the ligand 805 complex which only possessed 2 hydrogen bonds 3 times during the 5 ns of simulation. Despite these shortfalls on average throughout the simulation ligand 805 maintained more hydrogen bonds than ligand 6610. We could not perform a thorough investigation on the nature of the hydrogen bonds that were observed throughout the simulation because of time constraints.

When the binding energies of the first 50 ps of simulation were computed, ligand 805 appeared to have a better affinity for the protease model than ligand 6610. The binding energies appeared to be much lower than those observed from docking experiments. In order to further refine the computed binding energy incorporation of electrostatic potentials is anticipated to improve the precision in a later study. Experimental data would assist in the validation of the simulation approach, justifying the protocol we observed to select the lead candidate and interrogate its stability.

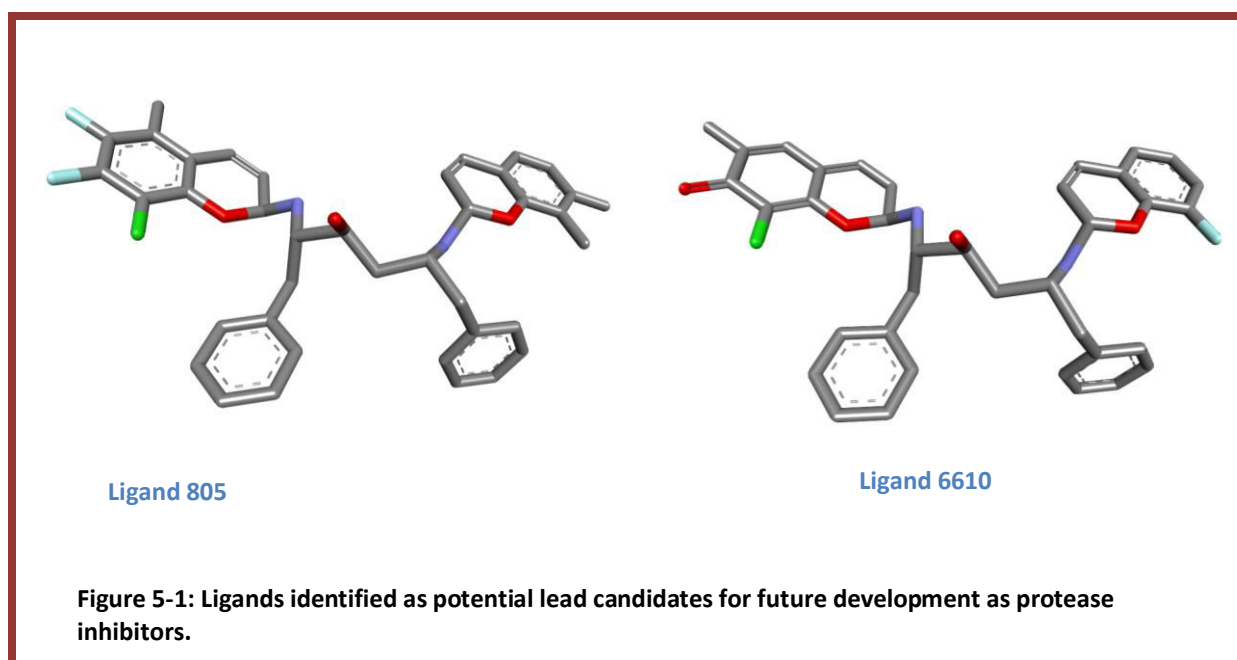
Chapter 5: Conclusion and Summary

Onywera et al. (2012) through an hit-identification experiment identified a *bis*-coumarin hit as a potential hit for optimisation as a protease inhibitor. This hit can be expanded into a library of analogues based on functionalised heterocycles. The Morita-Baylis-Hilman reaction gives rise to a variety of heterocyclic systems. By incorporating these MBH accessible heterocycles a wide diversity of analogues can be accessed. Synthesis of this vast array of analogues is nearly impossible given the size of the library and time constraints ever present in drug discovery pipelines.

Cheminformatic approaches allowed us to rapidly construct a virtual library of 17752 analogues based on the MBH accessible analogues. The Open Babel suite gave access to commands that were used to identify suitable sites and make necessary substitutions in a program that employed a markov chain recursive procedure. CheS-mapper 2.0 was utilized in order to analyze the chemical space being explored by the virtual library by employing up to 13 CDK toolkit chemical descriptors into a principle component analysis (PCA) that defined the 3D space coordinates. When the DrugBank database was coupled with a random sample of our virtual library it was evident from the clustering that we were undertaking an exhaustive search of analogues that possessed optimum potential as protease inhibitors.

In silico high-throughput virtual screening was employed in order to screen the virtual database for ligands that had strong affinity for the protease model dataset. HTVS of the 17750 ligands was performed by *Autodock Vina* on a cluster of 96 processors. Automated docking experiments were performed against a homology model built from the Stanford database Consensus C protease sequence in order to select the top 200 performing ligands. An exhaustive screen was employed over the entire model data set in order to reduce the 200 top performing ligands to 3 that could be visually inspected for their constitutional integrity and optimum pharmacological profile.

Ligand 805 and ligand 6610 in Figure 5-1 were identified as potential lead candidates based on a statistical comparison of their binding affinities across the entire data set of protease models.



Although binding affinities from docking experiments identified ligand 6610 as the superior lead candidate, molecular dynamic studies that analysed the stability of ligand-protein complexes showed that ligand 805 formed stable complexes. Protein and ligand atom fluctuations, protein and ligand deviations from the starting structures, hydrogen bonding network stabilities and binding free energy calculations confirmed this proposal.

Chemical synthesis of ligand 805 and ligand 6610 would allow *in vivo* analysis before medicinal chemistry techniques could be applied to optimise their pharmacokinetic properties such as solubility and metabolism.

In conclusion, our approach allowed us to rapidly diversify the *bis*-coumarin hit in order to further search the chemical space within its vicinity before high-throughput screening was used to identify a few lead candidates. Simulations were incorporated in order to further analyse the ligand-protein complexes.

References

- Abbas, W., & Herbein, G. (2012). Molecular Understanding of HIV-1 Latency. *Advances in Virology*, 2012, 574967. doi:10.1155/2012/574967
- ACS. (2015). Chemical Substances - CAS REGISTRY. Retrieved January 20, 2015, from <http://www.cas.org/content/chemical-substances>
- Adcock, S. A., & McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical Reviews*, 106, 1589–1615.
- Ahmed, S. M., Kruger, H. G., Govender, T., Maguire, G. E. M., Sayed, Y., Ibrahim, M. a a, ... Soliman, M. E. S. (2013). Comparison of the molecular dynamics and calculated binding free energies for nine FDA-approved HIV-1 PR drugs against subtype B and C-SA HIV PR. *Chemical Biology & Drug Design*, 81(2), 208–18. doi:10.1111/cbdd.12063
- Aitken, S. C., Bronze, M., Wallis, C. L., Stuyver, L., Steegen, K., Balinda, S., ... Schuurman, R. (2013). A pragmatic approach to HIV-1 drug resistance determination in resource-limited settings by use of a novel genotyping assay targeting the reverse transcriptase-encoding region only. *Journal of Clinical Microbiology*, 51(6), 1757–61. doi:10.1128/JCM.00118-13
- Al-Khindi, T., Zakzanis, K. K., & van Gorp, W. G. (2011). Does antiretroviral therapy improve HIV-associated cognitive impairment? A quantitative review of the literature. *Journal of the International Neuropsychological Society : JINS*, 17(6), 956–69. doi:10.1017/S1355617711000968
- Amidon, G. L., Lennernäs, H., Shah, V. P., & Crison, J. R. (1995). A theoretical basis for a biopharmaceutic drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharmaceutical Research*, 12(3), 413–20. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/7617530>
- Babine, R. E., & Bender, S. L. (1997). Molecular Recognition of Protein – Ligand Complexes: Applications to Drug Design, 2665(96).
- Bai, Y., Xue, H., Wang, K., Cai, L., Qiu, J., Bi, S., ... Liu, K. (2013). Covalent fusion inhibitors targeting HIV-1 gp41 deep pocket. *Amino Acids*, 44(2), 701–13. doi:10.1007/s00726-012-1394-8
- Baylis, A., & Hillman, M. (1972). German Patent 2155113. *Chemical Abstracts*, 77, 34174q.
- 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), 43–56.
- Bourne, Y., Talley, T. T., Hansen, S. B., Taylor, P., & Marchot, P. (2005). Crystal structure of a Cbtx-AChBP complex reveals essential interactions between snake alpha-neurotoxins and nicotinic receptors. *The EMBO Journal*, 24(8), 1512–22. doi:10.1038/sj.emboj.7600620

- Brik, A., & Wong, C.-H. (2003). HIV-1 protease: mechanism and drug discovery. *Organic & Biomolecular Chemistry*, 1(1), 5–14. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12929379>
- Cai, Y., Yilmaz, N. K., Myint, W., Ishima, R., & Schiffer, C. a. (2012). Differential Flap Dynamics in Wild-type and a Drug Resistant Variant of HIV-1 Protease Revealed by Molecular Dynamics and NMR Relaxation. *Journal of Chemical Theory and Computation*, 8(10), 3452–3462. doi:10.1021/ct300076y
- Cambiano, V., Bertagnolio, S., Jordan, M. R., Pillay, D., Perriens, J. H., Venter, F., ... Phillips, A. (2014). Predicted levels of HIV drug resistance: potential impact of expanding diagnosis, retention, and eligibility criteria for antiretroviral therapy initiation. *AIDS (London, England)*, 28 Suppl 1, S15–23. doi:10.1097/QAD.0000000000000082
- Chang, C., Trylska, J., Tozzini, V., & McCammon, J. A. (2007). Binding pathways of ligands to HIV-1 protease: coarse-grained and atomistic simulations. *Chemical Biology & Drug Design*, 69(1), 5–13. doi:10.1111/j.1747-0285.2007.00464.x
- Chang, M. W., Ayeni, C., Breuer, S., & Torbett, B. E. (2010). Virtual screening for HIV protease inhibitors: A comparison of AutoDock 4 and Vina. *PLoS ONE*, 5(8), 1–9. doi:10.1371/journal.pone.0011955
- Coman, R. M., Robbins, A. H., Fernandez, M. a, Gilliland, C. T., Sochet, A. a, Goodenow, M. M., ... Dunn, B. M. (2008). The contribution of naturally occurring polymorphisms in altering the biochemical and structural characteristics of HIV-1 subtype C protease. *Biochemistry*, 47(2), 731–43. doi:10.1021/bi7018332
- Damond, F., Worobey, M., Campa, P., Farfara, I., Colin, G., Matheron, S., & Robertson, D. L. (2004). Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification. *AIDS Research and Human Retroviruses*, 20(6), 666–672.
- Das, K., & Arnold, E. (2013). HIV-1 reverse transcriptase and antiviral drug resistance. Part 1. *Current Opinion in Virology*, 3(2), 111–8. doi:10.1016/j.coviro.2013.03.012
- De Cock, K. M., Jaffe, H. W., & Curran, J. W. (2012). The evolving epidemiology of HIV/AIDS. *AIDS*, 26, 1205 – 1213.
- Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology*, 9(1), 71. doi:10.1186/1741-7007-9-71
- Fenner, L., Reid, S. E., Fox, M. P., Garone, D., Wellington, M., Prozesky, H., ... Egger, M. (2013). Tuberculosis and the risk of opportunistic infections and cancers in HIV-infected patients starting ART in Southern Africa. *Tropical Medicine & International Health : TM & IH*, 18(2), 194–8. doi:10.1111/tmi.12026
- Feynman, R. (1985). *QED: The strange Theory of Light and Matter*. Princeton, NJ: Princeton University Press.
- Foulkes, J. E., Prabu-Jeyabalan, M., Cooper, D., Henderson, G. J., Harris, J., Swanstrom, R., & Schiffer, C. A. (2006). Role of invariant Thr80 in human immunodeficiency virus type 1 protease

- structure, function, and viral infectivity. *Journal of Virology*, 80(14), 6906–16. doi:10.1128/JVI.01900-05
- Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., ... Hahn, B. H. (1999). Origin of HIV-1 in the chimpanzee *Pan troglodytes*. *Nature*, 397(6718), 436–41. doi:10.1038/17130
- Gerschenson, M., Kim, C., Berzins, B., Taiwo, B., Libutti, D. E., Choi, J., ... Murphy, R. L. (2009). Mitochondrial function, morphology and metabolic parameters improve after switching from stavudine to a tenofovir-containing regimen. *The Journal of Antimicrobial Chemotherapy*, 63(6), 1244–50. doi:10.1093/jac/dkp100
- Goodsell, D. S., Morris, G. M., & Olson, A. J. (1996). Automated docking of flexible ligands: applications of AutoDock. *Journal of Molecular Recognition : JMR*, 9(1), 1–5. doi:10.1002/(SICI)1099-1352(199601)9:1<1::AID-JMR241>3.0.CO;2-6
- Gütlein, M., Karwath, A., & Kramer, S. (2014). CheS-Mapper 2.0 for visual validation of (Q) SAR models. *Journal of Cheminformatics*, 6(41), 1–18.
- Hamers, R. L., Wallis, C. L., Kityo, C., Siwale, M., Mandaliya, K., Conradie, F., ... de Wit, T. F. R. (2011). HIV-1 drug resistance in antiretroviral-naïve individuals in sub-Saharan Africa after rollout of antiretroviral therapy: a multicentre observational study. *The Lancet Infectious Diseases*, 11(10), 750–9. doi:10.1016/S1473-3099(11)70149-9
- Hann, M. M., & Oprea, T. I. (2004). Pursuing the leadlikeness concept in pharmaceutical research. *Current Opinion in Chemical Biology*, 8, 255–263. doi:10.1016/j.cbpa.2004.04.003
- Hazuda, D. J., Anthony, N. J., Gomez, R. P., Jolly, S. M., Wai, J. S., Zhuang, L., ... Emini, E. (2004). A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 11233–11238.
- Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends in Molecular Medicine*, 18(3), 182–92. doi:10.1016/j.molmed.2011.12.001
- Hemelaar, J., Gouws, E., Ghys, P. D., & Osmanov, S. (2006). Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS (London, England)*, 20(16), W13–23. doi:10.1097/01.aids.0000247564.73009.bc
- Hou, T. J., & Xu, X. J. (2003). ADME Evaluation in Drug Discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors. *Journal of Chemical Information and Computer Sciences*, 43, 2137–2152. doi:10.1021/ci034134i
- Hou, T., Wang, J., Li, Y., & Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods: I. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Comput. Sci.*, 51(1), 69–82. doi:10.1021/ci100275a
- ImageMagickStudio. (2015). ImageMagick. Retrieved February 03, 2015, from <http://imagemagick.org/index.php>

- Johnson, A. A., Ray, A. S., Hanes, J., Suo, Z., Colacino, J. M., Anderson, K. S., & Johnson, K. A. (2001). Toxicity of antiviral nucleoside analogs and the human mitochondrial DNA polymerase. *The Journal of Biological Chemistry*, 276(44), 40847–40857. doi:10.1074/jbc.M106743200
- Johnson, V., Calvez, V., Gunthard, H. F., Paredes, R., Pillay, D., Shafer, R. W., ... Richman, D. D. (2013). Update of the drug resistance mutations in HIV-1: March 2013. *Topics in Antiviral Medicine*, 21(1), 6–14. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23596273>
- Kim, J. T., Hamilton, A. D., Bailey, C. M., Domaoal, R. A., Wang, L., Anderson, K. S., & Jorgensen, W. L. (2006). FEP-guided selection of bicyclic heterocycles in lead optimization for non-nucleoside inhibitors of HIV-1 reverse transcriptase. *J Am Chem Soc*, 128, 15372–15373.
- Kipp, D. R., Hirschi, J. S., Wakata, A., Goldstein, H., & Schramm, V. L. (2012). Transition states of native and drug-resistant HIV-1 protease are the same. *Proceedings of the National Academy of Sciences of the United States of America*, 109(17), 6543–8. doi:10.1073/pnas.1202808109
- Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., & Detours, V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *British Medical Bulletin*, 58, 19–42. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11714622>
- Kosakovsky Pond, S. L., & Smith, D. M. (2009). Are all subtypes created equal? The effectiveness of antiretroviral therapy against non-subtype B HIV-1. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, 48(9), 1306–9. doi:10.1086/598503
- Kozal, M. J. (2009). Drug-resistant human immunodeficiency virus. *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, 15 Suppl 1, 69–73. doi:10.1111/j.1469-0691.2008.02687.x
- Kuhn, B., Mohr, P., & Stahl, M. (2010). Intramolecular hydrogen bonding in medicinal chemistry. *Journal of Medicinal Chemistry*, 53, 2601–2611. doi:10.1021/jm100087s
- Kumar, A., & Herbein, G. (2014). The macrophage: a therapeutic target in HIV-1 infection. *Molecular and Cellular Therapies*, 2(10), 1–15.
- Lawn, S. D., Török, M. E., & Wood, R. (2011). Optimum time to start antiretroviral therapy during HIV-associated opportunistic infections. *Current Opinion in Infectious Diseases*, 24(1), 34–42. doi:10.1097/QCO.0b013e3283420f76
- Lederberg, J. (1965). TOPOLOGICAL MAPPING OF ORGANIC MOLECULES. *Proceedings of the National Academy of Sciences of the United States of America*, 53(1), 134–9. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=219445&tool=pmcentrez&rendertype=abstract>
- Lemey, P., Pybus, O. G., Rambaut, A., Drummond, A. J., Robertson, D. L., Roques, P., ... Vandamme, A.-M. (2004). The molecular population genetics of HIV-1 group O. *Genetics*, 167(3), 1059–68. doi:10.1534/genetics.104.026666
- Lever, A. M. L., & Berkhout, B. (2008). 2008 Nobel prize in medicine for discoverers of HIV. *Retrovirology*, 5(1966), 91. doi:10.1186/1742-4690-5-91

- Lieberman-Blum, S. S., Fung, H. B., & Bandres, J. C. (2008). Maraviroc: a CCR5-receptor antagonist for the treatment of HIV-1 infection. *Clinical Therapeutics*, 30(7), 1228–50. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/18691983>
- Lipinski, C. a., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 64, 4–17. doi:10.1016/j.addr.2012.09.019
- Louis, J. M., Aniana, A., Weber, I. T., & Sayer, J. M. (2011). Inhibition of autoprocessing of natural variants and multidrug resistant mutant precursors of HIV-1 protease by clinical inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 108(22), 9072–7. doi:10.1073/pnas.1102278108
- Medina, D. J., Tsai, C. H., Hsiung, G. D., & Cheng, Y. C. (1994). Comparison of mitochondrial morphology, mitochondrial DNA content, and cell viability in cultured cells treated with three anti-human immunodeficiency virus dideoxynucleosides. *Antimicrobial Agents and Chemotherapy*, 38(8), 1824–1828. doi:10.1128/AAC.38.8.1824
- Medina-Franco, J. L., Giulianotti, M. a., Welmaker, G. S., & Houghten, R. a. (2013). Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discovery Today*, 18(May), 495–501. doi:10.1016/j.drudis.2013.01.008
- Métifiot, M., Marchand, C., & Pommier, Y. (2013). HIV integrase inhibitors: 20-year landmark and challenges. *Advances in Pharmacology (San Diego, Calif.)*, 67, 75–105. doi:10.1016/B978-0-12-405880-4.00003-2
- Miller, B. R., Mcgee, T. D., Swails, J. M., Homeyer, N., Gohlke, H., & Roitberg, A. E. (2012). MMPBSA.py : An Efficient Program for End-State Free Energy Calculations.
- Morita, K., Suzuki, Z., & Hirose, H. (1968). A tertiary Phosphine-catalyzed reaction of acrylic compounds with aldehydes. *Bulletin of The Chemical Society of Japan*, 41(11), 2815–2815. doi:10.1246/bcsj.41.2815
- Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16), 2785–91. doi:10.1002/jcc.21256
- Mosebi, S., Morris, L., Dirr, H. W., & Sayed, Y. (2008). Active-site mutations in the South african human immunodeficiency virus type 1 subtype C protease have a significant impact on clinical inhibitor binding: kinetic and thermodynamic study. *Journal of Virology*, 82(22), 11476–9. doi:10.1128/JVI.00726-08
- Murray, J. M., Kelleher, a. D., & Cooper, D. a. (2011). Timing of the Components of the HIV Life Cycle in Productively Infected CD4+ T Cells in a Population of HIV-Infected Individuals. *Journal of Virology*, 85(20), 10798–10805. doi:10.1128/JVI.05095-11
- Nakimuli-Mpungu, E., Bass, J. K., Alexandre, P., Mills, E. J., Musisi, S., Ram, M., ... Nachega, J. B. (2012). Depression, alcohol use and adherence to antiretroviral therapy in sub-Saharan Africa: a systematic review. *AIDS and Behavior*, 16(8), 2101–18. doi:10.1007/s10461-011-0087-8

- Nicolaou, K. C. (2014). Advancing the Drug Discovery and Development Process. *Angewandte Chemie (International Ed. in English)*, n/a–n/a. doi:10.1002/anie.201404761
- NIH. A reference guide for prescription HIV-1 medications. , Pub. L. No. 14-7628 (2014).
- Nyoni, D., Lobb, K. A., & Kaye, P. T. (2013). Dual-Catalyst Acceleration of Tandem Disulfide Cleavage and Baylis–Hillman Synthesis of 2 H -1-Benzothiopyran Derivatives. *Synthetic Communications*, 43(13), 1837–1841. doi:10.1080/00397911.2012.673449
- 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), 33. doi:10.1186/1758-2946-3-33
- Onywera, H., Lobb, K., & Tastan Bishop, O. (2012). Influence of Non-Synonymous Sequence Mutations on the Architecture of HIV-1 Clade C Protease Receptor Site: Docking and Molecular Dynamics Studies. *MSc Thesis*, (December), 1–164.
- Ott, M., & Verdin, E. (2013). Three rules for HIV latency: location, location, and location. *Cell Host & Microbe*, 13(6), 625–6. doi:10.1016/j.chom.2013.05.016
- Palmer, S., Alaeus, A., Albert, J., & Cox, S. (1998). Drug Susceptibility of Subtypes A, B, C, D and E Human Immunodeficiency Virus Type 1 Primary Isolates. *AIDS Research and Human Retroviruses*, 14(2), 157–162.
- Park, H., Lee, J., & Lee, S. (2006). Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins*, 65(3), 549–54. doi:10.1002/prot.21183
- Peat, T. S., Dolezal, O., Newman, J., Mobley, D., & Deadman, J. J. (2014). Interrogating HIV integrase for compounds that bind - A SAMPL challenge. *Journal of Computer-Aided Molecular Design*, 28, 347–362. doi:10.1007/s10822-014-9721-7
- Peng, J., Huang, X., Jiang, L., Cui, H.-L., & Chen, Y.-C. (2011). Tertiary amine-catalyzed chemoselective and asymmetric [3 + 2] annulation of Morita-Baylis-Hillman carbonates of isatins with propargyl sulfones. *Organic Letters*, 13(17), 4584–7. doi:10.1021/ol201776h
- PhRMA. (2013). 2013 Biopharmaceutical Research Industry Profile. *Pharmaceutical Research and Manufacturers of America*.
- Plika, V., Testa, B., & van de Waterbeemd, H. (Eds.). (1996). *Lipophilicity in Drug Action and Toxicology*. Weinheim, Germany: Wiley-VCH Verlag GmbH. doi:10.1002/9783527614998
- Qian, X., Xu, X., Li, Z., & Frontera, A. (2003). Intramolecular noncovalent force in cyclic amidines: Nonbonded interaction between carbon atoms and heteroatoms. *Chemical Physics Letters*, 372, 489–496. doi:10.1016/S0009-2614(03)00395-6
- Rahman, A. (1964). Correlations in the Motion of Atoms in Liquid Argon. *Physical Review*, 136(2A), A405–A411. doi:10.1103/PhysRev.136.A405
- Reymond, J. L., & Awale, M. (2012). Exploring chemical space for drug discovery using the chemical universe database. *ACS Chemical Neuroscience*, 3, 649–657. doi:10.1021/cn3000422

- Rhee, S.-Y., Kantor, R., Katzenstein, D. a, Camacho, R., Morris, L., Sirivichayakul, S., ... Shafer, R. W. (2006). HIV-1 pol mutation frequency by subtype and treatment experience: extension of the HIVseq program to seven non-B subtypes. *AIDS (London, England)*, 20(5), 643–651.
- Ruddigkeit, L., Van Deursen, R., Blum, L. C., & Reymond, J. L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52, 2864–2875. doi:10.1021/ci300415d
- Sainski, A., Cummins, N., & Badley, A. (2014). HIV Life Cycle - HIV Clinical Manual. *Antimicrobe*. Retrieved January 15, 2015, from <http://www.antimicrobe.org/hiv02.asp>
- Schames, J. R., Henchman, R. H., Siegel, J. S., Sotriffer, C. A., Ni, H., & McCammon, J. A. (2004). Discovery of a novel binding trench in HIV integrase. *Journal of Medicinal Chemistry*, 47, 1879–1881.
- Schramm, V. L. (2013). Transition states, analogues, and drug development. *ACS Chemical Biology*, 8, 71–81. doi:10.1021/cb300631k
- Schwab, F., & van Gunsteren, W. F. Zagrovic, B. (2008). Computational study of the mechanism and the relative free energies of binding of anticholesteremic inhibitors to squalene-hopene cyclase. *Biochemistry*, 47, 2945 – 2951.
- Sepkowitz, K. (2001). AIDS - The First 20 years. *The New England Journal of Medicine*, 344(23), 1764–1772.
- Shafer, R. W. (2006). Rationale and uses of a public HIV drug-resistance database. *The Journal of Infectious Diseases*, 194 Suppl , S51–S58.
- Sharp, P. M., & Hahn, B. H. (2010). The evolution of HIV-1 and the origin of AIDS. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 365(1552), 2487–94. doi:10.1098/rstb.2010.0031
- Shen, C. H., Tie, Y., Yu, X., Wang, Y. F., Kovalevsky, A. Y., Harrison, R. W., & Weber, I. T. (2012). Capturing the reaction pathway in near-atomic-resolution crystal structures of HIV-1 protease. *Biochemistry*, 51, 7726–7732. doi:10.1021/bi3008092
- Silva, S. Da, Alan, W., & Vranken, W. F. (2012). ACPYPE - AnteChamber PYthon Parser interface, 1–8. doi:10.1186/1756-0500-5-367
- Sinko, W., Lindert, S., & Mccammon, J. A. (2013). Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chemical Biology and Drug Design*, 81, 41–49. doi:10.1111/cbdd.12051
- Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. (2014). Computational Methods in Drug Discovery. *Pharmacological Reviews*, 66(January), 334–395. doi:10.1016/j.vascn.2010.02.005
- Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacological Reviews*, 66(1), 334–95. doi:10.1124/pr.112.007336

- Stanley, S. a, Grant, S. S., Kawate, T., Iwase, N., Shimizu, M., Wivagg, C., ... Hung, D. T. (2012). Identification of Novel Inhibitors of M. tuberculosis Growth Using Whole Cell Based High-Throughput Screening.
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43, 493–500. doi:10.1021/ci025584y
- Tatem, A. J., Hemelaar, J., Gray, R. R., & Salemi, M. (2012). Spatial accessibility and the spread of HIV-1 subtypes and recombinants. *AIDS (London, England)*, 26(18), 2351–60. doi:10.1097/QAD.0b013e328359a904
- Tebit, D. M., & Arts, E. J. (2011). Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *The Lancet Infectious Diseases*, 11(1), 45–56. doi:10.1016/S1473-3099(10)70186-9
- Teni. (2014, January 1). Immunological and clinical progress of HIV/AIDS patients on antiretroviral therapy at a health center in Addis Ababa, Ethiopia. *Archives of Pharmacy Practice*. Medknow Publications and Media Pvt. Ltd. doi:10.4103/2045-080X.128371
- Testa, B., Crivori, P., Reist, M., & Carrupt, P. a. (2000). The influence of lipophilicity on the pharmacokinetic behavior of drugs: Concepts and examples. *Perspectives in Drug Discovery and Design*, 19, 179–211. doi:10.1023/A:1008741731244
- Thompson, M. A., Aberg, J. A., Cahn, P., Montaner, J. S. G., Rizzardini, G., Telenti, A., ... Schooley, R. T. (2010). Antiretroviral Treatment of Adult HIV Infection 2010. *The Journal of American Medical Association*, 304(3).
- 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, 455–461. doi:10.1002/jcc
- Trylska, J., Tozzini, V., Chang, C., & McCammon, J. A. (2007). HIV-1 protease substrate binding and product release pathways explored with coarse-grained molecular dynamics. *Biophysical Journal*, 92(12), 4179–87. doi:10.1529/biophysj.106.100560
- Usach, I., Melis, V., & Peris, J.-E. (2013). Non-nucleoside reverse transcriptase inhibitors: a review on pharmacokinetics, pharmacodynamics, safety and tolerability. *Journal of the International AIDS Society*, 16(1), 1–14. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3764307&tool=pmcentrez&rendertype=abstract>
- Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45, 2615–2623. doi:10.1021/jm020017n
- Walmsley, S. (2007). Protease inhibitor-based regimens for HIV therapy: safety and efficacy. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 45 Suppl 1, S5–13; quiz S28–31. doi:10.1097/QAI.0b013e3180600709

- Wang, R., Fu, Y., & Lai, L. (1997). A New Atom-Additive Method for Calculating Partition Coefficients. *Journal of Chemical Information and Computer Sciences*, 2338(96), 615–621. doi:10.1021/ci960169p
- Wei, Y., & Shi, M. (2013). Recent advances in organocatalytic asymmetric Morita-Baylis-Hillman/aza-Morita-Baylis-Hillman reactions. *Chemical Reviews*, 113(8), 6659–90. doi:10.1021/cr300192h
- Wensing, A. M., Calvez, V., Günthard, H. F., Johnson, V. a, Paredes, R., Pillay, D., ... Richman, D. D. (2014). 2014 update of the drug resistance mutations in HIV-1. *Topics in Antiviral Medicine*, 22, 642–50. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/25101529>
- Wensing, A., van Maarseveen, N. M., & Nijhuis, M. (2010). Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Research*, 85(1), 59–74. doi:10.1016/j.antiviral.2009.10.003
- Wereszczynski, J., & McCammon, J. A. (2012). Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. *Quarterly Reviews of Biophysics*, 45(1), 1–25. doi:10.1017/S0033583511000096
- WHO. (2014). Global Health Observatory. *HIV/AIDS:Global situation and trends*.
- Williams, S. a, & Greene, W. C. (2007). Regulation of HIV-1 latency by T-cell activation. *Cytokine*, 39(1), 63–74. doi:10.1016/j.cyto.2007.05.017
- Yang, M., Jiang, X., & Jiang, N. (2013). Protonation state and free energy calculation of HIV-1 protease-inhibitor complex based on electrostatic polarisation effect. *Molecular Physics*, 112(February 2015), 1659–1669. doi:10.1080/00268976.2013.857050
- Zhang, S. (2011). Computer-aided drug discovery and development. *Methods in Molecular Biology (Clifton, N.J.)*, 716, 23–38. doi:10.1007/978-1-61779-012-6_2
- Zhao, M., Wei, Y., & Shi, M. (2011). *The Chemistry of the Morita-Baylis-Hillman Reaction*. (J. Spivey, Ed.) (RSC Cataly.). Cambridge: The Royal Society of Chemistry. Retrieved from <http://pubs.rsc.org/en/content/chapter/bk9781849731294-00001/978-1-84973-129-4#!divabstract>

Appendix

APPENDIX A – SCRIPTS

1. Library Construction (Authors K. Lobb and L. Sigauke)

```
#include "multiple.h"

MoleculeList corelist;
MoleculeList backbonelist;
MoleculeList layerl1list;

//MoleculeIterator corelistiterator;
//MoleculeIterator backbonelistiterator;
//MoleculeIterator layerl1listiterator;

////////////////////////////////////
////
////////////////////////////////////
////
////////////////////////////////////
////
class result
{
//this class contains a molecule and an array telling about the Hydrogens
public:
    result(){resetall();};
    ~result(){};

    OpenBabel::OBMol themolecule;
    int accounting[1000];

    void setacc(int value,int index){accounting[index]=value;}
    int getacc(int index){return accounting[index];}
    void resetall();
    int count(int level);
    void print();
    result & operator= (result &other);
    void add(OpenBabel::OBMol &other, int hydrogenatomtodelete, int level);
    void save();
};

////////////////////////////////////
////
////////////////////////////////////
////
////////////////////////////////////
////
////////////////////////////////////
////

void fix(result thisresult,int depth)
{
    int number1 = thisresult.count(1);
    int number2 = thisresult.count(2);

    if(number1>0)
    {
        int theatom=0;
        for(OpenBabel::OBMolAtomIter a(thisresult.themolecule);a;++a)
        {
            if((*a).IsHydrogen())&&(thisresult.getacc(theatom)==1)
            {
                MoleculeIterator corelistiterator;
                corelistiterator = corelist.begin();
            }
        }
    }
}
```

```

        while(corelistiterator != corelist.end())
        {
            OpenBabel::OBMol newsubst>(*corelistiterator);
            result newresult;
            newresult=thisresult;
            newresult.add(newsubst,theatom,2);
            fix(newresult,depth+1);
            corelistiterator++;
        }
        theatom++;
    }
}
else if(number2>0)
{
    int theatom=0;
    for(OpenBabel::OBMolAtomIter a(thisresult.themolecule);a;++a)
    {
        if((*a).IsHydrogen())&&(thisresult.getacc(theatom)==2)
        {
            MoleculeIterator layerliterator;
            layerliterator = layerliterator.begin();
            while(layerliterator != layerliterator.end())
            {
                OpenBabel::OBMol newsubst(*layerliterator);
                result newresult;
                newresult=thisresult;
                newresult.add(newsubst,theatom,3);
                fix(newresult,depth+1);
                layerliterator++;
            }
            theatom++;
        }
    }
else
{
    for(OpenBabel::OBMolAtomIter a(thisresult.themolecule);a;++a)
    {
        if((*a).IsHydrogen())
        {
            thisresult.themolecule.DeleteAtom(&*a);
        }
    }
    thisresult.themolecule.PerceiveBondOrders();
    //thisresult.themolecule.ConnectTheDots();
    thisresult.themolecule.DeleteHydrogens();
    thisresult.themolecule.AddHydrogens();
    thisresult.themolecule.PerceiveBondOrders();
    thisresult.themolecule.ConnectTheDots();

    //thisresult.themolecule.AddHydrogens();

    thisresult.print();
    thisresult.save();
}
}

int main()
{
    OpenBabel::OBConversion obconversion;
    OpenBabel::OBMol core1,core2,core3,core4,core5,core6,core7;
    OpenBabel::OBMol backbone2,backbone3,backbone4;
    OpenBabel::OBMol
    layer11,layer12,layer13,layer14,layer15,layer16,layer17,layer18,layer19;

```

```

obconversion.SetInFormat("xyz");
obconversion.ReadFile(&core1,"project/core/chromenes_core.xyz");
obconversion.ReadFile(&core2,"project/core/indolizine_A_core.xyz");
obconversion.ReadFile(&core3,"project/core/quinolone_core.xyz");
obconversion.ReadFile(&core4,"project/core/coumarin_core.xyz");
obconversion.ReadFile(&core5,"project/core/indolizine_core.xyz");
obconversion.ReadFile(&core6,"project/core/thiochromene_core.xyz");
obconversion.ReadFile(&core7,"project/core/quinoline_core.xyz");

obconversion.ReadFile(&backbone2,"project/backbone/2_chain.xyz");
obconversion.ReadFile(&backbone3,"project/backbone/3_chain.xyz");
obconversion.ReadFile(&backbone4,"project/backbone/4_chain.xyz");

obconversion.ReadFile(&layer11,"project/layer1/BrH.xyz");
obconversion.ReadFile(&layer12,"project/layer1/CCOH.xyz");
obconversion.ReadFile(&layer13,"project/layer1/ClH.xyz");
obconversion.ReadFile(&layer14,"project/layer1/FH.xyz");
obconversion.ReadFile(&layer15,"project/layer1/OH.xyz");
obconversion.ReadFile(&layer16,"project/layer1/CCH.xyz");
obconversion.ReadFile(&layer17,"project/layer1/CH.xyz");
obconversion.ReadFile(&layer18,"project/layer1/COH.xyz");
obconversion.ReadFile(&layer19,"project/layer1/HH.xyz");

// corelist
corelist.push_back(core1);
corelist.push_back(core2);
corelist.push_back(core3);
corelist.push_back(core4);
corelist.push_back(core5);
corelist.push_back(core6);
corelist.push_back(core7);
// backbonelist
backbonelist.push_back(backbone2);
backbonelist.push_back(backbone3);
backbonelist.push_back(backbone4);
// layer1list
layer1list.push_back(layer11);
layer1list.push_back(layer12);
layer1list.push_back(layer13);
layer1list.push_back(layer14);
layer1list.push_back(layer15);
layer1list.push_back(layer16);
layer1list.push_back(layer17);
layer1list.push_back(layer18);
layer1list.push_back(layer19);

result resultinitial;

resultinitial.resetall();
resultinitial.add(backbone2,0,1);
resultinitial.print();
fix(resultinitial,1);
}

////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////
////////////////////////////////////

```

```

void result::save()
{
    OpenBabel::OBConversion obconversion;
    obconversion.SetOutFormat("smi");
    std::stringstream buffer;
    obconversion.Write(&themolecule,&buffer);

    std::string filepart;
    buffer >> filepart;
    std::string filename="product/"+filepart + ".pdb";

    std::cout << "now writing to file " << filename << std::endl;
    obconversion.SetOutFormat("pdb");
    obconversion.WriteFile(&themolecule,filename);
}

void result::add(OpenBabel::OBMol &other, int hydrogenatomtodelete, int level)
{
    //rotation/translation tricky
    if(themolecule.NumAtoms()>0)
    {
        //we need to know the CH (or NH vector that is proceeding)
        OpenBabel::OBAtom * h = themolecule.GetAtom(hydrogenatomtodelete+1);
        //h is easy, because we specify hydrogenatomtodelete
        OpenBabel::OBAtom * c;
        //go through the rest of the atoms in mol, and if not hydrogen and close to
the
        //hydrogen we know the other atom on the bond
        for(OpenBabel::OBMolAtomIter a(themolecule);a;a++)
        {
            if(!((&*a)->IsHydrogen())&&(AtomDistance(&*a,h)<1.2)){c=(&*a);}
        }
        //now we know C and H that is going to be added to

        //for the substituent you made it easy for me. The H is atom 1
        //now look for the next atom
        OpenBabel::OBAtom * hs = other.GetAtom(1);
        OpenBabel::OBAtom * cs;
        for(OpenBabel::OBMolAtomIter a(other);a;a++)
        {
            if(!((&*a)->IsHydrogen())&&(AtomDistance(&*a,hs)<1.2)){cs=(&*a);}
        }
        //At this point we have our substituent XH vector
        //create vector from, and vector to
        int contacts=badcontacts(&themolecule,&other);
        std::cout << " we have bad contacts " << contacts << std::endl;
        Vector3 from = Vector3(hs->GetX(),hs->GetY(),hs->GetZ());
        Vector3 to = Vector3(0.000,0.000,0.000);
        //move the substituent to the origin before rotation
        Translate(&other,from,to);

        //Rotate the substituent appropriately using the CH and XH vectors
        Vector3 main=Vector3(c->GetX()-h->GetX(),c->GetY()-h->GetY(),c->GetZ()-h->GetZ());
        main.normalize();
        Vector3 subst=Vector3(hs->GetX()-cs->GetX(),hs->GetY()-cs->GetY(),hs->GetZ()-cs->GetZ());
        subst.normalize();
        Rotate(&other,main,-subst);

        //translate to the appropriate place
        to = Vector3(c->GetX()+0.297*(h->GetX()-c->GetX()),c->GetY()+0.297*(h->GetY()-c->GetY()),c->GetZ()+0.297*(h->GetZ()-c->GetZ()));
        from = Vector3(-hs->GetX(),-hs->GetY(),-hs->GetZ());
        Translate(&other,from,to);
    }
}

```

```

        int minimumangle=0;
        int minimumbad=10000;
        //      for(int i=20;i<=360;i+=20)
        //      {
        //          Vector3 axis=Vector3(hs->GetX()-cs->GetX(),hs->GetY()-cs->GetY(),hs-
>GetZ()-cs->GetZ());
        //          BondRotate(&other,axis,0.1);
        //
        if(badcontacts(&themolecule,&other)<minimumbad){minimumangle=i;minimumbad=badcontac
ts(&themolecule,&other);}
        //      if(badcontacts(&themolecule,&other)<5){i=400;}
        //      }
        //      std::cout << "the best bad contacts we get is " << minimumbad << "for angle
" << minimumangle << std::endl;

        //          Vector3 axis=Vector3(hs->GetX()-cs->GetX(),hs->GetY()-cs->GetY(),hs-
>GetZ()-cs->GetZ());
        //          BondRotate(&other,axis,(float)minimumangle);

    }

    int currentatom=0;
    for(OpenBabel::OBMolAtomIter a(other);a;a++)
    {
        //std::cout << "Adding atoms to the molecule, atom number " << currentatom <<
std::endl;
        if((currentatom>=1)|| (level==1))
        {
            OpenBabel::OBAtom newatom;
            newatom.Duplicate(&*a);
            themolecule.AddAtom(newatom);
        }
        currentatom++;
    }
    if(level>1)accounting[hydrogenatomdelete]=-1;
    currentatom=0;
    for(OpenBabel::OBMolAtomIter a(themolecule);a;a++)
    {
        if(((&*a)-
>IsHydrogen())&&(accounting[currentatom]==0))accounting[currentatom]=level;
        currentatom++;
    }
};

void result::print()
{
    std::cout << "molecular mass " << themolecule.GetExactMass() << std::endl;
    for(int i=0;i<50;i++)std::cout << accounting[i];
    std::cout << std::endl;
}

void result::resetall()
{
    for(int i=0;i<1000;i++)setacc(0,i);
}

int result::count(int level)
{
    int thevalue=0;
    for(int i=0;i<1000;i++){if(accounting[i]==level)thevalue++;};
    return thevalue;
}

result & result::operator= (result &other)
{

```

```

        themolecule=other.themolecule;
        for(int i=0;i<1000;i++)
        {
            accounting[i]=other.getacc(i);
        }
        return *this;
    };

void Translate(OpenBabel::OBMol * thesubst, Vector3 from, Vector3 to)
{
    //subst_h will be translated to a point on the CH bond to create a longer bond
    for the new CC^M
    Vector3 displacement = to-from;

    for(OpenBabel::OBMolAtomIter a(thesubst);a;a++)
    {
        float x=a->GetX()+displacement.x;
        float y=a->GetY()+displacement.y;
        float z=a->GetZ()+displacement.z;
        a->SetVector(x,y,z);
    }
}

void Rotate(OpenBabel::OBMol * thesubst, Vector3 mainvector_norm, Vector3
substvector_norm)
{
    //vectors in opposite directions for main the CH vector, for subst the
    HC vector
    Vector3 main=mainvector_norm;
    Vector3 subst=substvector_norm;
    main.normalize();
    subst.normalize();
    Matrix3 mymatrix;

    Vector3 axis = Vector3::cross(main,subst);
    axis.normalize();
    float angle =
    Math::radiansToDegrees(std::acos(Vector3::dot(main,subst)));

    mymatrix.rotate(axis,180.0-angle);

    for(OpenBabel::OBMolAtomIter a(thesubst);a;a++)
    {
        Vector3 atomvector=Vector3(a->GetX(),a->GetY(),a->GetZ());
        Vector3 finalvect=atomvector*mymatrix;
        a->SetVector(finalvect.x,finalvect.y,finalvect.z);
    }
}

int badcontacts(OpenBabel::OBMol * mol1,OpenBabel::OBMol * mol2)
{
    int count=0;
    for(OpenBabel::OBMolAtomIter a(mol1);a;a++)
    {
        for(OpenBabel::OBMolAtomIter b(mol2);b;b++)
        {
            if(AtomDistance((&a),(&b))<1.6) count++;
        }
    }
    return count;
}

```

```

void BondRotate(OpenBabel::OBMol * thesubst, Vector3 axis, float angle)
{
    Matrix3 mymatrix;
    OpenBabel::OBAtom * hs = thesubst->GetAtom(1);
    Vector3 from = Vector3(hs->GetX(), hs->GetY(), hs->GetZ());
    Vector3 to = Vector3(0.000, 0.000, 0.000);
    //move the substituent to the origin before rotation
    Translate(thesubst, from, to);

    axis.normalize();
    mymatrix.rotate(axis, 180.0-angle);

    for(OpenBabel::OBMolAtomIter a(thesubst); a; a++)
    {
        Vector3 atomvector=Vector3(a->GetX(), a->GetY(), a->GetZ());
        Vector3 finalvect=atomvector*mymatrix;
        a->SetVector(finalvect.x, finalvect.y, finalvect.z);
    }
    Translate(thesubst, to, from);
}

double AtomDistance(OpenBabel::OBAtom * atoma, OpenBabel::OBAtom * atomb)
{
    double deltax=atoma->GetX()-atomb->GetX();
    double deltax=atoma->GetY()-atomb->GetY();
    double deltaz=atoma->GetZ()-atomb->GetZ();
    deltax=deltax*deltax;
    deltax=deltax*deltax;
    deltax=deltaz*deltaz;
    double total=deltax+deltay+deltaz;
    total=std::sqrt(total);
    return total;
}

```

2. Optimized docking (L. Sigauke)

a. Vina docking of 61 protease models with 200 ligands.

```

import os, sys, time

#names=open("names_ligand.txt", "w")

protein_array=os.listdir('proteins')
ligands_array=os.listdir('ligands')

"""
for protein in protein_array:
    os.system('cp proteins/'+protein+'
'+ 'newnameproteins/'+protein.split('.')[0]+' .pdbqt')
"""

for protein in protein_array:
    if protein.endswith('.pdbqt'):
        os.system('mkdir '+protein[:-6])
        os.chdir(protein[:-6])
        os.system('mkdir dockings')
        os.system('mkdir logfiles')
        for ligand in ligands_array:

```

```

if ligand.endswith('.pdbqt'):

    jobfile=open(protein[:-6]+'_'+ligand[:-8]+".job","w")
    jobfile.write("#!/bin/csh\n#\n#\n")
    jobfile.write("#PBS -N "+protein[:-6]+'_'+ligand[:-8]+" \n")
    jobfile.write("#PBS -l nodes=1:ppn=1,walltime=1:00:00\n")
    jobfile.write("#PBS -q batch\n")
    jobfile.write("#PBS -d

/home/lester/kev/HIV1/fresh_dock/complete_dock/optimized_docking/"+protein[
:-6]+"/ \n")
    jobfile.write('vina --receptor ../proteins/'+protein+' --
ligand ../ligands/'+ligand+' --out dockings/'+protein[:-6]+'_'+ligand[:-
8]+' .pdbqt --center_x 22.560 --center_y 1.955 --center_z 10.232 --size_x 15
--size_y 15 --size_z 15 --log logfiles/'+protein[:-6]+'_'+ligand[:-8]+' .log
--cpu 4 --exhaustiveness 4\n\n')
    jobfile.close()
    os.system("qsub "+protein[:-6]+'_'+ligand[:-8]+".job")
    time.sleep(2)
os.chdir('..')

#names.close()

```

b. Extraction of relevant data from Vina logfile

```

import sys, os, re, heapq

### what are the names of the folders that I created
names = os.listdir('logfiles')

#####
#Prepare dictionary of ligand binding energies
d = {}

for result in names:

    if result.endswith('log'):
        value = 0
        log = open("logfiles/"+result,"r")
        state = True
        for line in log:

            if line[:3] == '---':
                state = False

            if state == False:

                if line[1] == " ":
                    newline = re.split('\s*',line)
                    if value == 0: #nothing added yet
                        value = float(newline[2]) #first value from the logfile
                    elif float(newline[2]) >= value: #
                        value = value
                    else: #
                        value = float(newline[2])
                    d[result] = value
                if line[0] == "W":
                    state = True
        log.close()

#####

```



```
#Sort through the dictionary 'd' to generate the top 100 ligands.
```

```
from collections import Counter
```

```
topset = open("topall_optimized.txt","w")
for item in d.items():
    string = ' '.join(str(i) for i in item)
    topset.write(string)
    topset.write('\n')
topset.close()
```

```
docs = dict(Counter(d).most_common()[::-200-1::-1])
top200 = open('top200_optimized.txt','w')
for item in docs.items():
    string = ' '.join(str(i) for i in item)
    top200.write(string)
    top200.write('\n')
top200.close()
```

c. Analysis of all log files to calculate docking statistics

```
import os, sys, re
```

```
interaction = open('topall_optimized.txt','r')
energylist = []
keys = []
lists = {}
element = []
d = {}
```

```
for line in interaction:
    separated = re.split('\s*',line) # e.g.
    ['5046_fligand_6906_docked.log', '-10.3', '']
    name = separated[0][:-4] # e.g. 301812_fligand_8322_docked
    ligand = name.split('_')[1] # e.g. 301812
    energyvalues = float(separated[1]) # e.g. -9.00
```

```
if ligand in keys:
    keys = keys
    d[ligand] = d[ligand] + [energyvalues]
```

```
else:
    keys.append(ligand)
    #element= [energyvalues]
    d[ligand] = [energyvalues]
```

```
#print d
```

```
#####
##### Preliminary Statistical analysis.
#energylist.append(float(separated[1])) # this is how to create a list
of energy values.
#energylist.sort()
#print energylist
```

```
import numpy
```

```

statsdoc = open("statsdoc.txt","w")
statsdoc.write("#LigandName, #Med, #Mean, #StandardDeviation \n")
for i in d:
    med_value = numpy.median(d[i])
    mean_value = numpy.mean(d[i])
    std_value = numpy.std(d[i])

    statsdoc.write("%s, %.3f, %.3f, %.3f \n" % (i, med_value, mean_value,
std_value))
statsdoc.close()

```

3. Protein Ligand Molecular Dynamics (L. Sigauke)

```

import os, sys

os.system('mkdir jobfiles')
os.system('mkdir pythonscripts')
os.system('mkdir outputs')
os.system('mkdir errors')
#### define a function:
def dynamic_process(protein):

    if protein[-4:] == '.pdb':
        ### creates a job
        name = protein[:-4]
        os.system('mkdir '+name)
        os.chdir(name)
        jobfile = open('../jobfiles/'+name+'_ligprot.job','w')
        jobfile.write('#!/bin/bash \n')
        jobfile.write('#PBS -V \n')
        jobfile.write('#PBS -N '+name+'\n')
        jobfile.write('#PBS -q tf \n')
        jobfile.write('#PBS -d .\n')
        jobfile.write('#PBS -e
/home/lester/kev/HIV1/dynamicstudies/lig_prot/errors/'+name+'_error.txt
\n')
        jobfile.write('#PBS -o
/home/lester/kev/HIV1/dynamicstudies/lig_prot/outputs/'+name+'_output.txt
\n')
        jobfile.write('#PBS -l
nodes=1:ppn=24,walltime=99:00:00,mem=16gb\n')
        jobfile.write('python pythonscripts/'+name+'_ligprotdyn.py \n')
        jobfile.close()
        ### creates a script
        python = open('../pythonscripts/'+name+'_ligprotdyn.py','w')
        python.write('import os, sys, time \n')
        python.write('\n')
        python.write('\n')
        python.write('name = '+'"+name+"'\n')
        python.write('os.system("mkdir "+name+"/dynamics_ligprot")\n')

        python.write('os.system("cp -r mdp "+name+"/dynamics_ligprot")\n')

        python.write('os.chdir(name+"/dynamics_ligprot")\n\n')
        python.write('moldyn = "../..\n')
        python.write('echothing = "6610h 1" \n')
        python.write('os.system("cp ../6610/6610h.acpype/6610h_GMX.itp
6610h.itp")\n') #create topology for ligand

```

```

python.write('os.system("cp ../../6610/6610h.acpype/6610h_NEW.pdb
.")\n') #bring _NEW.pdb into visibility
python.write('os.system("mkdir equil heat em prod ions")\n')
python.write('os.system("g_pdb2gmx -f
"+moldyn+"/protease_models/"+name+".pdb -o "+name+"_proc.pdb -water spce -
ff amber03")\n')
python.write('os.system("grep -h ATOM "+name+"_proc.pdb
6610h_NEW.pdb > complex.pdb") \n') #merge the
python.write('os.system("sed -i \'/#include
\\"amber03.ff\\forcefield.itp\\\'/a \\'#include \\'6610h.itp\\\'\'
topol.top")\n')
python.write('os.system("echo "+echothing+" >> topol.top")\n')
#echo
python.write('os.system("g_editconf -f complex.pdb -o
"+name+"_box.gro -c -d 1.0 -bt cubic")\n') #simulation box - g_editconf
python.write('os.system("g_genbox -cp "+name+"_box.gro -cs
spc216.gro -o "+name+"_solv.gro -p topol.top")\n') #solvate the box -
g_genbox
python.write('os.system("g_grompp -f mdp/ions.mdp -c
"+name+"_solv.gro -p topol.top -o ions/ions.tpr")\n') # Neutralization
g_grompp
python.write('os.system("echo 15 | g_genion -s ions/ions.tpr -o
ions/"+name+"_ions.gro -p topol.top -pname NA -nname CL -nn 1")\n') #echo
15
python.write('os.system("g_grompp -f mdp/em.mdp -c
ions/"+name+"_ions.gro -p topol.top -o em/em.tpr")\n') #Minimize the system
g_grompp
python.write('os.system("g_mdrun -v -deffnm em/em")\n') # rn
minimization g_
python.write('os.system("g_grompp -f mdp/heat.mdp -c em/em.gro -p
topol.top -o heat/heat.tpr")\n') #Heat the system g_
python.write('os.system("g_mdrun -deffnm heat/heat")\n') # rn
heating g_
python.write('os.system("g_grompp -f mdp/equil.mdp -c heat/heat.gro
-t heat/heat.cpt -p topol.top -o equil/equil.tpr")\n') #g_short
equilibration
python.write('os.system("g_mdrun -deffnm equil/equil")\n') #g_ru
python.write('os.system("g_grompp -f mdp/500ps_prod.mdp -c
equil/equil.gro -t equil/equil.cpt -p topol.top -o
prod/500ps_prod.tpr")\n') #g_Longer dynamics
python.write('os.system("g_mdrun -deffnm prod/500ps_prod")\n') #g
run
python.write('os.system("echo 1 | g_trjconv -f prod/500ps_prod.trr
-s prod/500ps_prod.tpr -o prod/500ps_prod.pdb")\n') #g_trjconv

python.close()
#check path situation
os.chdir('../')
return

proteinlist = os.listdir('protease_models')

for model in proteinlist:
    if model != '':
        ##### loop through a list
    of proteins
        dynamic_process(model)
    ### call the function for each one.
    os.chdir('pythonscripts')
    os.system("chmod 777 "+model[:-4]+"_ligprotodyn.py")

```

```

    os.chdir('../')
    os.system("qsub jobfiles/"+model[:4]+"_ligprot.job")
# submit a job to the cluster

```

```

###

```

4. Analysis of Molecular Dynamic simulations (L. Sigauke)

a. Quality and Structural features

```

import os, sys

```

```

os.system('mkdir qualityash')
os.system('echo 14 | g_energy -f prod.edr -o qualityash/temperature.xvg')
os.system('echo 22 | g_energy -f prod.edr -o qualityash/density.xvg')
os.system('echo 1 | g_mindist -f prod.xtc -s prod.tpr -od
qualityash/minimal-periodic-distance.xvg -pi')
os.system('echo 1 | g_rmsf -f prod.xtc -s prod.tpr -o qualityash/rmsf-all-
atom.xvg -ox qualityash/average.pdb -oq qualityash/bfactors.pdb')
os.system('echo 13 | g_rmsf -f prod.xtc -s prod.tpr -o qualityash/rmsf-
lig.xvg -ox qualityash/average_lig.pdb -oq qualityash/bfactors_lig.pdb')
os.system('echo 1 1 | g_rms -f prod.xtc -s prod.tpr -o qualityash/rmsd-all-
atom-vs-start.xvg')
os.system('echo 3 3 | g_rms -f prod.xtc -s prod.tpr -o qualityash/rmsd-
backbone-vs-start.xvg')
os.system('echo 1 1 | g_rms -f prod.xtc -s qualityash/average.pdb -o
qualityash/rmsd-all-atom-vs-average.xvg')
os.system('echo 3 3 | g_rms -f prod.xtc -s qualityash/average.pdb -o
qualityash/rmsd-all-atom-vs-average.xvg')
os.system('echo 13 13 | g_rms -f prod.xtc -s prod.tpr -o qualityash/rmsd-
ligand-vs-start.xvg')
os.system('echo 13 13 | g_rms -f prod.xtc -s qualityash/average_lig.pdb -o
qualityash/rmsd-lig-vs-average_lig.xvg')
os.system('echo 1 | g_gyrate -f prod.xtc -s prod.tpr -o qualityash/radius-
of-gyration.xvg')
os.system('echo 13 | g_gyrate -f prod.xtc -s prod.tpr -o qualityash/radius-
of-gyration_ligand.xvg')
os.system('mkdir structural')
os.system('echo 1 1 | g_sas -f prod.xtc -s prod.tpr -o structural/solvent-
accessible-surface.xvg -oa structural/atomic-sas.xvg -or
structural/residue-sas.xvg')
os.system('echo 1 1 | g_hbond -f prod.xtc -s prod.tpr -num
structural/hydrogen-bonds-intra-protein.xvg')
os.system('echo 1 13 | g_hbond -f prod.xtc -s prod.tpr -num
structural/hydrogen-bonds-protein-lig.xvg')
os.system('g_saltbr -f prod.xtc -s prod.tpr -t 1.0 -sep')
os.system('g_rama -f prod.xtc -s prod.tpr -o structural/ramachandaran.xvg')
os.system('')
os.system('mkdir dynamic_prop')
os.system('echo 4 4 | g_covar -s prod.tpr -f prod.xtc -o
dynamic_prop/eigenvalues.xvg -v dynamic_prop/eigenvectors.trr -ascii
dynamic_prop/covariances.dat')
#os.system('mkdir binding')

```

b. Binding Energy via mmpbsa

```
import os, sys

os.system('mkdir pythonscripts')
os.system('mkdir outputs')
os.system('mkdir errors')

#### define a function:
def dynamic_process(protein):

    if protein[-4:] == '.pdb':
        name = protein[:-4]
        os.system('mkdir '+name)
        os.chdir(name)
        ### creates a script
        python = open('../pythonscripts/'+name+'_ligprotdyn.py', 'w')
        python.write('import os, sys, time \n')
        python.write('\n')
        python.write('\n')
        python.write('name = '+'\"'+name+\"'\n')
        #python.write('os.system("mkdir "+name+"/dynamics_ligprot")\n')

        pathtoprod =
'/home/lester/kev/HIV1/dynamicstudies/lig_prot/50ps/805_lig/'+name+'/dynam
cs_ligprot/prod'
        python.write('os.chdir("+name+")\n')
        python.write('os.system("cp -r '+pathtoprod+'/*.xtc
'+pathtoprod+'/*.tpr '+pathtoprod+'/*.gro ../'+name+'")\n')
        #python.write('os.system("ssh tf")\n') #change to the tf shell
        pathto =
"/home/lester/kev/HIV1/dynamicstudies/lig_prot/50ps/805_analysis/"+name #
enter the proteins folder
        python.write('os.system("echo q | g_make_ndx -f prod.gro -o
prod.ndx")\n') # make .ndx file
        python.write('os.system("echo 1 13 | g_mmpbsa -f prod.xtc -s
prod.tpr -n prod.ndx -pdie 2 -decomp")\n') # calculation of potential
energy in vacuum
        python.write('os.system("echo 1 13 | g_mmpbsa -f prod.xtc -s
prod.tpr -n prod.ndx -i ../polar.mdp -nomme -pbsa -decomp")\n') #
calculation of polar solvation energy
        python.write('os.system("echo 1 13 | g_mmpbsa -f prod.xtc -s
prod.tpr -n prod.ndx -i ../apolar_sasa.mdp -nomme -pbsa -decomp -apol
sasa.xvg -apcon sasa_contrib.dat")\n') # calculation of SASA-only non-polar
solvation energy
        python.write('os.system("cp ../MnPbSaStat.py .")\n')
        python.write('os.system("python26 MnPbSaStat.py -m energy_MM.xvg -p
polar.xvg -a sasa.xvg")\n')

    python.write('os.chdir("/home/lester/kev/HIV1/dynamicstudies/lig_prot/50ps/
805_analysis/")\n')
    #python.write('os.system("exit")\n')

    python.close()
    #check path situation

os.chdir('/home/lester/kev/HIV1/dynamicstudies/lig_prot/50ps/805_analysis/'
) #get out of protein folder
return
```

```

proteinlist = os.listdir('../805_lig/protease_models')
modellist = os.listdir('.')

#print modellist

for model in proteinlist: #### loop through a list of proteins
    if model != '':
        if model[:-4] in modellist:
            pass
        else:
            dynamic_process(model)
### call the function for each one.
os.system("chmod 777 "+model[:-4]+"_ligprotdyn.py")
os.system("python pythonscripts/"+model[:-4]+"_ligprotdyn.py")

```

APPENDIX B – RAW DATA

1. Docking Data

a. Sample from top 200 protein – Ligand Interactions

```

5014_fligand_536_docked.log -11.4
5046_fligand_16219_docked.log -11.3
5261_fligand_5043_docked.log -11.2
514412_fligand_9533_docked.log -11.1
514412_fligand_13509_docked.log -11.1
5045_fligand_13509_docked.log -11.3
501424_fligand_513_docked.log -11.5
302112_fligand_392_docked.log -11.6
3021_fligand_7839_docked.log -11.1
305152_fligand_3000_docked.log -11.0
504612_fligand_805_docked.log -11.2
5046_fligand_16404_docked.log -11.1
52423_fligand_5043_docked.log -11.1
5166_fligand_3206_docked.log -11.5
5045_fligand_15733_docked.log -11.2
3051_fligand_3206_docked.log -11.0
305152_fligand_8973_docked.log -11.2
508012_fligand_5411_docked.log -11.1
501424_fligand_8560_docked.log -11.3
5242_fligand_8560_docked.log -11.1
5207_fligand_805_docked.log -11.3
503252_fligand_6071_docked.log -11.3
302112_fligand_1758_docked.log -11.6
1HXB_RenumberedResidue_fit_fligand_7440_docked.log -11.0
305152_fligand_6130_docked.log -11.3
509424_fligand_15965_docked.log -11.2
5014_fligand_12957_docked.log -11.1
5166_fligand_1758_docked.log -11.6
509424_fligand_16879_docked.log -11.1
5261_fligand_14631_docked.log -11.5

```

b. Statistical analysis

#LigandName, #Med, #Mean, #StandardDeviation

15454, -9.100, -9.068, 0.880
8626, -9.700, -9.544, 0.806
10069, -9.700, -9.605, 0.898
137, -9.700, -9.579, 0.781
6724, -9.350, -9.318, 0.839
8014, -9.600, -9.448, 0.805
9773, -9.600, -9.390, 0.929
17258, -9.600, -9.363, 0.869
691, -9.700, -9.644, 0.788
12679, -9.300, -9.235, 1.071
11388, -9.300, -9.327, 0.671
13565, -9.600, -9.521, 0.724
16315, -9.800, -9.845, 0.721
542, -9.600, -9.573, 0.795
12174, -9.400, -9.352, 0.907
11466, -9.500, -9.489, 0.761
17277, -9.700, -9.703, 0.712
5065, -9.300, -9.308, 0.968
3837, -9.300, -9.134, 1.119
16464, -9.600, -9.594, 0.612
8921, -9.750, -9.603, 0.845
9066, -9.400, -9.329, 0.775
16469, -9.500, -9.469, 0.863
671, -9.650, -9.597, 0.790
1152, -9.400, -9.221, 1.033
10206, -9.850, -9.655, 0.972
122, -9.600, -9.497, 0.812
10205, -9.700, -9.585, 0.883
5990, -9.450, -9.365, 0.788
7043, -9.550, -9.561, 0.761
9989, -9.400, -9.232, 0.806
11573, -9.300, -9.263, 0.879
6096, -9.700, -9.668, 0.833
7440, -10.000, -10.008, 0.682
2552, -9.500, -9.442, 1.125
3561, -9.400, -9.361, 0.753
7445, -9.450, -9.292, 0.799
6657, -9.600, -9.348, 0.824
6564, -9.800, -9.626, 0.866
16879, -9.800, -9.665, 0.856
15696, -9.200, -9.173, 0.885
4103, -9.500, -9.305, 1.115
12383, -9.450, -9.342, 0.887
536, -9.300, -9.300, 0.908
10056, -9.700, -9.589, 0.817
299, -9.400, -9.345, 0.817
11479, -9.750, -9.600, 0.790

2. Molecular Dynamics

a. Binding Energy analysis of 6610 with protease model 305152

#Complex Number: 1

=====

SUMMARY

=====

van der Waal energy	=	-296.401	+/-	10.049 kJ/mol
Electrostatic energy	=	-81.661	+/-	13.459 kJ/mol
Polar solvation energy	=	284.878	+/-	11.103 kJ/mol
SASA energy	=	-28.951	+/-	0.892 kJ/mol
SAV energy	=	0.000	+/-	0.000 kJ/mol
WCA energy	=	0.000	+/-	0.000 kJ/mol
Binding energy	=	-122.135	+/-	14.430 kJ/mol

=====

END

=====

c. Binding Energy analysis of 805 with protease model 305152

#Complex Number: 1

=====

SUMMARY

=====

van der Waal energy	=	-318.640	+/-	17.987 kJ/mol
Electrostatic energy	=	-104.061	+/-	8.387 kJ/mol
Polar solvation energy	=	309.489	+/-	13.767 kJ/mol
SASA energy	=	-29.252	+/-	0.952 kJ/mol
SAV energy	=	0.000	+/-	0.000 kJ/mol
WCA energy	=	0.000	+/-	0.000 kJ/mol
Binding energy	=	-142.464	+/-	20.194 kJ/mol

=====

END

=====