RHODES UNIVERSITY
Where leaders learn

Application of machine learning, molecular modelling and structural data mining against antiretroviral drug resistance in HIV-1

A thesis by

Olivier Serge André
SHEIK AMAMUDDY

Department of Biochemistry & Microbiology

Submitted in fulfilment of the
requirements for the degree of

Doctor of Philosophy in Bioinformatics

February 2019

Abstract

Millions are affected with the Human Immunodeficiency Virus (HIV) world wide, even though the death toll is on the decline. Antiretrovirals (ARVs), more specifically protease inhibitors have shown tremendous success since their introduction into therapy since the mid 1990's by slowing down progression to the Acquired Immune Deficiency Syndrome (AIDS). However, Drug Resistance Mutations (DRMs) are constantly selected for due to viral adaptation, making drugs less effective over time. The current challenge is to manage the infection optimally with a limited set of drugs, with differing associated levels of toxicities in the face of a virus that (1) exists as a quasispecies, (2) may transmit acquired DRMs to drug-naive individuals and (3) that can manifest class-wide resistance due to similarities in design. The presence of latent reservoirs, unawareness of infection status, education and various socio-economic factors make the problem even more complex. Adequate timing and choice of drug prescription together with treatment adherence are very important as drug toxicities, drug failure and sub-optimal treatment regimens leave room for further development of drug resistance. While CD4 cell count and the determination of viral load from patients in resource-limited settings are very helpful to track how well a patient's immune system is able to keep the virus in check, they can be lengthy in determining whether an ARV is effective. Phenosense assay kits answer this problem using viruses engineered to contain the patient sequences and evaluating their growth in the presence of different ARVs, but this can be expensive and too involved for routine checks. As a cheaper and faster alternative, genotypic assays provide similar information from HIV *pol* sequences obtained from blood samples, inferring ARV efficacy on the basis of drug resistance mutation patterns. However, these are inherently complex and the various methods of *in silico* prediction, such as Geno2pheno, REGA and Stanford HIVdb do not always agree in every case, even though this gap decreases as the list of resistance mutations is updated. A major gap in HIV treatment is that the information used for predicting drug resistance is mainly computed from data containing an overwhelming majority of B subtype HIV, when these only comprise about 12% of the worldwide HIV infections. In addition to growing evidence that drug resistance is subtype-related, it is intuitive to hypothesize that as subtyping is a phylogenetic classification, the more divergent a subtype is from the strains used in training prediction models, the less their resistance profiles would correlate.

For the aforementioned reasons, we used a multi-faceted approach to attack the virus in multiple ways. This research aimed to (1) improve resistance prediction methods by focusing solely on the available subtype, (2) mine structural information pertaining to resistance in order to find any exploitable weak points and increase knowledge of the mechanistic processes of drug resistance in HIV protease. Finally, (3) we screen for protease inhibitors amongst a database of natural compounds [the South African natural compound database (SANCDB)] to find molecules or molecular properties usable to come up with improved inhibition against the drug target. In this work, structural information was mined using the Anisotropic Network Model, Dynamics Cross-Correlation, Perturbation Response Scanning, residue contact network analysis and the radius of gyration. These methods failed to give any resistance-associated patterns in terms of natural movement, internal correlated motions, residue perturbation response, relational behaviour and global compaction respectively. Applications of drug docking, homology-modelling and energy-minimization for generating features suitable for machine-learning were not very promising, and rather suggest that the value of binding energies by themselves from Vina may not be very reliable quantitatively. All these failures lead to a refinement that resulted in a highly-sensitive statistically-guided network construction and analysis, which leads to key findings in the early dynamics associated with resistance across all PI drugs. The latter experiment unravelled a conserved lateral expansion motion occurring at the flap elbows, and an associated contraction that drives the base of the dimerization domain towards the catalytic site's floor in the case of drug resistance. Interestingly, we found that despite the conserved movement, bond angles were degenerate. Alongside, 16 Artificial Neural Network models were optimised for HIV proteases and reverse transcriptase inhibitors, with performances on par with Stanford HIVdb. Finally, we prioritised 9 compounds with potential protease inhibitory activity using virtual screening and molecular dynamics (MD) to additionally suggest a promising modification to one of the compounds. This yielded another molecule inhibiting equally well both opened and closed receptor target conformations, whereby each of the compounds had been selected against an array of multi-drug-resistant receptor variants. While a main hurdle was a lack of non-B subtype data, our findings, especially from the statistically-guided network analysis, may extrapolate to a certain extent to them as the level of conservation was very high within subtype B, despite all the present variations. This network construction method lays down a sensitive approach for analysing a pair of alternate phenotypes for which complex patterns prevail, given a sufficient number of experimental units. During the course of research a weighted contact mapping tool was developed to compare renin-angiotensinogen variants and packaged as part of the MD-TASK tool suite. Finally the functionality, compatibility and performance of the MODE-TASK tool were evaluated and confirmed for both Python2.7.x and Python3.x, for the analysis of normals modes from single protein structures and essential modes from MD trajectories. These techniques and tools collectively add onto the conventional means of MD analysis.

Declaration

The research described in this thesis was carried out in the context of my PhD research in Bioinformatics, which started on 31st March 2016 to end on the 15th of December 2018 under the supervisions of Prof Özlem Tastan Bishop and Prof Nigel T. Bishop. I, Olivier Sheik Amamuddy, declare that this thesis submitted to Rhodes University is wholly my own work and has not previously been submitted for a degree at this or any other institution.

Signature: O. SheikAmamuddy

Date: 05/02/2019

Acknowledgement

I am thankful for my project supervisors for all the assistance and environment they created for my self-development and improvement in bioinformatics. Being part of the RUBi research group has made me grow a lot as a research scientist. Special thanks to the *Center for High Performance Computing* for supercomputing facilities and the *National Research Foundation* (grant number U24HG006941) for funding the research, without which this project would not have been feasible.

Contents

List of Figures

List of Tables

List of abbreviations

Å	Angstrom
ABC	Abacavir
ACE	Angiotensin Converting Enzyme
AIDS	Acquired immune deficiency syndrome
AMBER	Assisted Model Building with Energy Refinement
ANM	Anisotropic network model
ANN	Artificial neural network
ANRS	Agence Nationale de Recherches sur le SIDA
APOBEC3	Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3
ART	Antiretroviral therapy
ARV	Antiretroviral
ASP	Aspartic acid
ATV	Atazanavir
AZT	Zidovudine
BC	Betweenness centrality
BCC	Bond charge correction
BFGS	Broyden-Fletcher-Goldfarb-Shanno
CA	Capsid
CAV	Coxsackievirus
CCR5	Cysteine-Cysteine Chemokine Receptor 5
CD4	Cluster of differentiation 4
CHPC	Centre for high performance computing
CNN	Convolutional neural network
COBI	Cobicistat
COM	Center of mass
CRF	Circulating Recombinant Form
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSV	Comma-separated values
CV	Cross-validation
CXCR4	(C-X-C Motif Chemokine Receptor 4
DBSCAN	Density-based spatial clustering of applications with noise
DC3	Diffusa Cyclotide-3 (a cyclopeptide)
DCC	Dynamic Cross Correlation

DDI	Didanosine
DNA	Deoxyribonucleic acid
DOPE	Discrete optimized protein energy
DRIN	Dynamic residue interaction network
DRM	Drug resistant mutation
DRN	Dynamic residue network
DRV	Darunavir
DTG	Dolutegravir
EFV	Efavirens
EM	Energy minimization
ENF	Enfuvirtide
ENM	Elastic network model
ETR	Etravirine
EVD	Eigenvalue decomposition
EVG	Elvitegravir
FATHMM	Functional Analysis through Hidden Markov Models
FDA	Food and Drug Administration
FN	False negative
FP	False positive
FPV	Fosamprenavir
FTC	Emtricitabine
GLY	Glycine
GNM	Gaussian Network Model
GUI	Graphical User Interface
HAART	Highly active antiretroviral therapy
HIV	Human immunodeficiency virus
HTVS	High-throughput virtual screening
HUMA	Human Mutation Analysis
IBA	Ibalizumab
IC	Inhibitory concentration
IDV	Indinavir
ILE	Isoleucine
IN	Integrase
INI	Integrase Inhibitor
KDa	Kilodaltons
KNN	k-Nearest Neighbors
LINCS	Linear Constraint Solver
LOOCV	Leave-one-out cross-validation
LPV	Lopinavir
LSTM	Long short-term memory
MA	Matrix (protein)

MAR	Maximum adjacency ratio
MD	Molecular dynamics
MDR	Multidrug-resistant
MDS	Multidimensional scaling
ML	machine learning
MLR	Multiple linear regression
MM-GBSA	Molecular Mechanics/Generalized Born Surface Area
MM-PBSA	Molecular Mechanics/Poisson-Boltzmann Surface Area
MSE	Mean squared error
MVC	Maraviroc
NA	Not applicable
NC	Nucleocapsid
NFV	Nelfinavir
NMA	Normal mode analysis
NMR	Nuclear magnetic resonance
NNRTI	Non-nucleoside reverse-transcriptase inhibitor
NRTI	Nucleoside reverse-transcriptase inhibitor
NVP	Nevirapine
PBC	Periodic boundary condition
PBMCs	Peripheral blood mononuclear cell
PC	Principal component
PCA	Principal components analysis
PHE	Phenylalanine
PI	Protease inhibitor
PIC	Protein Interactions Calculator
PME	Particle mesh Ewald
PR	Protease
PROVEAN	Protein Variation Effect Analyzer
PRS	Perturbation-response scanning
PSEP	Position-specific evolutionary preservation
QMEAN	Qualitative Model Energy ANalysis
QSAR	Quantitative Structure-Activity Relationship
RAAS	Renin-angiotensin-aldosterone system
RAL	Raltegravir
RAS	Renin-angiotensin system
RBF	Radial basis function
RCN	Residue contact network
RF	Random forest
RIN	Residue interaction network
RMSD	Root mean squared deviation
RMSF	Root mean squared fluctuation

RNA	Ribonucleic acid
RNAse	Ribonuclease
RNN	Recurrent neural network
RPV	Rilpivirine
RT	Reverse transcriptase
RTI	Reverse transcriptase inhibitor
RTV	Ritonavir
RUBi	Research Unit in Bioinformatics
SD	Steepest descent
SNP	Single-nucleotide polymorphism
SNV	Single-nucleotide variant
SPC	Simple point-charge
SQV	Saquinavir
SSH	Secure Shell
STAT3	Signal transducer and activator of transcription 3
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVD	singular value decomposition
SVM	Support Vector Machine
TDF	Tenofovir
THR	Threonine
TIM	Triosephosphate isomerase
TIP	Transferable intermolecular potential
TN	True negative
TO	Topological overlap
TP	True positive
TPV	Tipranavir
t-SNE	t-Distributed Stochastic Neighbour Embedding
UNAIDS	Joint United Nations Programme on HIV and AIDS
URF	Unique Recombinant Form
VAL	Valine
VAPOR	Variant Analysis Portal
VP1	Viral protein 1
VP2	Viral protein 2
VP3	Viral protein 3
VP4	Viral protein 4
WHO	World Health Organization
WT	Wild Type

List of tools, libraries and web servers

ACPYPE	https://github.com/llazzaro/acpype
AutoDockTools	http://autodock.scripps.edu/resources/adt
AutoDock Vina	http://vina.scripps.edu/
Avogadro	https://avogadro.cc/
EMBOSS	http://emboss.sourceforge.net/
igraph (R)	https://igraph.org/r/
Jalview	http://www.jalview.org/
GNU Parallel	https://www.gnu.org/software/parallel/
GROMACS	http://www.gromacs.org/
JupyterLab	https://github.com/jupyterlab/jupyterlab
MATLAB	https://www.mathworks.com/products/matlab.html
Matplotlib	https://matplotlib.org/
MD-TASK	https://github.com/RUBi-ZA/MD-TASK
MDTraj	http://mdtraj.org
MODE-TASK	https://github.com/RUBi-ZA/MODE-TASK
MODELLER	https://salilab.org/modeller/
MUSCLE	https://www.ebi.ac.uk/Tools/msa/muscle/
NetworkX	https://networkx.github.io/
nglview	https://github.com/arose/nglview
NumPy	http://www.numpy.org/
Open Babel	http://openbabel.org
pandas	https://pandas.pydata.org/
PDB	http://www.rcsb.org/
PDB2PQR	https://github.com/Electrostatics/apbs-pdb2pqr
PyMOL	https://pymol.org
ProDy	http://prody.csb.pitt.edu/
Python	https://www.python.org/
R	https://www.r-project.org/
SciPy	https://www.scipy.org/
SWISS-MODEL	https://swissmodel.expasy.org/
SANCDB	https://sancdb.rubi.ru.ac.za/
Stanford HIVdb	https://hivdb.stanford.edu/
SHIVA	http://shiva.heiderlab.de/

VEGA	https://nova.disfarm.unimi.it
VMD	https://www.ks.uiuc.edu/Research/vmd/
X-Score	http://www.umich.edu/~shaomengwanglab/software/xtool/

Research outputs

Publications and contributions:

1. **Sheik Amamuddy O**, Bishop Nigel T and Tastan Bishop Ö. "Characterizing early drug resistance-related events using geometric ensembles from HIV protease dynamics." *Scientific Reports*, 2018 December 18. doi: 10.1038/s41598-018-36041-8.

   Contributions: I designed and conducted this work under the guidance of my supervisors. They assisted me in improving the first article draft before submission and also after the peer-review process.

2. Ross CJ, Nizami B, Glenister M, **Sheik Amamuddy O**, Atilgan AR, Atilgan C and Tastan Bishop Ö "MODE-TASK: Large-scale protein motion tools" *Bioinformatics*, 2018 May 29. doi: 10.1093/bioinformatics/bty427.

   Contributions: I tested and evaluated the performance of the command line version of the MODE-TASK suite, recommended changes and fixed some minor typographical mistakes in some scripts before evaluating the graphical user interface functionality.

3. **Sheik Amamuddy O**, Bishop NT and Tastan Bishop Ö. "Improving fold resistance prediction of HIV-1 against protease and reverse transcriptase inhibitors using artificial neural networks" *BMC Bioinformatics*, 2017 August 15. doi: 10.1186/s12859-017-1782-x.

   Contributions: I trained separate neural network models for predicting drug resistance in HIV protease subtype B, and wrote the first draft of the article before correction by my supervisors and also corrected the peer-reviewed manuscript with their recommendations.

4. Brown DK, Penkler DL, **Sheik Amamuddy O**, Ross C, Atilgan AR, Atilgan C and Tastan Bishop Ö. "MD-TASK: a software suite for analyzing molecular dynamics trajectories." *Bioinformatics*, 2017 May 31. doi: 10.1093/bioinformatics/btx349. PMID: 28575169.

   Contributions: I refined and generalised previous code written for building weighted contact maps from MD trajectories to be incorporated in the MD-TASK tool suite and generated MD trajectories for a mutant and wild-type HIV protease for the contact maps displayed in the article.

5. Brown DK, **Sheik Amamuddy O** and Tastan Bishop Ö. "Structure-Based Analysis of Single Nucleotide Variants in the Renin-Angiotensinogen Complex." *Global Heart*, 2017 Mar 13. pii: S2211-8160(17)30006-6. doi: 10.1016/j.gheart.2017.01.006. PMID: 28302554.

Contributions: I performed molecular dynamics simulations and computed weighted contact networks, RMSD and RMSF metrics on homology models provided by Dr David Brown.

Poster presentations:

1. **Sheik Amamuddy O**, Bishop Nigel T and Tastan Bishop Ö. “Artificial Neural Networks: Applications in HIV drug resistance research”. 10th H3Africa consortium meeting.
2. **Sheik Amamuddy O**, Lobb K and Tastan Bishop Ö. “High-throughput in silico screening of natural compounds to accelerate the discovery of novel protease inhibitors of HIV-1 (subtype C)”. The Tenth Annual CHPC National Meeting and Conference, East London International Convention Centre. East London, South Africa, 2016.

Oral presentation:

1. **Sheik Amamuddy O**, Bishop Nigel T and Tastan Bishop Ö. “Screening SANCDB for compounds with potential inhibitory activity against HIV-1 subtype C proteases from South African patients undergoing Lopinavir treatment”. 25th South African Society of Biochemistry and Molecular Biology (SASBMB) Congress, 2016. East London, South Africa, 2016.

Thesis overview

The thesis is divided into two main parts. In Part I, the research problem is motivated, before proceeding to explain the state-of-the-art in HIV treatment, which therefore requires an explanation the viral life-cycle as it is a main working form of controlling the virus, in the wait of a cure. The main research thematic is that of finding novel *in silico* ways of tackling and improving our understanding of the drug-resistance problem. In Part II, three side projects performed in collaboration with various members of the RUBi research laboratory are elaborated, for which the extent of my contributions are listed in the **Publications and contributions** section.

In Chapter 1, an introduction is provided on the viral life-cycle, its general structural, genomic and taxonomic organisation, in addition to the current treatment strategies before focusing on the viral protease.

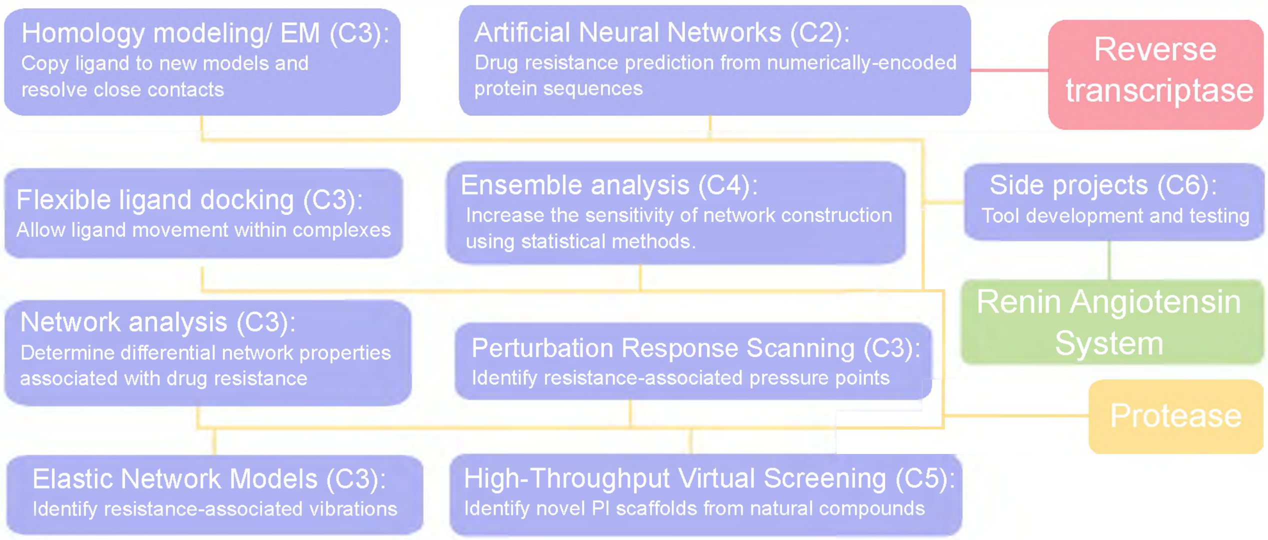
In chapter 2, based on available HIV drug resistance data sets from the Stanford HIVdb and the proportion of subtypes present therein, we aimed to improve the performance of drug resistance prediction for HIV proteases and reverse transcriptases in HIV. Ideally, a more general prediction method was desired, unfortunately we ended up building models for subtype B only, due to lack of sufficient labelled non-subtype B data. To do so, we applied various filtering approaches for the construction of neural network models for the improvement of drug resistance prediction in HIV.

In Chapter 3, several structural characteristics of the viral protease were mined to search for a possible resistance-specific signal in an attempt to find a more generalisable property that would hopefully extrapolate to non-subtype B HIV strains.

A failure to see any resistance-related signal in the previous chapter lead to the development of a more sensitive method, designed by improving the residue contact network approach by coupling statistical tests to network construction across an ensemble of MD trajectories for extracting short, but well-conserved motions generally associated with drug resistance, as presented in Chapter 4.

In Chapter 5, another facet of drug resistance in HIV protease is investigated via high throughput virtual screening for the discovery of novel drug scaffolds using the South African natural compound database. As opposed to single variant targeting, a diverse set of protease variants obtained from the Stanford HIVdb was used to guide the screening process to target sequences from darunavir-failing patients.

Finally, in Part 2, we present side projects showcasing the tools MD-TASK and MODE-TASK, which are sets of scripts developed/tested in collaboration with RUBi members, and also showcase the application of network analysis onto MD data, using the renin-angiotensinogen complex - an important hypertension-related drug target.



Homology modeling/ EM (C3):
Copy ligand to new models and resolve close contacts
Artificial Neural Networks (C2):
Drug resistance prediction from numerically-encoded protein sequences
Reverse transcriptase
Flexible ligand docking (C3):
Allow ligand movement within complexes
Ensemble analysis (C4):
Increase the sensitivity of network construction using statistical methods.
Side projects (C6):
Tool development and testing
Renin Angiotensin System
Network analysis (C3):
Determine differential network properties associated with drug resistance
Perturbation Response Scanning (C3):
Identify resistance-associated pressure points
Protease
Elastic Network Models (C3):
Identify resistance-associated vibrations
High-Throughput Virtual Screening (C5):
Identify novel PI scaffolds from natural compounds

Figure 1: Schematic of the thesis. Coloured lines depict the techniques/sections applied to the respective proteins, namely HIV protease (yellow) and reverse transcriptase (red), and the human renin-angiotensinogen complex (green). Chapters are abbreviated C1-C6, for Chapters 1 to 6.

Part I

The main research: Targeting drug resistance in HIV

Chapter 1

Introduction to the viral pathogen

1.1 HIV/AIDS: A still unresolved problem

Discovering the Human Immunodeficiency Virus (HIV) in the early 1980's as causative agent of the Acquired Immunodeficiency Syndrome (AIDS) was a result of independent work done by Robert Gallo and Luc Montagnier. Viral isolation was done by Montagnier's research group while the association to the syndrome was found by Robert Gallo's group [2]. The death rate due to HIV had been increasing very rapidly until the year 1995, when the drug saquinavir was introduced in antiretroviral therapy, resulting a sudden decline despite a steadily increasing number of new cases of people living with the virus [3]. According to the World Health Organisation, an estimated 36.7 million of people were living with HIV/AIDS in 2016, while 1 million died the same year worldwide [4]. The development and use of protease inhibitors (PIs) in therapy is a very good example of what can be achieved with the help of computer-guided rational drug design [5]. Evolution and progression in HIV treatment is quite unique in the history of medicine [6], in the sense that despite the development of several drugs with HIV inhibitory activity, the virus consistently mutates and selects for drug resistance mutations under the selective pressure of antiretroviral therapy. As such there is no current HIV cure [7]. At the time of writing there are 32 FDA-approved inhibitors, with the exclusion of combination products and retracted ones [8]. Due to continuous adaptation and differing levels of toxicity to patients, this pool of drugs is in fact limited. Complications arising from PI usage can be relatively benign, with side-effects such as skin rashes and gastrointestinal dysfunctions, but can also be quite severe with possibilities of intra-cranial haemorrhages [9] and hepatotoxicity [10]. HIV continuously mutates and adapts to every drug it is exposed to during treatment with antiretrovirals (ARVs). Relatively recent work by Cueva and colleagues in fact report "extremely high" mutation rates in HIV-1, on the order of 4×10^{-3} for each base per cell, from *in vivo* assays using peripheral blood mononuclear cells (PBMCs) [11]. In the same work, they show that it is the host apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3 (APOBEC3), which accounts for most of what is termed "hypermutation" - 98%, against 2% mutations caused by error-prone replication [12] resulting from the lack of proof-reading activity from the viral reverse transcriptase. Opinions diverge with respect to the impact of hypermutation as some report complete viral inactivation [13] while others suggest its use by the virus in a mechanism for evading recognition by the host immune system [14], even being referred to as a facilitator in the development of drug resistance [15]. This said, the retrovirus tolerates an astounding amount of genomic modification while remaining viable, even though fragility has been observed in the capsid protein, where multiple mutations resulted in replication-defective virions [16] *in vitro*. This genetic malleability with the associated complexity of drug resistance mutation (DRM) patterns have consistently thwarted any attempt at finding a permanent cure. In the wait of such, the virus can be slowed down by increasing the diversity of ARVs and/or improving the use of current drugs to maintain appropriate viral suppression. Uninhibited, the virus is estimated to produce 10^8 to 10^9 new virions per day [17]. Fortunately, after being approved by the FDA for use in treatment 1995, the first PI saquinavir [18] resulted in a significant drop in the number HIV-related deaths [19]. Since then, many other PIs have been designed along the similar peptidomimetic scaffolds until the development of tipranavir, which is the first non-peptidomimetic [20, 21]. Due to the similarities in drug design, it is not uncommon for some strains to be multi-drug resistant after the long history of using ARVs [22, 23]. Strains harbouring DRMs from drug-treated patients can be passed on to drug-naive individuals via various infection routes and can thus reduce treatment options. In fact, a 2003-2015 study reported ARV drug resistance in an estimated 8.3% to 15% of treatment-naive individuals [24]. The establishment of viral latency (viral genome integration) within an infected person marks the potential for chronic infection [25] and the impossibility of ARV drugs to completely clear out the virus [26] from the infected individual. Numerous strategies are used to assist in slowing down the permanent impairment of the immune system (Acquired Immune Deficiency Syndrome) [27] and are discussed in section 1.4. The damages associated with HIV are not limited to human physiology, but also extend to patient mental and social well-being, which are very much intertwined with confounding factors such as poverty and level of education [28] that can certainly play a role in getting access to ARVs in treatment. All these factors make the virus a constantly elusive target against treatment.

In this work, we focus on the improvement of current treatment approaches and finding characterising features that can assist in improving the understanding of drug resistance. We therefore investigate *in silico* various methods of optimizing current ARV treatment, highlight potentially strong elements associated with drug resistance and additionally propose new potential molecular scaffolds for the design of novel PIs via high-throughput virtual screening. In the following sections we describe in varying levels of detail important characteristics of the retrovirus and provide information on the state-of-the-art in current treatment strategies and research challenges.

1.2 The life cycle of HIV: Opportunities for viral suppression

Knowledge of HIV's life cycle is crucial for the development effective HIV counter-measures. The virus is mainly targeted by interfering at critical points in its life-cycle via antiretroviral therapy, and the increasing number of solved HIV protein structures has been very beneficial in accelerating rational drug discovery in HIV. The mature retroviral virion (Figure 1.1) is shielded by a host-

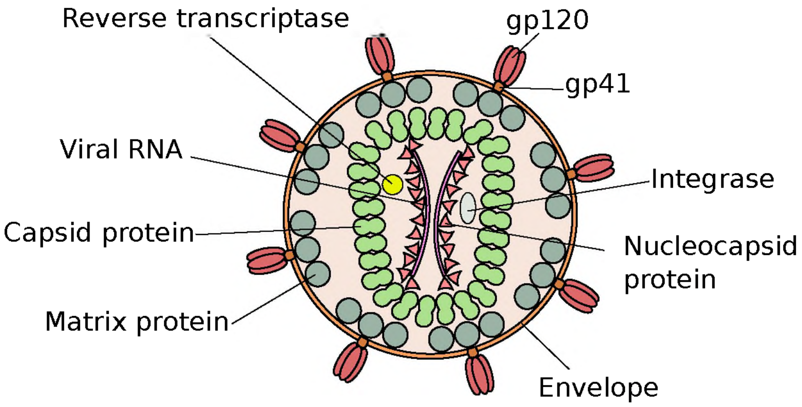


Figure 1.1: A detailed view of the mature viral particle. The capsid proteins (in green) mainly form hexameric assemblies and few pentameric "defects" at the top and bottom of the capsid, forming a protective fullerene cone around the genomic material [31]. Adapted from [32].

cell-derived bilayered outer envelope decorated by transmembrane glycoprotein complexes [29] and other host proteins such as the major histocompatibility antigens [30]. Immediately found underneath is the matrix shell, composed of an estimated 2000 copies of monomeric matrix (MA) proteins, which in addition other functions [33], assists in guiding immature viral poly-proteins to the plasma membrane prior to viral assembly [34]. Located further inside is the conically-shaped protective structure referred to as the capsid (CA), which is also composed of about 2000 monomers [34], which together house viral genetic material, stabilizing nucleocapsid (NC) proteins, essential enzymes (protease (PR), reverse transcriptase (RT) and integrase (IN)) together with several accessory proteins [29]. The genomic information is packaged as two copies of single-stranded RNA molecules stabilized in the form of ribonucleoproteins by NC proteins [29]. The NC has been shown to play a vital role in stabilizing and protecting nucleic acid sequences from various forms of degradation [35, 36]. Its genome is approximately 9.7 kilobases in length (Figure 1.2) and encodes fifteen distinct proteins [37] that are terminally-flanked on each side of the genome by long terminal repeats [29]. These fifteen proteins can be simply regrouped into structural, enzymatic

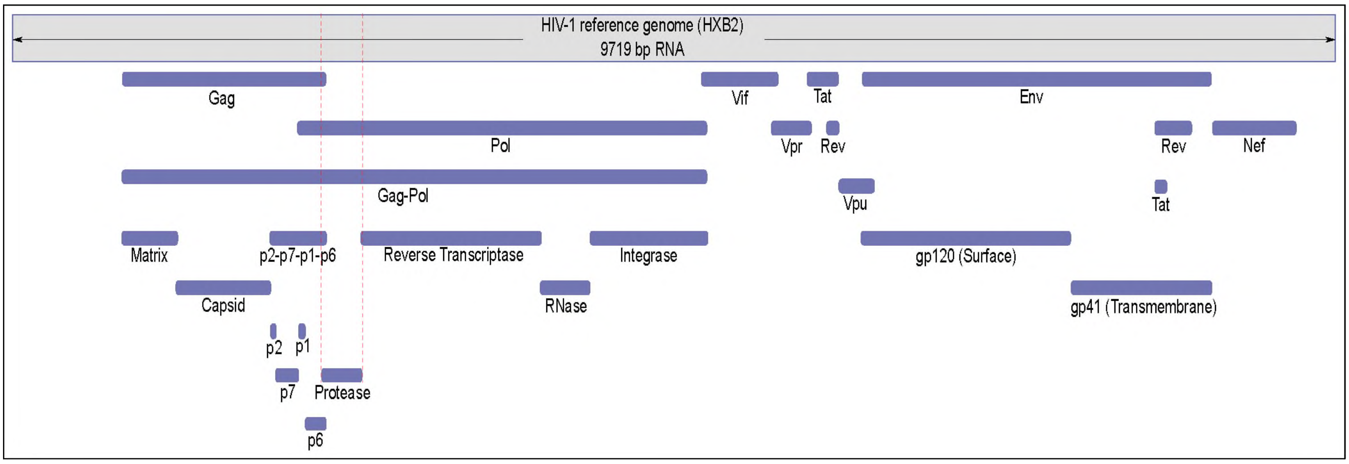


Figure 1.2: Components of the HIV-1 proteome, with respect to the HXB2 reference genome. Adapted from information generated by the HIV-1 genome browser from the Los Alamos National Laboratory.

and accessory proteins. The six structural proteins comprise the four Gag proteins (MA, CA, NC and p6) and two Env proteins (glycoproteins gp120 and gp41). The three enzymatic proteins (PR, RT and IN) are all found on the Pol poly-protein and are only active in dimeric form - RNAse forms part of RT heterodimer [38]. The enzymatic and structural proteins are synthesized as polyproteins of Gag, Pol-Gag and Env, which require catalytic cleavage in a process of maturation. Remaining six proteins can be classified as accessory (Vif, Vpr, Vpu and Nef) proteins [39] and as regulators (Tat and Rev) of gene expression [29], even though some authors classify all of them as accessory on the main basis that they were observed not to be essential for replication in cell cultures [40].

The life cycle of HIV is spent in multiple forms, namely (1) as a mature infective viral particle, (2) as a cell parasite relying on the host translational machinery for replication and (3) as a DNA fragment integrated in the host genome in wait of activation. Figure 1.3 depicts the various stages in the life-cycle of HIV, which are explained in this current paragraph. Cellular infection begins upon recognition of the envelope glycoprotein gp120 by the CD4 receptor molecule [42], which is found mainly at the surface of T-helper cells, monocytes, macrophages and dendritic cells [41]. Further interaction with a host co-receptor, namely the chemokine receptor 5 (CCR5) or the C-X-C chemokine receptor 4 (CXCR4) leads to membrane fusion and release of the full capsid into the host cell cytoplasm [32]. HIV strains may have specific co-receptor requirements, for instance using only CCR5 or CXCR4, while others can use both, by a phenomenon termed co-receptor tropism [41]. After capsid digestion, the single-stranded viral RNA is subsequently reverse-transcribed and converted to double-stranded DNA by the enzyme reverse transcriptase (RT) [17, 43]. A pre-integration complex (PIC) is formed involving HIV IN and the viral DNA

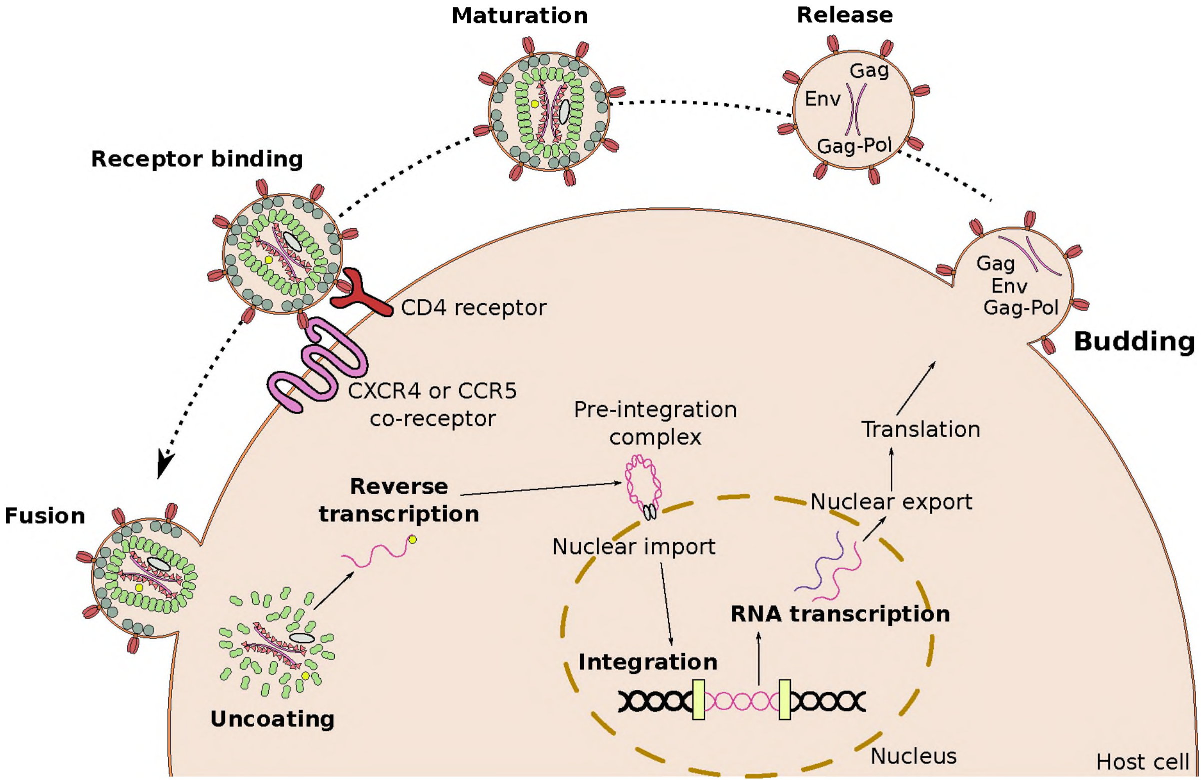


Figure 1.3: Various stages in the HIV life-cycle in and outside of the host cell, going through the process of binding to CD4 receptors and co-receptors, which lead to membrane fusion, exposure of the viral genetic material to the host cellular machinery, before undergoing nuclear integration and later releasing new virions. Adapted from [32], [41] and [42]

to be transported through the nuclear pore complex. Integrase then nicks the host genomic DNA to permanently insert the proviral genome [44]. Upon host cell activation [17], a number of viral messenger RNA is transcribed and exported out of the nucleus for translation by the host cellular machinery. This results in the production of viral poly-proteins Gag, Pol-Gag and Env that are directed to the cell inner surface together with viral RNA waiting to be released from the host cell. Finally, the virion undergoes a maturation stage during which viral aspartyl protease (PR) cleaves the proteins precursors into their functional forms. The exact mechanism for proteolysis by the retroviral protease is not known, however one of the proposed models is shown in subsection 1.5 describing features of the enzyme.

1.3 HIV classification: Strength in genetic diversity

Owing to its high mutation rate and genetic heterogeneity [45], HIV has been classified using multiple criteria, including phylogenetic relationships, genome architectures, clinical traits, virulence, infectivity and their geographic distribution [46]. At the top-level HIV is classified as two types on the basis of virulence, namely as HIV-1 and HIV-2 [47], with the former generally being more virulent and common worldwide. Both types are further delineated into groups. HIV type 1 is subdivided into 4 groups M, N, O and P, each linked to an independent cross-species transmission on the basis of their separate origins from non-human primates [48]. HIV-2 is composed of groups A-H [49]. As the focus of this research is on type-1, we describe HIV-1 classifications in more detail. Group M (where "M" stands for major or main) is the one responsible for the majority of HIV infections, while the remaining ones are mainly present in Central Africa, with a generally lower prevalence [50]. Underneath group classification, both HIV types are further differentiated into genetically-distinct clusters known as subtypes, which are a result of founder effects that have emerged at different time points in the past [48]. A founder event is described as the situation whereby a small group of individuals colonizes a new environment, which may then lead to adaptation and a reduction in genetic variation - a phenomenon known as the “founder effect” [51]. In HIV-1, this has lead to the description of 9 distinct subtypes (A-D, F-H, J and K) [49]. A depiction for the generally-accepted classification for HIV strains is shown in Figure 1.4.

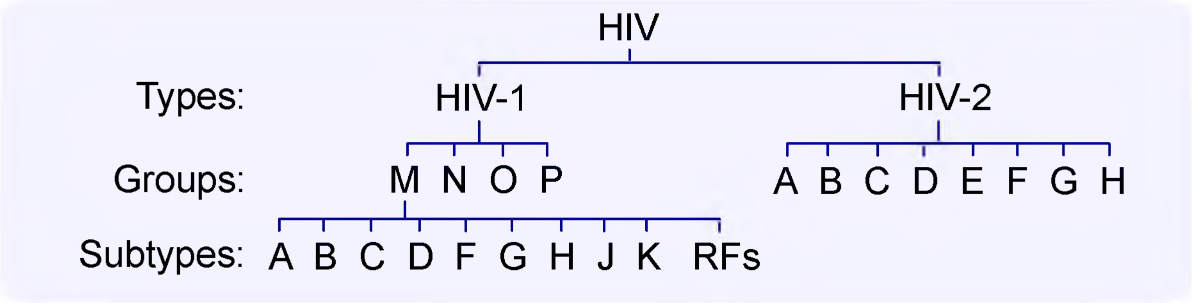


Figure 1.4: Classification of HIV strains into types, groups and subtypes. Recombinant Forms (RFs) comprise the strain hybrids. Adapted from [49]

HIV protein residue differences ranging from 8% to 30% can be observed within a given subtype, while variations on the order of 17%-42% between subtypes can prevail depending on the subtype and genomic locus under consideration [52]. Inter-strain HIV hybrids, termed unique or circulating recombinant forms (URFs or CRFs, respectively) also form part of the viral population, being referred as such based on their frequency in human population [53]. With a genetic diversity epicentre located in Central Africa, recombinant forms are widespread [49] and their presence is being increasingly felt in certain parts of the world, such as Thailand, China, Brazil and West Africa [54, 55], where they appear to be displacing the founder subtypes [54]. This said, Hemelaar reported that subtype C accounts for nearly 50% of the HIV-infected population worldwide, followed by subtypes A and B, as shown in Figure 1.5. It should be emphasized that strain subtyping has been done in slightly different ways over time, using the *env* [53], *gag* [56], partial sequences [57, 58], combinations of *gag* and *env* [59], even complete genomes [60] and later *pol* sequence data due to the availability of PR and RT sequences from drug resistance testing data [61, 62]. According to Robertson, the large majority of subtypes adopt consistent clustering patterns regardless of genomic locus, even though a certain fraction of the subtypes have been found discordant [63, 53]. HIV's extant genetic diversity and constant evolution pose a serious threat for the perpetuation of current virological suppression approaches used for treatment of patients. In order to maintain a wide array of treatment approaches, it is necessary to keep on developing novel control strategies that will resist their evolutionary adaptation mechanisms. The next section elaborates on some of these strategies.

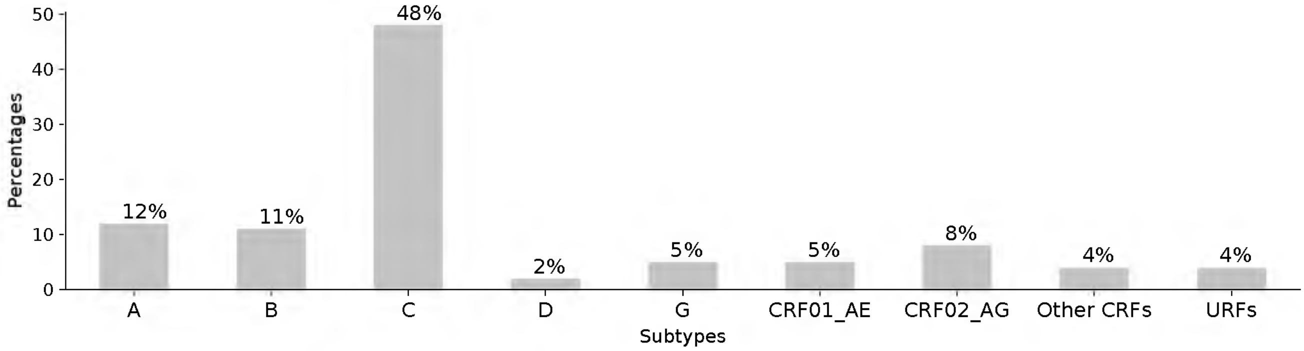


Figure 1.5: HIV subtype distributions worldwide (2004-2007 data), adapted from [49].

1.4 Current state-of-the-art in the fight against HIV

Due to the tenacity and resilience of the virus, approaches targeting HIV are manifold and there is currently no permanent cure [64]. One approach consists in the development of a vaccine by searching for broadly-neutralizing antibodies, which have shown success only in animal models and not in humans [65]. Another area of research is focused on genome-inactivation of the persistent retroviral DNA by using the CRISPR/Cas9 system via gene therapy [66]. Recent work showed successful inactivation of integrated proviral HIV DNA in latently-infected human cell lines, namely in Jurkat cells [67] and HeLa T-cell lines [68]. A main obstacle for its success is the lack of an effective delivery system for the CRISPR reagents to infected cells [66]. Most of current therapies in use involve the prescription of antiretroviral inhibitor combinations in the form of a Highly-Active Antiretroviral Therapy (HAART) [69] that target and interfere with essential steps in the life-cycle of HIV. These include drugs interfering with nucleoside reverse transcriptases (NRTIs), non-nucleoside reverse transcriptases (NNRTIs), proteases (PIs), fusion, entry and integrases (IN) [70]. The modes of action for the different classes of inhibitors are described in Table 1.1. In order to raise the barrier against drug resistance, ARVs may be prescribed together with low levels of pharmacological enhancers, such as the drug ritonavir [71] or its analogue cobicistat (COBI), which mainly inhibit cytochrome P450 drug metabolism [72] and are especially used in second and third line regimens [41]. The sheer number of drug targets and molecules (Table 1.1) exploited over the years relates to the amount of effort required to keep up with this constantly-adapting single pathogen. Unfortunately, associated toxicities can be a limitation of drug efficacy in certain cases. Various treatment strategies of prescribing these drugs have been evaluated for prolonging their usable lifetimes, for which some opinions diverge in the scientific community, as resistance mechanisms are not completely understood. Additionally, diagnosis and treatment do differ on the basis of resources availability. For instance, monitoring CD4 cell counts and the use of viral load assays can guide clinicians as to how patients respond to ARV treatment [73], however assistance from sequence-based methods greatly facilitates and speeds up the drug selection process [74]. Certain regimen recommendations can vary according the guidelines being followed [75], some examples of which are those from the International AIDS Society (IAS), the World Health Organisation (WHO), the European AIDS Clinical Society (EACS) and the US Department of Health and Human Services (DHHS) [76]. In an effort to improve our understanding of outcomes from complex mutation patterns, a panel of experts from the International Antiviral Society-USA (IAS-USA) regularly updates HIV drug resistance information with curated data in order to better guide HIV clinical practitioners [77, 78, 79, 80]. This said, a variety of ARV formulations based on nano/micro-particle drug delivery systems are being evaluated as a means to enable sustained slow-release and improve the ARVs half-life, with the objective of targeting HIV reservoir areas (central nervous system, gut-associated lymphoid tissue, lymph and spleen) [81]. Meanwhile the use of preventive medication such as the pre-exposure prophylactic (PrEP) drug holds a lot of promise in limiting the incidence of new HIV cases, especially for high-risk individuals [82]. In parallel, initiatives such as the Stanford HIVdb store and make publicly-available information and tools to assist in performing drug resistance predictions from sequence data [83] thus expediting progress in the field of HIV drug resistance research. While phenotypic assays can be used to reliably predict drug resistance for making more informed decisions in the treatment of HIV patients, they are expensive and laborious. Cheaper *in silico* proxies (genotypic assays) are instead routinely used when sequence data is available, using web servers such as the Stanford HIVdb, REGA and ANRS. As more resistance-associated mutations are uncovered over time, the algorithms used to assess the impact of drug resistance mutations are updated [84]. At the same time, independent research groups also strive to improve current prediction performances, especially as machine learning tools mature and as more labelled data becomes available. Further details around *in silico* genotype-based predictions are given in chapter 2, where we present the development of neural network models to improve predictability in HIV PR and RT.

Table 1.1: Description of inhibitors used in the treatment of HIV patients.

ARV class	**Mechanism of action**	**Drug**	**Additional information**
PI	Competitively inhibit the HIV protease active site, which prevents cleavage of the precursor proteins, thus hindering viral maturation [85].	atazanavir (ATV) darunavir (DRV) fosamprenavir (FPV) indinavir (IDV) lopinavir (LPV) nelfinavir (NFV) ritonavir (RTV) saquinavir (SQV) tipranavir (TPV)	amprenavir precursor. Pharmacokinetic booster.
NRTI	Competitively inhibit HIV reverse transcriptase [86] and interrupts DNA synthesis by incorporating viral DNA, leading to chain termination [41].	tenofovir (TDF) lamivudine (3TC) abacavir (ABC) emtricitabine (FTC) zidovudine (AZT) didanosine (DDI) stavudine (D4T)	In certain circumstances, some of these drugs have been associated with serious side-effects, sometimes being fatal [87].
NNRTI	Bind away from the active site, causing conformational changes to inhibit reverse transcriptase [41].	efavirens (EFV) etravirine (ETR) nevirapine (NVP) rilpivirine (RPV)	
INI	Prevent the formation of covalent bonds between the host and viral DNA [70].	raltegravir (RAL) dolutegravir (DTG) elvitegravir (EVG)	
Fusion inhibitor	Interferes with the fusion of viral and cell membranes by binding with part of the gp41 glycoprotein [88].	enfuvirtide (ENF)	Limited use in multi-drug resistant patients mainly due to the requirement for subcutaneous injections and the frequency of painful side reactions [71].
Entry inhibitor	Selectively binds to the human CCR5 receptor to prevent gp120 attachment [70]	maraviroc (MVC)	Ineffective against HIV showing CXCR4 or dual tropism [70]
Post attachment Inhibitor	Prevents viral transmission from cell-cell fusion by binding domain 2 of the CD4 receptor, hindering post-attachment of HIV viral particles [89].	ibalizumab (IBA)	Fortnightly intra-venal injections.

1.5 Characteristics of the chosen drug target: HIV protease

Antiretrovirals of the PI class have come out as an essential part of HAART [69] changing the tides in HIV treatment since the introduction of saquinavir, the first in line amongst a series of first-generation PIs to be approved by the FDA [90]. The target enzyme is a 22 kDa [91] homodimer composed of two 99-residue long chains, with an approximate 2-fold rotational symmetry [92, 93], in other words the dimer appears very similar when rotated 180 degrees. It forms part of the aspartic peptidase domain super-family of proteins, which is widely distributed across multiple domains of life, encompassing organisms such as vertebrates, plants, protozoa going through prokaryotes and certain viruses [94]. In HIV, the enzyme is crucial for the proteolytic cleavage of precursor poly-proteins and hence is vital for the process of viral maturation [95]. The enzyme relies on a pair of catalytic aspartic acid residues situated at position 25 from each monomer [92] to catalyse the breakdown of asymmetric peptide substrates [96]. Each ASP25 residue forms part a catalytic triad (ASP25-THR26-GLY27), which are stabilized as loops by a network of hydrogen bonds to form a conformation referred to as the fireman's grip [97]. Overall the monomer topology is mainly composed of a β-barrel and an α-helix, which is usually observed in pepsin-like proteases such as renin and cathepsin D from the super-family of acid proteases [98]. These secondary arrangements have been further described as functional units in HIV protease, akin to a mechanical system actioned by levers. These mainly consist of the flaps, cantilever and a fulcrum, as shown in Figure 1.6. Less-structured segments comprising the elbow, 10's, 60's and 80's loop regions mainly connect to β-strands. Being highly-selective, the aspartyl protease achieves catalysis by

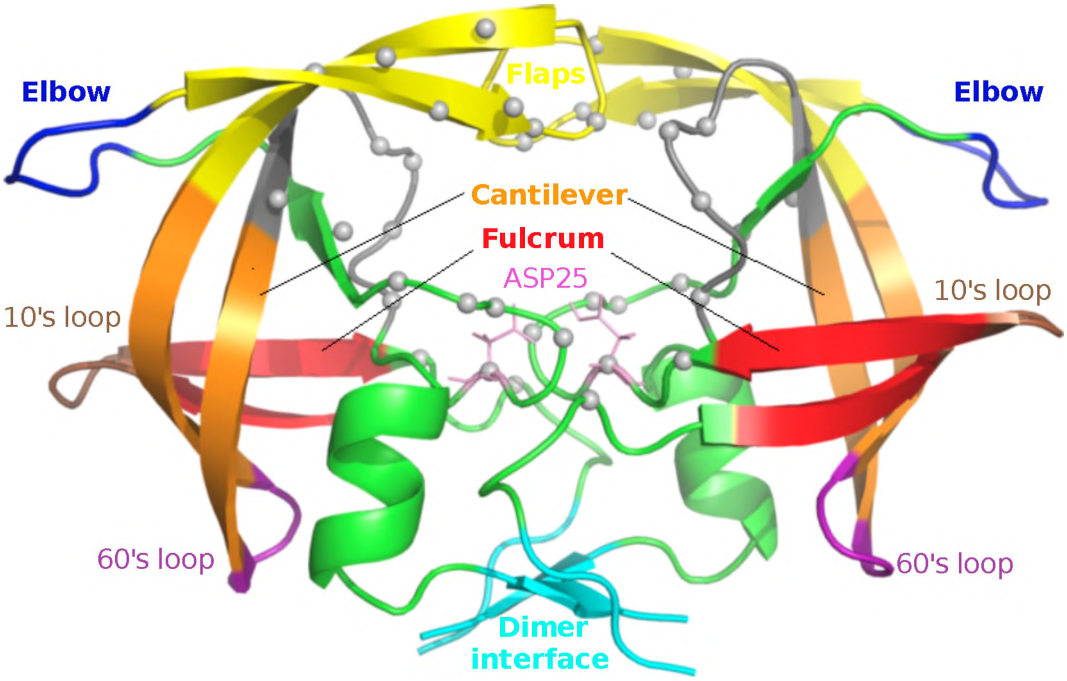
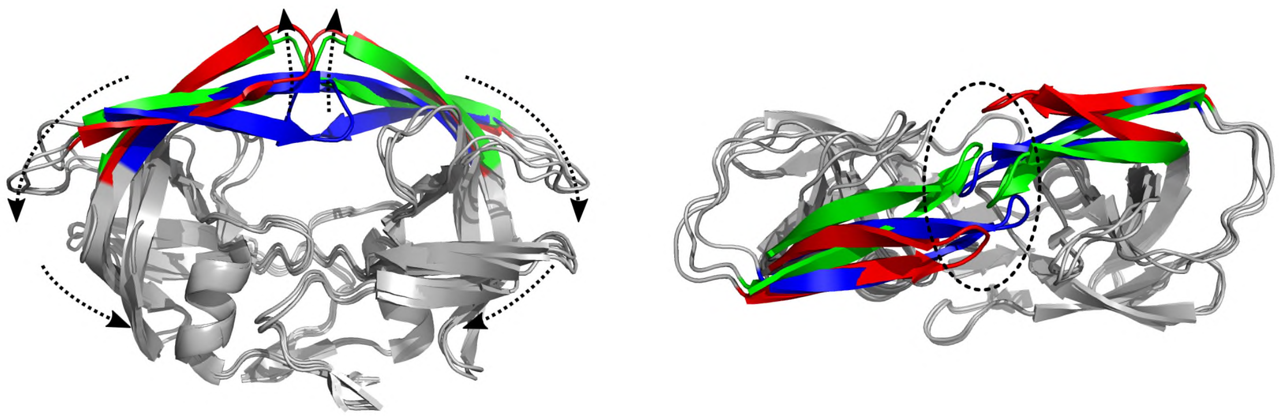


Figure 1.6: Mapping of functional elements from HIV-1 protease. The coloured segments depict the flaps (yellow), cantilever (orange), fulcrum (red), elbows (blue), dimer interface (cyan) and 10's loop (brown), 60's loop (purple) and the 80's loop (grey). Grey spheres denote residues from the binding cavity while the catalytic aspartates are coloured pink. Figure re-used from [99]

performing a nucleophilic attack on the scissile amide linkages within their peptide substrates [17]. The exact mechanism of proteolytic cleavage is not known, but is proposed to occur as shown in Figure 1.8, via a concerted mechanism involving the polarization of water by the catalytic residues, causing a nucleophilic attack on the targeted peptide bond. The substrate cleavage site itself is an octapeptide which bears no clear motif conservation, and is usually represented as P_4-P_3-P_2-$P_1 \downarrow P'_1$-P'_2-P'_3-P'_4, where the down arrow represents the scissile bond [100] amongst the residues denoted by P, which each correspond to cavity regions referred to as sub-sites. Due to discordances in the exact definition of cavity sub-sites [101, 102, 103], the terminology is not used in this study. The use of protein mimicry (peptidomimetics) was a natural choice for inhibiting the catalysis of peptide bonds in the process of rational drug design for the development of substrate-like PIs [104] based on a hydroxyethylene or hydroxyethylamine substructure [105]. This long tunnel is formed by the assembly of two identical subunits [106] and can allow the flaps to move by up to 7 Å away from each other upon substrate association [97]. The closed receptor conformation is typically observed in the presence of a bound substrate or inhibitor, while the opened conformation allows ligand entrance and release [95]. A semi-opened conformation is also described as an intermediate between the two, as shown in Figure 1.7 with green flaps in the midst of the other two conformations. The strategy of designing PIs which compete for the active site by masquerading as the transition state of the natural protein has proved to be exceptionally successful in achieving tightly-bound PI complexes of high-affinity [95]. Out of all the FDA-approved PIs, tipranavir is the only non-peptidomimetic drug, being based from lead optimization of coumarin and pyrone derivatives [111]. As opposed to the peptide-mimicking counterparts, non-peptidomimetic drugs do not require an interfacial water molecule for stabilization by hydrogen bonds around the ILE50 residues [112].



(a) Side view.

(b) Top view.

Figure 1.7: Depictions of the closed (PDB ID: 4TVG [107], blue flaps), semi-opened (PDB ID: 1HHP [108], green flaps) and wide open (PDB ID: 1TW7 [109], red flaps) HIV protease conformations superimposed, with top and side views, showing the general direction of motion involved in receptor opening (left) and the inter-flap spacing (right). Adapted from [110]

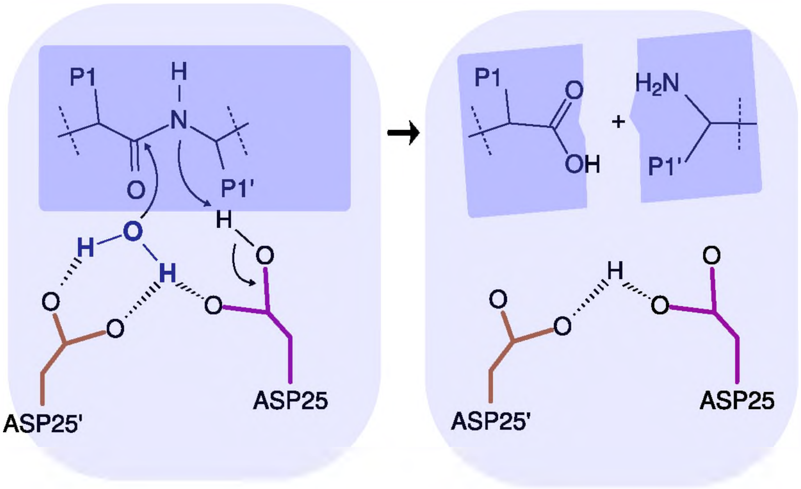


Figure 1.8: A proposed mechanism by Jaskólski and co-workers for the concerted proteolytic cleavage of peptide bonds by the HIV protease. Catalytic aspartates are shown in purple and brown, while the water molecule is coloured in dark blue. In this model, water is polarized by catalytic aspartate oxygen atoms to cause a nucleophilic attack the peptide bond, following which a proton is transferred from one of the aspartates to the N-terminal amine group of the cleaved peptide. Adapted from [113, 17]

1.6 Challenges in HIV research

Our target pathogen is a tremendously resilient retrovirus endowed with an efficient, rapid and constant adaptation, owing to various survival strategies. (1) Its existence as a diverse viral pool within infected individuals means that any class of enzymatic inhibitor has to remain potent against a large number variants in order to impart adequate virological suppression. (2) Its permanent establishment as part of the host genome as latent reservoirs forces a life-long treatment regimen upon patients, who have to face the additional challenges of drug tolerability, toxicity and eventually adherence. Due to latency period, many infected individuals may not be aware of their infection status, thus allowing the virus to prevail unnoticed. Such events facilitate the spread of drug experience to drug-naive individuals thus narrowing initial treatment options. (3) Inter-subtype viral recombinations (mosaicism) form novel hybrids [114], which add to the resilience and complexity of the quasispecies dynamics.

While genotype-based drug resistance prediction approaches are manifold, they are all mainly based on a single subtype, which is not reflective of the worldwide distribution, and hence provides an opportunity for improvement, which is unfortunately limited by the insufficient diversity in publicly-available labelled drug resistance datasets. DRM patterns are complex and thus disagreements between various genotypic methods can occur, even though such differences tend to lessen with time. Viral sequences are typically generated using Sanger sequencing or high-throughput technologies, for which there is a trade-off amongst several factors involving sequence diversity, quality and cost for the assays, given non-homogeneous samples simultaneously containing low and high-abundance variants in the form of a quasispecies [115, 116]. Template re-sampling errors can also originate from the polymerase chain reaction, which can make samples appear more homogeneous by repetitively amplifying original templates [117, 115]. Independent to these biological and experimental factors, socio-economic problems such as lack of education and reduced access to ARVs can contribute to the unabated proliferation of the virus.

1.7 Research objectives

Our main objectives consisted in contributing to the body of research in the fight against ARV drug resistance in HIV. While initially targeting non-B HIV strains, the amount of labelled publicly-available data was insufficient for our purposes. Therefore, a significant portion of the work was laid on finding very strongly-conserved signals from subtype B instead, in the hopes that such could extrapolate or establish a methodology applicable to our initial target strains should such information be available later.
In Part I, we do so by aiming to accomplish the following:

1. Improve the accuracy of drug resistance prediction in HIV-1 protease and reverse transcriptase for FDA-approved ARV drugs.

2. Seek for a characteristic resistance-associated signal from structural data in HIV proteases able to tell apart resistant sequences from susceptible ones with high performance and precision.

3. Discover novel scaffolds for use in the design of anti-HIV ARV drugs using high-throughput *in silico* screening of natural compounds.

Part II deals with several side projects performed in collaboration with members of the RUBi research group, pertaining to method and tool development in structural bioinformatics.

Chapter 2

Improving fold resistance prediction of HIV-1 against protease and reverse transcriptase inhibitors using artificial neural networks

This chapter draws from and reproduces certain figures and tables used in the publication listed below. Credit for the reproduced material is given as citations in the respective figure and table captions.

- **Sheik Amamuddy O**, Bishop NT and Tastan Bishop Ö. "Improving fold resistance prediction of HIV-1 against protease and reverse transcriptase inhibitors using artificial neural networks" *BMC Bioinformatics*, 2017 August 15. doi: 10.1186/s12859-017-1782-x.

2.1 Introduction

In this chapter, we introduce the use of neural networks for the improvement of prediction of drug resistance in subtype B HIV-1, more specifically against protease and reverse transcriptase inhibitors. This method was chosen as a cheap, fast and accurate way of improving over what is used to assist physicians in prescribing more optimal regimens to patients. HIV patients would typically have HIV enzymes sequenced from blood samples followed by a genotypic assay [118], which characterizes the effectiveness of the antiretroviral to prescribe. The gold standard is the phenotypic assay [119] whereby the sequenced HIV *pol* gene fragment is cloned into a recipient HIV virus later used in cell cultures exposed to different drug concentrations to determine an inhibitory concentration. The concentration required to inhibit viral proliferation is compared against that of the wild type virus. Depending on the determined fold resistance ratio obtained with respect to defined cut-off values, a sequence may be deemed susceptible, resistant or partially (intermediate) resistant to a given drug. Further classification denominations are used by different prediction methods. A variety of genotypic resistance prediction methods exist and are usually implemented as web servers. Some examples of which are Geno2pheno [119], Stanford

HIVdb [83], REGA [120], the "Agence Nationale de Recherches sur le SIDA" (ANRS), SHIVA [121], amongst many others. Geno2pheno defines a drug resistance state (susceptible or resistant, with a probability score) using machine learning techniques (decision trees and support vector machines) trained on genotype and phenotype correlations [122]. HIVdb assigns a drug resistance state (susceptible, potential low-level, low-level, intermediate or high-level) by obtaining a total drug resistance score from the of sum penalty scores derived from literature for each residue difference from the reference B subtype for each separate ARV drug [123]. ANRS defines a drug resistance state (susceptible, possible resistance or resistant) using a set of rules derived largely from genotpe/phenotype correlations based from a large database of ARV-failing patients to produce a tabulated list of resistance-associated mutations [124]. REGA is a complex algorithm that defines a drug resistance state (susceptible, intermediate or resistant) based on a set of rules derived from mutations reported to be associated with resistance or reduced therapeutic effect [123]. The algorithm takes into account DRM interactions and is valid across HIV subtypes despite showing reduced effectiveness in cases of multi-drug resistant HIV [123]. SHIVA is a recently-developed web server, which specialises in interpretation of drug resistance from multiple sequences obtained via next-generation sequencing technologies for instance from a single patient and internally uses random forest models on their numerical-encoding (using the Kyte and Doolittle hydrophobicity scores) as input to give resistance predictions (resistance or susceptible) for each sequence [125, 121]. Results from these different prediction algorithms may have certain discordances in drug resistance predictions, even though this trend should be decreasing as more resistance data and information becomes available. HIV DRMs are nevertheless updated via regular publications [77, 78, 79, 80]. Accuracy of predictions can have a large impact on the patient well-being, as suboptimal prescriptions lead to faster development of DRMs, which make drugs less effective. As the pool of ARVs is limited, optimized drug prescription is a must, not only for improved patient lifestyle but also for reduced risks of viral transmission, for instance in cases of shared syringes amongst drug users and in mother to child transmissions. In addition, preliminary analysis of the Stanford HIV protease and reverse transcriptase datasets used for training models to predict ARV resistance mainly consist of subtype B, which represents only 11% of the global cases of HIV infections, while at least 9 other subtypes exist, with a high prevalence of subtype C followed by subtype A [49]. Recombinants also exist and can be unique or not, whereby the former are termed "Unique Recombinant Forms" while the latter are referred to as "Circulating Recombinant Forms" [48] and are numbered sequentially [126]. Subtyping is a phylogenetic grouping [53], and subtype B is only an HIV sub-classification. One can easily predict that the effects of unobserved subtype-specific mutations will not be taken into consideration and as such prediction methods will generalize less well, especially if the evaluated sequences are very divergent. We therefore push and support the idea of making available subtype-specific sequences so that the accuracy of prediction can be increased. In this chapter, we show that by applying filtering strategies involving the removal of non-B subtypes, that prediction accuracy can indeed be increased, though at the expense of non-B sequences.

Publicly-available sequence datasets labelled with drug resistance ratios are used to train artificial neural network models for available FDA-approved ARVs. The raw dataset has various sources of uncertainty, that if mitigated can improve the predictive performance. By decreasing the chances for technical variation via various filtering approaches and focusing the analysis on the majority subtype (subtype B), we improve drug resistance prediction accuracy in several cases for protease and reverse transcriptase inhibitors, when compared to similar work done by Shen and co-workers [127], the well-established Stanford HIVdb web server and a recent prediction server, known as SHIVA [121]. We highlight strong disagreements with respect to drug resistance classifications obtained from the SHIVA web server. Overall, we demonstrate an improved regression performance for the majority of HIV PIs and NRTIs when compared to recent models developed by Shen and co-workers [127] and further show competitive classification accuracies in comparison to the web servers Stanford HIVdb and SHIVA.

2.1.1 Machine learning solutions for complex problems

Since recent years there has been a growing interest in the use of machine learning (ML) to solve complex problems due to the increasing amounts of computational power and data availability. Many of these algorithms have been open-sourced via various libraries written in different programming languages, popular ones including scikit-learn [128], keras [129], TensorFlow [130], PyTorch [131], caret [132] for the R language and also as stand-alone tools such as LIBSVM [133]. While some algorithms are implemented similarly across languages, others can also be available in a particular language. For a brief description of what is available and how the algorithms are generally used in predictive modelling, we will describe some functionality from the open-source packages scikit-learn and caret together with the commercially-available software MATLAB. Depending on the format of data at hand, predictive models can either be built in a supervised or an unsupervised manner [134]. When available, labels can be used as target values to guide the learning process using matching features, in which case the training is referred to as being supervised. A multitude of supervised learning algorithms are available, including linear regression, Random Forests (RF), k-Nearest Neighbours (KNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), naïve Bayes, decision trees and ensemble methods, only to name a few. We now give a brief description for some of these algorithms. Linear regression is the simplest method which predicts an intercept and a coefficient for one variable to predict a response - multiple linear regression (MLR) extends the idea using more than one input variable [135]. RFs are a composition of various decision tree structures each built from a random subset from the training data and uses averaging to predict classes or continuous variables [128]. The KNN algorithm uses a predefined number of neighbours on the basis of a distance measure (eg. Euclidean, Minkowski, Hamming, etc) to infer the class of a sample, but can also be used in regression [128]. SVMs work by identifying an optimal high dimensional decision boundary (hyperplane) defined by a given kernel function (eg. linear, polynomial, sigmoid, radial basis) with the objective of maximizing a separating margin between some training samples defined as support vectors [136, 128, 137]. ANNs are a class of biologically-inspired algorithms, and are discussed in detail in subsection 2.1.4. Naïve Bayes methods are a class of highly-scalable algorithms based on the application of Bayes Theorem under strong assumptions of predictor independence to mainly emit classification probabilities [137]. Ensemble methods are built by combining predictions from different simpler

models, comprising of techniques such as bagging, boosting or stacking in order to decrease model variance - this definition also includes the random forest algorithms [138].
In the absence of any useful annotation, dimension reduction methods (for instance PCA or t-SNE [139]) in combination with clustering techniques (such as k-means, DBSCAN [140]) can be applied in an unsupervised manner to a set of features with the hope of finding structure from a given dataset. In the same way, hierarchical clustering can also be employed with a variety of linkage approaches before applying branch-cutting methods to find partitions within data points without explicitly-defined labels. In scikit-learn, these algorithms are laid out in a systematic object-oriented design, which requires creation of a parametrized model object, followed by applying a *fit* and *predict* function to perform the learning and prediction respectively on featurized samples. Training performance can then be assessed using separate functions, for instance regression and cross-validation. Proper data partitioning into training and testing sets is expected from the user. Similarly, the caret package from the R programming language uses a set of procedures, featuring a *train* and a *predict* function. ML algorithms are then simply defined as parameters to the train function together with cross-validation parameters. Same would be accomplished for a neural network using the *newff*, *train* and *sim* functions from the Neural Net add-on from MATLAB, and is used in this chapter.

2.1.2 Applications of neural networks in biological research

An artificial neural network (ANN) is internally nothing more than a complex regression algorithm amongst a list of various machine learning (ML) algorithms. ANNs take a series of input vectors, also know as features to match them up at an acceptable error tolerance to one or more expected target values (labels). Each feature set captures some description of a given input object, which can for instance be a protein sequence and the features would have some characteristic associated to it, for instance amino acid composition (as described in this chapter), as linearised 3D features (such as the Delaunay triangulation as used in Shen and co-workers [127]) or any other accurate descriptor of an object. Recent work has also used chemical descriptors as features to describe actual drugs in an attempt to infer biological activity from new compounds, in an approach known as the Quantitative Structure-Activity Relationship (QSAR) [135]. ANN-based models have found high-impact applications in the medical field, for instance in the non-invasive detection of malignant tumours from mammograms [141], identification of treatable diabetic retinopathy from tomography images [142], tuberculosis detection from radiographic images [143], modelling cellular growth and function from genotype [144], and many more.

2.1.3 Recent applications of ANNs specific to HIV research

ANNs have also been used in multiple contexts within the field of HIV research over the years, with interesting mentions of earlier studies on drug resistance in HIV-1 protease performed by Bonet and co-workers who trained recurrent neural networks with amino acid contact energies as features to predict resistance against 7 protease inhibitors [145], and work by Fogel and co-workers, who trained ANNs to identify dual-tropic HIV-1 (i.e. strains able to bind any of 2 co-receptors for cell entry) using various structural, biochemical and regional annotations as feature vectors for a dataset of 1559 HIV subtype B sequences [146]. We describe a few more recent use cases of ANNs in HIV research in the following paragraphs.

Otange and colleagues achieved clinically-acceptable performance ($R^2 > 0.9$) estimates of viral load concentrations using Raman spectral peaks obtained from HIV-1 p24 antigen spiked blood samples as inputs for training ANNs [147]. As an underlying technique, Raman spectroscopy detects analytes based on the scattering patterns of monochromatic light associated with characteristic molecular vibrational modes [147]. Whilst displaying a high performance, the viral titre estimation is additionally reagent-free, cheap and fast.

Dwivedi and Chouhan have employed a special kind of ANN, termed a radial basis function ANN to classify CRF and non-CRF HIV-1 strains with high accuracy using features of length 64 (to represent codon composition) from complete genomes with 10-fold CV [148].

Lu and co-workers applied one-hot encoding onto HIV-1 protease cleavage datasets (a form of binary feature representation comprising one non-zero bit within a longer vector of zeros) to train deep neural networks in order to predict cleavage sites of the HIV protease [149]. The deep learning models used in their study involved a Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN) and a Long Short Term Memory Network (LSTM). These ANN models all include a high number of layers, and even more complex node layouts as is the case for LSTMs. RNN and LSTMs gave highest classification accuracies, in the range of 92-96%.

Hu and colleagues developed a deep learning framework named DeepHINT, able to accurately predict HIV-1 genomic integration sites from host DNA, while also providing mechanistic information explaining the detected sites [150]. Known integration sites together with flanking upstream and downstream nucleotides with one-hot encoding were generated prior to training.

Barzegar and co-workers developed an ANN-based QSAR workflow for predicting HIV-1 reverse transcriptase inhibition $[-log(IC_{50})]$ based on a 1459 inferred molecular descriptors from 40 pyridinone scaffold derivatives [151]. A regression performance of 0.92% was observed from the test set.

Buiu and co-workers performed preliminary work to predict the neutralising ability (IC$_{50}$ values) independently for several antibodies using 4907 entire HIV-1 Env amino acid sequences, which were aligned and numerically-encoded for use as training vectors in feed forward ANNs [152]. The dataset was divided into 3 parts - 75% for training, 10% for testing and 15% for validation, repeated 100 times before returning the most generalisable model, based on the correlation value.

2.1.4 Neural network fundamentals

At the heart of neural networks is the application of a non-linear function to the dot product of the input and a series of weight matrices, followed by back-propagation of mean error. In their early beginnings, ANNs relied on binary step functions as “squashing” transforms from input data, which failed to give consistent results thus causing research interest in the field to decrease. However, replacing the hard step function by smooth non-linear functions turned the tides as training errors could be fed back into the model and minimized by including derivatives of the so-called sigmoid functions. These include the logistic, hyperbolic (inverse) tangent, softmax, in addition to linear transforms and the more recent Rectified Linear Unit (ReLU), which finds applications when the number of layers and nodes becomes very large (deep neural networks). In Figure 2.1, we show the simple layout for the construction of a feed-forward neural network. In this schematic (read from left to right), the input matrix containing 3 input features (vectors) are dotted against a weight matrix before being scaled by a sigmoid function. This forms the first layer of activations for 3 neurons. In order to produce 2 outputs per input vector, another weight matrix (with 2 columns) is dotted against this activated neuron and once again transformed - in this case a linear function is used by some implementations.

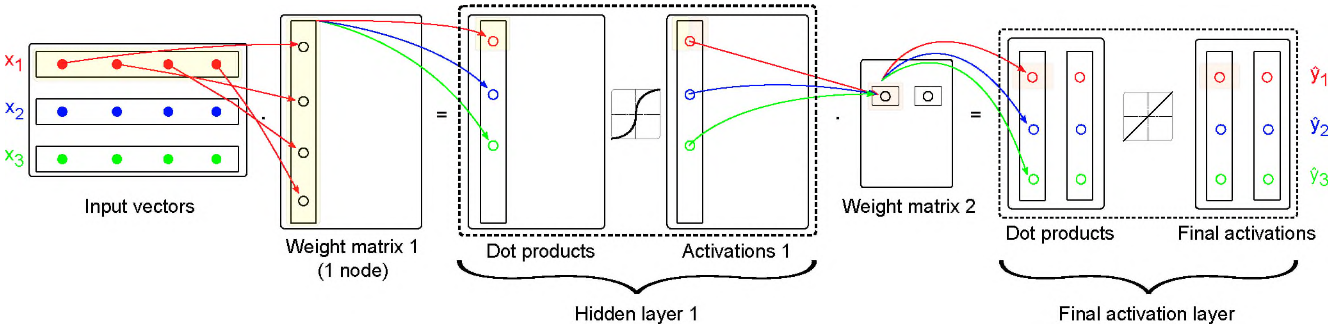


Figure 2.1: Example of a feed-forward ANN architecture with 1 hidden layer of 1 node, used to predict 2 target values from each input feature. Every single dot is a real number and is coloured to show the flow of dot products from each input vector to subsequent layers. A sigmoid and a linear function are also included. In-place operations are surrounded by dotted lines. Two neurons are also shown in yellow and red hues. The first column from the inputs are biasing coefficients, which are typically assigned a value of one.

After the first pass, initial prediction values, here labelled as ($\hat{y}_i$), are generated. In order to adjust them towards expected output values (for instance y_i), the back-propagation algorithm is used to minimize the mean square error (equation 2.6). Many variations of the gradient descent algorithm are available to address certain weaknesses that may be encountered, such as saddle points [153] and learning rates which can either lead to slow convergence [154] or even overshoot the error surface. Some examples of minimization algorithms (optimizers) for the error function are the stochastic gradient descent with momentum [155], Adagrad [156], Adadelta [157], RMSprop, Adam [158] amongst many more [154]. They all mainly affect the basic gradient descent algorithm. The generic algorithm is defined below (equation 2.1):

$$w' = w - \alpha . \frac{\partial P}{\partial w} \tag{2.1}$$

where P is the error function and w' is the updated weight after subtracting the derivative with respect to the weight. Parameter α is the learning rate, a factor which specifies the rate of convergence towards a set target error. In order to implement this weight update, the error is back-propagated from the last layer towards the left-most weight matrix, by using the analytical derivative, which involves the chain rule to unwrap the derivatives from function compositions. Going against the gradient along a logistic or hyperbolic tangent makes this process convenient

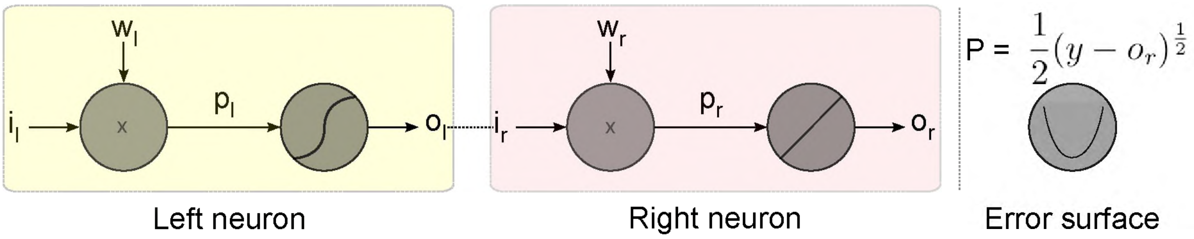


Figure 2.2: Two selected neurons showing the flow of input across the multiplication operator and two activation functions (logistic and linear). The letters i, w, p, o and y represent the input, weight, product, activation and the expected output respectively. The subscripts l and r jointly refer to being an attribute of the left and right neurons. Note that i_r is synonymous to o_l. Adapted from Patrick Winston's OCW lecture material on Artificial Intelligence [159].

as the updated weights are obtained by directly subtracting multiples of the same input weights. As an example, we show the back-propagation mechanics using the two highlighted neurons and a logistic transfer function that are used to apply equation 2.1 for weight updates from Figure 2.1. The two neurons are redrawn in Figure 2.2: In order to calculate the first order partial derivatives of the cost function P with respect to all weights, the latter are considered separately, as they belong to separate layers. Starting from right to left, we evaluate $\frac{\partial P}{\partial w_r}$ and then $\frac{\partial P}{\partial w_l}$, only stopping when the concerned weight is reached as shown in equations 2.2 and 2.3. In order to increase the depth of the neural network, equation 2.2 is applied to the final layer, while 2.3 is used to all the preceding layers. Reused computations are coloured in blue font. After one back-propagation sweep, the updated weight information is fed forward and the mean squared error is calculated. The cycle is iterated for a certain number of epochs until a defined maximum error tolerance.

$$\frac{\partial P}{\partial w_r} = \frac{\partial P}{\partial o_r} \times \frac{\partial o_r}{\partial p_r} \times \frac{\partial p_r}{\partial w_r} \tag{2.2}$$

$$\frac{\partial P}{\partial w_l} = \frac{\partial P}{\partial o_r} \times \frac{\partial o_r}{\partial p_r} \times \frac{\partial p_r}{\partial o_l} \times \frac{\partial o_l}{\partial p_l} \times \frac{\partial p_l}{\partial w_l} \tag{2.3}$$

As feature components can be of different magnitudes, these are typically scaled using various methods to improve the rate of convergence towards the target error. Two examples of such methods comprise the standard scaler (equation 2.4), which normalizes the data according to the normal distribution, with mean of zero while the minmax scaler (equation 2.5) scales data in the range [0,1]. The σ function is given in equation 2.8

$$\text{z-score}(x) = \frac{x - \bar{x}}{\sigma(x)} \tag{2.4}$$

$$\text{minmax}(x) = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2.5}$$

2.1.5 Evaluation of model performance

ML models are mainly trained to emit real numbers or integers, in which case they are termed regressors and classifiers respectively, the choice of which directly influences the evaluation metric to be used. Additionally, a trained model has to generalize well given unseen data, and not only predict training samples with high accuracy. The latter case is referred to as over-fitting. For this reason, prior to fitting the data is first partitioned into 2 or 3 subsets reserved for training, validation and testing (depending on the ML model implementation). As this may be problematic when the number of data points (samples) is limited, clever approaches have been designed to make maximum use of available data. Cross-validation (CV) methods such as the Leave-One-Out (LOOCV) and k-fold CV do just this. LOOCV trains n models using $(n-1)$ samples, while evaluating against the single unseen sample each time. Similarly, in the case of k-fold CV data is randomly partitioned into k unique partitions of size (n/k) before training k times on each of the $(k-1)$ folds to test on the single unseen partition each time. In both LOOCV and k-fold CV, n and k performance values are evaluated respectively, from which the mean and variance can be calculated. As k is a smaller subset of n, it can be deduced that LOOCV demands much more rounds of computation, but finds application when the number of samples is limited.

The actual performance values used in classification and regression problems are various, however a few simple ideas borrowed from the field of statistics are routinely used for performing comparisons - metrics based on the deviation from averages in the case of real data and the use of contingency table metrics in the case of discrete data. Some of the functions used in regression problems include the mean squared error

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 \tag{2.6}$$

where y_i and $\hat{y}_i$ are the actual and predicted values respectively for an i^{th} sample. The coefficient of determination (R^2) is also widely used. One way of computing this value is by squaring the correlation as such

$$R^2 = \left[\sum_{i=0}^{N}\frac{(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sigma(y)\sigma(\hat{y})}\right]^2 \tag{2.7}$$

where σ is the root mean squared deviation function, as shown in the following equation.

$$\sigma(y) = \sqrt{\frac{\sum_{i=0}^{N}(y_i - \bar{y})^2}{N-1}} \tag{2.8}$$

For classification problems, commonly-employed metrics include the sensitivity (true positive rate), specificity (true negative rate) and accuracy. However in our case we measure accuracy's complement - the rate of misclassification. The metrics are calculated thus:

$$sensitivity = \frac{TP}{TP + FN} \tag{2.9}$$

$$specificity = \frac{TN}{TN + FP} \tag{2.10}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.11}$$

$$misclassification\ rate = 1 - accuracy \tag{2.12}$$

where the letters T and F refer to $True$ and $False$ while letters P and N refer to *positives* and *negatives*. Sensitivity and specificity effectively give the proportions of correctly-classified positives and negatives over the actual positive and negative samples respectively, while classification accuracy gives the total proportion of correctly-classified samples. Above all, any prediction model will only at best be reflective of the quality and amount of training samples used to teach it. Hence, an understanding of the problem at hand and circumstances of data collection are very often key in making drastic improvements in the quality of computational predictions. Here we use biological knowledge of HIV and features of the data to pry out and discard data points likely to increase model variance, while limiting potential bias from under-sampling.

2.2 Methods

2.2.1 Preprocessing PR and RT drug resistance data

Unfiltered datasets were retrieved from Stanford HIVdb for proteases and reverse transcriptases. These datasets were chosen over the pre-filtered ones as they contained subtype information. Preliminary checks could not match certain labels from the unfiltered dataset to the filtered one. Each entry of the records was uniquely identifiable by a sequence identifier, corresponding to a compacted sequence representation dependent on a consensus protease or reverse transcriptase sequence of the B subtype. Fold resistance ratios for each drug class were also available. Any residue deviation from the consensus were labelled by the characters '.' (no residue), '#' (insertion), '∼' (deletion), '*' (stop codon). Further, residue mixtures were denoted by two or more residues for a given position. Entries with incomplete fold resistance ratios were retained in order to increase the number of data points. To keep a total length of 99 residues (for proteases) and 240 (for reverse transcriptases) of a relatively higher quality, entries with indels were removed. Reverse transcriptase sequences were restricted to 240 residues to be on par with the filtered version of the dataset from Stanford HIVdb. Additional characters found during exploratory data analysis of the datasets (ˆ, X, d and l) were also filtered out. Sequences with mixtures were expanded to retrieve all possible sequence variations up to defined cut-offs, which were introduced initially for computational efficiency in cases where some entries would display over a million variant combinations. Non-B sequences were discarded as they were in large minority in the datasets. Each sequence entry, was then represented as sequence with a corresponding fold resistance value (the label) for the matching antiretroviral for each of the drug classes used. Every residue composing each sequence was then numerically-encoded in order to construct a feature vector representing the viral sequences, as shown in Table 2.1. The matching sequence fold resistance values were placed in an output vector, with each entry corresponding to a feature vector. The approach used bears some similarity to the method used by [160] where codons were utilized with an earlier version of the dataset [161]. Major differences here are that our numbering method has no specific biological meaning; also more recent datasets were used.

2.2.2 The training strategy

An Artificial Neural Network was optimized for each of the FDA-approved drugs (8 protease inhibitors and 10 reverse transcriptase inhibitors) mainly by varying the topologies of the hidden layers from 1 to 3 layers, with permutations of 2, 4, 6, 8 and 10 nodes per layer. For each drug, the dataset was partitioned into 3 subsets, namely: 70% were used for training, 15% for validation and 15% for testing. The testing set was not seen during training and was used to evaluate out-of-sample performance. When over-fitting was detected due to low performance in the testing set, the number of layers was reduced to one and the number of nodes were varied (from 5 to 20). In order to mitigate sources of technical variation such as sequencing error coming from the input sequences, filtering strategies were evaluated in order to decrease out-of-sample error, namely (1) the filtering of rare variants, (2) the use of Principal Components Analysis for detecting potential outliers (3) the use of sequence expansion cut-off values to limit the amount of technical variation and (4) the use of error analysis to remove samples consistently giving large errors despite the use of different random weight initializations. Rare variants are defined here as amino acid residues present only once at any homologous position across all aligned sequences used in entire dataset for each drug. Absolutely conserved positions were ignored from training to improve the rate of convergence.

Table 2.1: Amino acid numerical encoding

residue	A	R	N	D	B	C	E	Q	Z	G	H	I	L	K	M	F	P	S	T	W	Y	V
encoding	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22

As a preparation for MATLAB's neural network functions, both the input features and the output vector were transposed to be used in a feed-forward neural network model based on Levenberg-Marquardt's algorithm [162, 163]. By default the *newff* command scales the input features and removes invariant columns. Errors calculated from the differences between actual outputs and final network activations were back-propagated to adjust weights using the gradient descent algorithm. Back-propagation was repeated for a maximum of 1000 cycles, unless a minimum gradient of 1e-7 was reached or upon 6 consecutive validation failures.

2.2.3 Assessing model performance

The predictive performance of trained models were assessed using both regression and classification performance metrics, namely via the coefficient of determination (R^2) and misclassification rates respectively. In order to show any signs of over-fitting, the k-fold cross-validation was used with 5-folds over the complete dataset giving the average error and its standard deviation. These values were compared against similar metrics evaluated from the Random Forest (RF) and K-Nearest Neighbour (KNN) models developed by Shen and co-workers referred to as protocols B and C respectively [127]. As an additional quality control, the detailed breakdown of regression performances was provided for separate subsets of the complete data giving R^2 regression performances over (1) the whole dataset, (2) the validation set and (3) the test set. Misclassification rates were computed by applying appropriate cut-off values from our ratio predictions. Predictions from

Stanford HIVdb were obtained by submitting a GraphQL queries together with FASTA-formatted back-translated protein sequences (one of the coding sequences from each peptide) via the Sierra web service. Back-translation was done using *backtranseq* tool from the EMBOSS suite [164]. JSON-formatted predictions (where classes were labelled S, I and R for susceptible, intermediate and resistant respectively) were parsed and converted to numeric form 0, 2 and 1 for computing discordances. SHIVA predictions were obtained by submitting the protein sequence via their web page with default settings. Returned CSV files with labels (0,1) corresponding to susceptible and resistant were used directly for evaluating classification accuracy. In order to accommodate differences between the number of classes evaluated by Stanford HIVdb and SHIVA (3 and 2 classes respectively), available binary cut-offs were obtained from the PhenoSense Assay, while values from a Stanford HIVdb R script [165] were used for 3-class performance evaluations (As shown in Table 2.2). In the case of SHIVA, where no binary cut-off is available for certain ARVs, a lower and an upper bound were evaluated for determining misclassification rates. The proportion of truly misclassified pairs (0,1 or 1,0) was used as a lower bound while ambiguous cases (2,0 or 2,1) were counted and set as an upper bound. In each case (predictions from our ANN models, SHIVA and Stanford HIVdb), matching cut-off values were applied to PhenoSense fold resistance ratios available from the dataset before evaluating discordances.

Table 2.2: Cut-off values used for classing drug resistance. Table re-used from [99].

		PhenoSense cut-offs		HIVdb cut-offs	
ARV class	**ARVs**	**lower bound**	**upper bound**	**lower bound**	**upper bound**
PIs	FPV	4	11	3	15
	ATV	5.2		3	15
	IDV	10		3	15
	LPV	9	55	9	55
	NFV	3.6		3	6
	SQV	2.3	12	3	15
	TPV	2	8	2	8
	DRV	10	90	10	90
NRTIs	3TC	3.5		5	25
	ABC	4.5	6.5	2	6
	AZT	1.9		3	15
	D4T	1.7		1.5	3
	DDI	1.3	2.2	1.5	3
	TDF	1.4	4	1.5	3
NNRTIs	EFV	3		3	10
	NVP	4.5		3	10
	ETR	2.9	10	3	10
	RPV	2.5		3	10

2.3 Results and Discussions

Initially a single ANN model was constructed per ARV class, i.e. one for the HIV PIs and another for the RTIs. However, very large mean squared errors were obtained, corresponding to moderate correlations. We note that correlation is not a very robust metric when thousands of samples are processed as errors are diluted over the number of training samples to give a wrong impression of accuracy. For this reason, the regression models were independently built - one for each drug and correlation was replaced by the coefficient of determination (R^2) as quality metric. It was also initially assumed that higher confidence sequences would be obtained by only considering cases for which drug fold resistance ratios were available against all members of a given class (RT and PR). Even though the odds of this assumption being right could not be completely ignored, this strategy consistently reduced the number of training samples, which would decrease generalizability of the prediction models downstream. Therefore, all sequences with at least one fold drug resistance ratio were considered to increase the information content via the more diverse pool of available sequence compositions. For this reason, several filtering criteria and approaches were implemented to reduce the incidence of low-confidence sequences.

2.3.1 Preliminary filtering and training

We begin by identifying structure in the data by examining the proportion of subtypes for each unique sequence entry and find an overwhelming proportion of subtype B sequences. Subtyping is a taxonomic denomination which clusters similar sequences and separates more divergent ones. On this basis we proceeded with increasing the accuracy for subtype B at the expense of decreased generalization to non-B subtypes. For each compactly-represented sequence from the raw dataset, if no ambiguous residue was present, it was reconstituted from 3 dimensional arrays. The first dimension corresponded to the sequence ID, the second represented residue positions and the third contained possible residues for each given residue position. Cartesian product was thus applied to each array of residues to reconstitute the encoded sequences. To facilitate and standardize feature construction, sequences with indel mutations were disregarded such that each protease was 99-residues in length. Same was done for reverse transcriptase sequence but were trimmed down to 240 residues as is done for the pre-filtered data available from Stanford HIVdb. During sequence reconstitution, some sequences were found to have several thousands to millions of possibilities. Training with this data can bias the models towards sequences from the dominating sequence IDs and unnecessarily increase model variance if these were technical variations. Therefore, a series of cut-off values were experimented on to control the maximum number of sequence combinations to use, namely 10, 20, 50, 100, 200, 300 and 1000. Values lower than 300 were too stringent, yielding too few sequences, which would then lower the information content and generalizability of the models downstream. Therefore, we proceeded with threshold values of 300 and 1000. Numerical sequence encoding and checks for uniqueness were then performed for each separate filter level and drug. Thereafter, residues occurring once across the whole dataset (termed *rare variant* in our manuscript) were systematically removed on the assumption that very rare sequences have a higher chance of being random (sequencing) errors. Further PCA analysis was performed on the encoded sequences together with their resistance labels to infer additional outlying samples and remove them. ANN models were then trained by varying the number of hidden layers (1-3) and nodes (2, 4, 6, 8, 10). In all cases random seeds were set and recorded for reproducibility. Training parameters giving the smallest mean squared errors were retained. Error-analysis was then employed by performing element-wise subtractions of the predicted drug fold resistance scores from the actual values to spot sequences with which the models have trouble fitting. In an attempt to mitigate these problems, the best model architecture was retained for each drug, the initial random seed was removed and models were generated a few times to spot samples with recurring high absolute error. These were removed only when occurring multiple times before putting the seed back. Final models were evaluated by both regression and classification metrics in order to perform comparisons against those produced by Shen and co-workers, SHIVA and Stanford HIVdb.

2.3.2 Further filtering and ANN architecture optimisations

Filtering criteria described in the Methods section yielded a different number of unique sequence IDs for each drug, ranging from 169 (RPV) to 1524 (NFV) as shown in Table 2.4. The number of allowed sequence variants is generally found to be a maximum of 1000 except for 2 PIs (LPV, NFV) and the NNRTI NVP. Irrespective of cut-off value RTIs displayed a higher average number of variations than PIs, as observed from the higher proportions of unique IDs from expanded sequences. No particular pattern was observed for rare variant filtering, being applied to 10 out of the 18 prepared drug datasets. PCA and error analysis identified 1-2 potential outlying sequences for the drugs ATV, DRV, IDV, TPV and ETR. For the three reverse transcriptase inhibitors ABC, AZT and RPV, lower mean R^2 values were initially obtained with relatively high variances. Therefore in these cases ANN architectures were screened afresh with only one layer of 5-20 nodes to rectify over-fitting, resulting in 14, 19 and 16 nodes respectively. For the remaining 15 drugs, only 2 models performed better with hidden 2 layers, the rest being optimal with 3 layers.

2.3.3 Regression performance

Our ANN models (henceforth referred as protocol A) generally improved on the performance obtained from models developed by Shen and co-workers [127] with higher mean R^2 values and smaller standard deviations, as shown in Figure 2.3. In the case of PIs (Figure 2.3a), the average performance is higher in all cases, the gap being less in the case of IDV and LPV due to the relatively high model variance. Largest improvements for the PIs were achieved for the drugs FPV, SQV and TPV with respective average R^2 differences of 0.117, 0.116 and 0.219 when compared to protocol B. For NRTIs, we obtained better overall performance, except in the case of 3TC, where the model was on par with protocol B. Appreciable improvements were obtained for AZT, DDI and TDF both in terms of R^2 average and variance. In the case of NNRTIs, observably higher performances were obtained with the drugs ETR and RPV. Our seemingly higher performance for RPV may be offset by the smaller number of sample points, which limits generalizability to more divergent samples. Protocol C outperforms ours in the case of NVP, while the model developed for EFV had a lower average R^2 compared to those of protocols B and C.

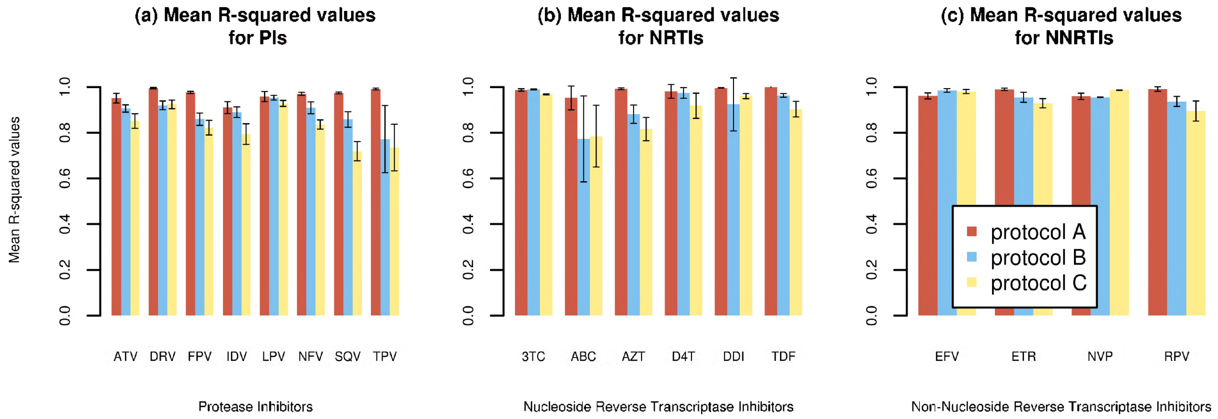


Figure 2.3: Regression training performances for PIs, NRTIs and NNRTIs. Our ANN models are in red while those from Shen and co-workers coloured in blue and yellow, representing the RF and KNN models respectively. Figure re-used from [99]

Further to our k-fold CV calculations, a breakdown of coefficient-of-determination values for the final models, given in Table 2.3 shows that over-fitting is minimal as the R^2 values are equally high for both data subsets used in model training and the test sets, for which data points were not seen during ANN construction. It is important to point out that these quality calculations are separate from the average values evaluated from k-fold CV.

2.3.4 Classification performance

In order to compare our models to additional prediction methods, our fold score predictions were converted to classes by applying cut-off values. All models were found to compare favourably in terms of misclassification rates against Stanford HIVdb, with respect to the initial PhenoSense labels (See Table 2.5). Same cannot be generally said for SHIVA, with exception of RPV where only 8.33% misclassification was obtained. TPV and ETR are unavailable from the SHIVA web server, but in all cases SHIVA seems to be very divergent from even the reference PhenoSense dataset given that the contingency tables were built in each case with respect to this reference. Main differences between our models and those from HIVdb should be attributed to the difference in filtering methods used on the original subtype-labelled fold resistance raw datasets.

Table 2.3: Performance values (R^2) of our final models for different data subsets. Table re-used from [99].

ARV Classes	**ARVs**	**Whole dataset**	**Validation subset**	**Test subset**
PIs	ATV	0.951	0.913	0.856
	DRV	0.991	0.991	0.989
	FPV	0.980	0.938	0.958
	IDV	0.899	0.816	0.842
	LPV	0.966	0.922	0.883
	NFV	0.975	0.924	0.939
	SQV	0.977	0.949	0.906
	TPV	0.989	0.995	0.943
NRTIs	3TC	0.995	0.988	0.985
	ABC	0.984	0.956	0.954
	AZT	0.994	0.979	0.985
	D4T	0.995	0.996	0.979
	DDI	0.997	0.997	0.992
	TDF	0.999	1.000	0.992
NNRTIs	EFV	0.976	0.905	0.967
	ETR	0.996	0.993	0.982
	NVP	0.962	0.939	0.927
	RPV	0.982	0.956	0.915

Table 2.4: ANN hidden layer architectures and filtering parameters. Table re-used from [99].

ARV class	ARVs	Topology	Fraction of unique sequence IDs	No. of allowed combi-nations	Rare variant filtering	No. of outliers removed
PIs	ATV	10x8x6	995/13625	<1000	yes	1
	DRV	8x8	590/10374	<1000	yes	2
	FPV	8x8x8	1429/17501	<1000	no	0
	IDV	8x6x10	1459/16977	<1000	yes	1
	LPV	10x8x10	1284/11019	<300	no	0
	NFV	10x10x10	1524/11929	<300	no	0
	SQV	10x10x8	1484/11509	<300	no	0
	TPV	10x6x8	698/11989	<1000	yes	2
NRTIs	3TC	10x10x6	1342/33181	<1000	yes	0
	ABC	14	1401/34016	<1000	no	0
	AZT	19	1358/33818	<1000	yes	0
	D4T	10x4x4	1365/34056	<1000	yes	0
	DDI	10x6x6	1368/34062	<1000	yes	0
	TDF	10x2	1130/29637	<1000	no	0
NNRTIs	EFV	10x6x10	1400/33906	<1000	yes	0
	ETR	8x2x10	448/11397	<1000	no	2
	NVP	10x10x4	1414/20348	<300	no	0
	RPV	16	169/2977	<1000	yes	0

Table 2.5: Misclassification rates for our ANN models, HIVdb and SHIVA. Table re-used from [99].

ARV class	**ARVs**	**ANN**	**HIVdb**	**SHIVA**
PIs	ATV	26.61	28.57	84.53
	DRV	2.98	22.57	32.41-53.49
	FPV	16.08	36.97	67.0-79.74
	IDV	34.29	26.19	81.92
	LPV	9.79	36.82	68.05-83.51
	NFV	25.23	20.36	80.84
	SQV	30.37	38.75	67.25-88.16
	TPV	9.07	39.88	NA
NRTIs	3TC	3.87	12.09	90.21
	ABC	6.53	33.78	50.76-72.25
	AZT	36.19	29.88	90.38
	D4T	7.31	44.07	79.15
	DDI	8.05	57.52	34.14-92.44
	TDF	5.39	37.2	37.36-66.53
NNRTIs	EFV	16.08	21.05	81.32
	ETR	6.58	13.21	NA
	NVP	24.87	9.4	73.97
	RPV	1.55	24.99	8.33

2.4 Conclusions

The main objective of this work was to improve the accuracy of fold drug resistance ratios to be as close to lab-determined values. We do so by training and optimising regression ANNs for each of the 16 FDA-approved ARVs (comprising PIs, NRTIs and NNRTIs) for which labelled data was obtained from unfiltered datasets available from the HIVdb. Our models compare favourably against the well-established Stanford HIVdb server and the models developed by Shen and co-workers [127], as shown by high classification accuracies and R^2 values. The main novelty here is in the series of filtering criteria used, which focused on subtype B while minimizing sources of technical error. Overall, we found that non-B data is insufficient from publicly-available datasets. The big question which remains is how these models will perform against phenotypic data from non-B subtype if such is made available. Should such labelled data be obtained, the strategy described in this chapter should apply similarly to produce competitive resistance prediction models.

Chapter 3

When binding energies fail: An arsenal of approaches to search for a hidden drug-resistance signal

3.1 Introduction

The approach described in Chapter 2 displayed a surprisingly high overall performance for drug resistance prediction in HIV protease and reverse transcriptase subtype B. This improvement comes with a decreased generalizability for other subtypes. We therefore attempt to fill in this gap by firstly evaluating protein ligand binding energies, still using the same resistance data. To test the correctness of this method we use subtype B to search for a highly-conserved pattern separating drug-resistance from susceptibility, in the hopes that it would be strong enough to extrapolate to non-B subtypes. However, as we will see binding energy fails to reveal any meaningful association with the level of drug resistance. We then test an array of different approaches to search for a clear differential signal, that if sufficiently conserved, might apply to non-B subtypes. Current ARV drug resistance prediction methods are mainly developed using subtype B, hence may tend to show reduced performance in non-B subtypes. A growing body of evidence suggests the influence of subtype on drug resistance [166, 167]. For instance, there is an increased rate of the early emergence of an RT major DRM K65R variant in subtype C when tenofovir is used in therapy [168]. Also, the RT N348I variant (not found amongst the 2017 DRM update [80]) was found to be frequent in patients failing first-line ART for subtype C HIV [169]. In a study assessing the suitability of resistance prediction algorithms in subtype C, it was found that the resistance was overestimated against the drugs etravirine and rilpivirine [170]. Another study based on viral load and CD4 cell counts found that non-B subtypes (A, C and D) differed amongst themselves in the selection of DRMs for patients not undergoing routine viral load checks and failing first-line NNRTI regimens [171]. Unfortunately the number of non-B resistance-labelled sequences from the Stanford HIVdb dataset is currently insufficient and so is its diversity for improving predictive models against any non-B HIV. This said, *in silico* techniques for structural analysis may be evaluated for their effectiveness in determining drug affinity to receptors, which may then be extrapolated to non-B subtypes if they prove to be correct enough. There are various ways of estimating drug binding free energies ranging from the computationally-cheaper virtual screening, to those that can improve on it further such as MM-PBSA/MM-GBSA [172], thermodynamic integrations [173], linear interaction energy [174], alchemical transformations [175, 176], potential of mean force calculations [177], which are however very computationally-expensive to compute even for moderately large sample sizes. A great amount of effort is focused on improving *in silico* methods used to evaluate drug performance as this is still an open research problem [178]. Many approaches have been tried and are still emerging, with some examples including: the combination of docking and molecular dynamics to predict PI drug resistance, with reported accuracies in the range of 72-83% [179]; the use of docking scores to predict SQV resistance in HIV [180]; the combination of 3D-QSAR and molecular docking to cross-validate the potency of DRV derivatives as potential protease inhibitor candidates [181]; the training of convolution neural networks using 3D representations of protein-ligand complexes to predict drug performance [182];the development of a trained Random Forest-based scoring function used to augment AutoDock Vina's semi-empirical score for a improved overall screening power [183]. The list is long, especially with the increasing accessibility of structural data and advent of more sophisticated machine learning algorithms.

In this work, we start by investigating simpler approaches using a relatively large number of labelled sequences in an attempt to mitigate the drop in accuracy for a computationally-tractable speed trade-off, to gradually escalate towards more involved approaches. Starting with (1) modelling and minimizing with ligands, we then proceed to (2) protein-ligand docking and (3) binding energy calculations during short molecular dynamics simulations of the complexes. Subsequently, several approaches are tested, involving Normal Mode Analysis (NMA), Dynamic Cross Correlation (DCC), Perturbation Response Scanning (PRS) and dynamic residue network analysis. No strongly-conserved differential signal could be picked up to identify the resistance but they helped in laying the foundations and understanding the system in order to get closer towards a working method, later described in Chapter 4. In each of the following subsections we give details of the methods used in this chapter.

3.1.1 Overview of homology modelling and energy-minimization

In our search for a ligand-binding performance metric from docking, protein receptors are required. However, not all of them are available from current macromolecular structure databases, such as the Protein Data Bank. In the absence of experimental data from X-Ray crystallography or NMR data, protein structures can be inferred via various *in silico* approaches such as *ab initio* methods for relatively small proteins (< 120 residues) [184] and template-based modelling [185]. Provided with solved high-resolution templates of sufficient similarity, high-quality models can be obtained for target sequences [186]. MODELLER is a very versatile library with which a most-probable protein 3D models can be inferred from experimentally-determined structures [185]. Given sufficiently similar sequences comparative modelling can predict structures with a root mean squared deviation within 1 Å of the actual structure [187]. The method begins by aligning a sequence of interest (the target) to that of a selected template [188], which already has a solved structure. Multiple spatial features including $C_\alpha - C_\alpha$ bond distances, angles, dihedrals and dihedral pairs at atom intervals (2, 4 and 8 atoms), atomic density and solvent accessibility are then extracted from the template to be transferred onto an initial candidate structure for the target [189]. A final model is obtained by a combination of restraint satisfaction and conjugate gradient energy minimization (EM), even resorting to molecular dynamics or simulated annealing depending on the degree of refinement required [190, 191]. For our purpose, we require the positioning of a drug inside the binding cavity. To achieve such a task, MODELLER copies the co-crystallized ligand to the target model as a rigid body, which implies that additional resolution of energies may be required to allow for more favourable ligand orientations. For this task energy minimization is explored using the steepest descent algorithm via the GROMACS tool. While the global potential energy minimum would be ideal, finding it through high-dimensional configurational search space is very laborious such that EM algorithms within GROMACS only stop at the nearest local minimum [192]. Available algorithms include that of the steepest descent (SD), the conjugate gradient and the limited-memory Broyden-Fletcher-Goldfarb-Shanno (l-BFGS) minimizer, which all vary in their convergence rates versus accuracy trade-offs [192]. We employ the SD algorithm, implemented in GROMACS as such:

$$\mathbf{r}_{n+1} = \mathbf{r}_n + \frac{\mathbf{F}_n}{\max\left(|\mathbf{F}_n|\right)} h_n \tag{3.1}$$

where $\mathbf{r}$ and $\mathbf{F}$ are vectors each of length 3N describing the atomic positions and forces along the x, y and z axes for each atom of a given system respectively. The scalar h_n is the initial step size in nm, which is scaled by a factor of 1.2 or 0.2 based on whether the difference next potential energy is lower or higher than its actual value, respectively [192]. The force vector is scaled by the absolute maximum of its components before applying the position update for all the atoms at once. The algorithm stops when either a user-defined gradient (ϵ) of the potential energy or a maximum number of steps is reached.

3.1.2 Molecular docking for estimating ligand affinity

Small molecule docking against receptors forms part of the preliminary steps for prioritizing lead molecules prior to the use of more computationally-demanding steps in the hopes of expediting the drug discovery process [193, 172, 194]. Diverse algorithms have been developed for that particular task and mainly involve finding the most energetically-favourable ligand binding pose using scoring and searching functions [195]. As reviewed by Kitchen and co-workers, ligand scoring can broadly be classified into force-field, empirical and knowledge-based methods [196]. An example of a force-field-based approach is the one used by AutoDock4 and is shown in equations 3.2 and 3.3, in which energy terms are calculated independently and summed. However, it is also semi-empirical [197]. The estimated binding free energy (ΔG) is a summation of the inter- and intra-molecular interactions in the bound and unbound forms, together with ligand conformational entropy (equation 3.2).

$$\Delta G = (V^{L-L}_{bound} - V^{L-L}_{unbound}) + (V^{P-P}_{bound} - V^{P-P}_{unbound}) + (V^{P-L}_{bound} - V^{P-L}_{unbound} + \Delta S_{conf}) \tag{3.2}$$

In equation 3.2, superscripts L and P refer to the ligand and protein respectively. The intra- and inter-molecular potential energy (V) terms calculated for each of the bound and unbound terms in the same equation correspond to the following:

$$V = W_{vdw}\sum_{ij}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond}\sum_{ij}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + \\ W_{elec}\sum_{ij}\left(\frac{q_i q_j}{\epsilon(r_{ij})r_{ij}}\right) + W_{sol}\sum_{ij}(S_i V_j + S_j V_i)e^{-r_{ij}^2/2\sigma^2} + W_{conf}N_{tors} \tag{3.3}$$

where the coefficients W are weighing terms calibrated to experimental binding constants [198]. The first and second terms are implementations of the Lennard-Jones potential for estimating van der Waals and directional hydrogen bonding contributions respectively, where the higher and lower power terms jointly denote repulsive and attractive components. The Coulombic term handles electrostatic interactions as a function of charge pairs (q_i and q_j) and their inter-atomic distance. Desolvation accounts for the hydration-related energy contributions in the form of implicit water [199]. Finally the entropic term gives an estimation of conformational entropy lost upon ligand binding [197].

An example of a non force-field-based (semi-empirical) scoring function is the one used by AutoDock Vina (henceforth referred to as Vina), whereby binding affinity is estimated via a series of empirically-determined weights (from actual ligand binding energies) applied to two Gaussian terms and three piecewise functions, which are mainly parametrised by inter-atomic distances [200, 201]. No interactions are evaluated beyond a distance of 8 Å. The Gaussian terms together with the repulsive term form the steric component of the potential [201]. Hydrogen bonding and hydrophobic interactions (equations 3.8 and 3.9) are applied conditionally via piecewise functions - i.e. a hydrogen bond is evaluated only between proton-donating and accepting atom types, while hydrophobic interactions are only inferred from pairs of hydrophobic atom types. The atomic distance is defined by the difference between their collective van der Waal radii (R_t) for the atom types and their centers, as shown in equation 3.5. The total energy is evaluated as the sum of inter- and intra-molecular contributions (c_{inter} and c_{intra} respectively) for the complex, as shown in equation 3.4, where h refers to each energy term.

$$c = c_{inter} + c_{intra} = \sum_{ij} h_{t_i t_j}(d_{ij}) \tag{3.4}$$

$$d_{ij} = r_{ij} - R_{t_i} - R_{t_j} \tag{3.5}$$

$$gauss_1(d_{ij}) = e^{-(d_{ij}/0.5\text{Å})^2} \tag{3.6}$$

$$gauss_2(d_{ij}) = e^{-((d_{ij}-3\text{Å})/2\text{Å})^2} \tag{3.7}$$

$$repulsion(d_{ij}) = \begin{cases} d_{ij}^2, & if d_{ij} < 0 \\ 0, & if d_{ij} \geq 0 \end{cases} \tag{3.8}$$

$$hydrophobic(d_{ij}) = \begin{cases} 1, & if\ d_{ij} < 0.5\text{Å} \\ linearly\ interpolated, & if\ 0.5\text{Å} < d_{ij} \leq 1.5\text{Å} \\ 0, & if\ d_{ij} > 1.5\text{Å} \end{cases} \tag{3.9}$$

$$hydrogen(d_{ij}) = \begin{cases} 1, & if\ d_{ij} < -0.7\text{Å} \\ linearly\ interpolated, & if\ \ -0.7\text{Å} < d_{ij} \leq 0\text{Å} \\ 0, & if\ d_{ij} > 0\text{Å} \end{cases} \tag{3.10}$$

In order to calculate the poses, conformational space (comprising of translational and rotational degrees of freedom) has to be visited, but for computational tractability various algorithms are used to avoid an exhaustive sampling [196]. Vina estimates ligand binding free energy via the weighted sum of terms comprising two Gaussian functions and terms for repulsion, hydrophobicity, hydrogen bonding and the number of rotatable bonds evaluated for inter- and intra-molecular interactions [200]. As such, in addition to the commonly used molecular mechanical approximations of chemical potentials, Vina is also partly ML-oriented. The best solutions for the ligand poses are then ranked from a search space initiated from a diverse set of randomly-mutated conformations by minimizing the scoring function using the Broyden-Fletcher-Goldfarb-Shanno algorithm [200]. Vina is generally more easily set-up and according to Trott and co-workers [200] is approximately an order of magnitude faster than its parent software AutoDock, yielding lower ligand RMSDs from the refined PDBind training set while still maintaining higher out-of-sample performance. For these reasons Vina was a tool of choice for assessing in this case not the ligand, but the indirect performance of its bound receptor in a large-scale experiment. We hypothesised a reduction in ligand affinity against drug-resistant receptors in comparison to drug-susceptible ones. Correlation tests of the binding energies against the actual fold drug resistance scores for the B subtype dataset would subsequently inform us on its suitability as a proxy for inferring the efficacy of each PI drug. Both the Pearson's and Spearman's correlations are used. While Pearson's coefficient is the covariance normalized by the product of the standard deviations of bivariate samples, Spearman uses the same mechanics on the ranks rather than the actual values. With a range of [-1, 1] jointly corresponding to negative and positive association (and zero being a lack thereof), Pearson's r is more affected by the magnitude of differences, while Spearman's ρ is affected to a much lesser extent. The tool X-Score, also used in this experiment, estimates binding affinity using a different scoring method from an already docked ligand pose, based on a consensus averaged score estimated from a set of 3 empirical scores pertaining to different aspects of hydrophobicity, namely HPScore, HMScore and the HSScore [202]. Each of the individual scores factors in van der Waals interactions, hydrogen bonding and deformation with different weights, differing at their hydrophobic term which separately account for "Hydrophobic Pairs", "Hydrophobic Match" and "Hydrophobic surface" [202].

3.1.3 Molecular dynamics simulations for conformational sampling

Molecular dynamics simulations are a way of investigating the dynamic properties of molecular systems over time. Due to the advent of increasingly efficient algorithms and cheaper cost of computations, MD simulations are being heavily used in research to investigate multiple biological phenomena, such as allostery [203], in drug discovery [176], protein dynamics in alternate phenotypes [204], which themselves encompasses a vast array of research applications. Different levels of simulation accuracies and theory can be applied to the biological systems under investigation via combinations of quantum and/or Newtonian representations of atomic models, for this work we only use the latter for computational efficiency given the large experimental size. Newtonian (molecular mechanical) models can be parametrised to mimic certain aspects of biology, such as ionic concentrations, temperature, pressure and to a limited extent the pH as well (via the recognition and use of specific protonation atom types in the GROMACS software for instance). The basis for atomic mobility is implemented by solving for Newton's second law of motion ($F = m\ddot{x}$) in a thermodynamic system defined by functions approximating potential energies. The simulations begin with a static structure to which velocities components (v_i) are randomly selected from a Maxwell-Boltzmann distribution generated by multiplying a random normally-distributed number in the range [0,1] by the standard deviation $\sqrt{\frac{kT}{m_i}}$ of the Boltzmann curve at the required temperature in Kelvin, where k is the Boltzmann constant. Velocities for all particles have to then be rescaled to match the required temperature and total energy [192] for the ensemble, due to the partial stochasticity. Residue and atom types for the force field define various atomic parameters such as the masses, partial charges, valences and bonding geometries for use in the thermodynamic system. After the first velocity updates from zero, the subsequent velocities and positions are calculated for each time step, for instance by applying integrators such as the leap-frog or Verlet. Energies are calculated using the potential function defined by the chosen force field equations. These functions cater for both bonded (bond stretching, bending, angles and dihedrals) and non-bonded (Lennard-Jones and Coulomb) interactions, which are very similar to the terms used for pose-scoring in molecular docking. In this thesis, the AMBER03 [205] forcefield (equation 3.11) is used for all MD simulations. The bond stretching (1-2 interactions) and bending (1-3 interactions) terms [206] are defined by their respective force constants k_b and k_θ and their deviations from equilibrium position (b_{eq} and θ_{eq}) along their respective harmonic potentials. V_n is the force constant while ϕ and γ are the dihedral and phase angles respectively for the third term (1-4 interactions). The last term (for the non-bonded interactions) is a combination of the Lennard-Jones and Coulomb potentials, where the A and B terms represent atomic repulsion and attraction, while partial charges are represented by q. The effect of simulation medium is controlled by the parameter ϵ, which is usually has the value of one in typically-used solvated environments [205].

$$\begin{aligned} E_{total} = \sum_{bonds} k_b(b - b_{eq})^2 + \sum_{angle} k_\theta(\theta - \theta_{eq})^2 + \\ \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \end{aligned} \tag{3.11}$$

Independently to the force field equations, restraint potentials (mainly distance, angle and dihedral) are also available to prevent large sudden deviations, for instance during preliminary temperature and pressure equilibration steps, prior to doing production MD runs. In order to make biomolecular simulations more life-like, these MD experiments are typically performed in explicitly-defined water molecules, which are mainly utilised as variations of the SPC (Single Point Charge) or TIP water models together with added ions. As the solvation introduces new forces to the structure under study, these atomic clashes are first resolved by employing energy-minimization algorithms (e.g. steepest descent and conjugate gradient) before initializing velocity components for each atom. The temperature and pressure equilibration algorithms (commonly those of Berendsen [207] and Parrinello-Rahman [208] respectively) are typically done prior to production runs to ensure appropriate thermodynamic properties for the target temperature and pressure of the system. Finally constraint functions (typically LINCS [209] or SHAKE [210] algorithms) are also included to correct for exceedingly large geometry deviations which may occur during equilibration and production runs [192].

3.1.4 Normal Mode analysis using an Elastic Network Model

Modal analysis is used in different domains of research ranging from the analysis of vibrational motions within objects engineered to perform critical functions, such as bridges and plane wings [211] to the study of functional motion in enzymes, viral and large protein assemblies [212]. The normal modes are obtained by decomposing a matrix of second partial derivatives of the potential energy for a given system, with respect to displacement along the x, y and z component axes [213]. Depending on the system size and/or level of refinement needed, this potential energy can be the result of applying a wave function (e.g. hybrid quantum mechanics [214]), a molecular mechanical force-field [215, 192] or more simply via a harmonic potential [216]. This second order derivative is generally represented as a square matrix, termed a Hessian (H), where V is the potential energy, subscripts i and j denote atom positions while the Cartesian coordinates are X, Y and Z:

$$H_{ij} = \begin{bmatrix} \frac{\partial V}{\partial X_i \partial X_j} & \frac{\partial V}{\partial X_i \partial Y_j} & \frac{\partial V}{\partial X_i \partial Z_j} \\ \frac{\partial V}{\partial Y_i \partial X_j} & \frac{\partial V}{\partial Y_i \partial Y_j} & \frac{\partial V}{\partial Y_i \partial Z_j} \\ \frac{\partial V}{\partial Z_i \partial X_j} & \frac{\partial V}{\partial Z_i \partial Y_j} & \frac{\partial V}{\partial Z_i \partial Z_j} \end{bmatrix} \tag{3.12}$$

which represents force coefficients $\frac{\partial F_x}{\partial x}$, where x here is a generalized coordinate for the displacement of any particle along an axis. In order to obtain the modes of motion, which are non-zero orthogonal basis vectors, the Hessian is decomposed. These characteristic vectors, termed eigenvectors are typically normalised to unit length and have corresponding eigenvalues (scalars) associated to them. Two popular matrix decomposition algorithms are the eigenvalue and singular value decomposition methods (EVD and SVD, respectively), which both yield the same eigenvectors for a square symmetric matrix. EVD factorisation by $U\Lambda U^T$ yields a diagonal matrix Λ of eigenvalues and another array for the matching eigenvectors U, while SVD factorisation provides the same information as identical left and right singular vectors, as represented by the following equation $U\Sigma V^T$. Elements from the diagonal matrix Σ are however equal to the square root of

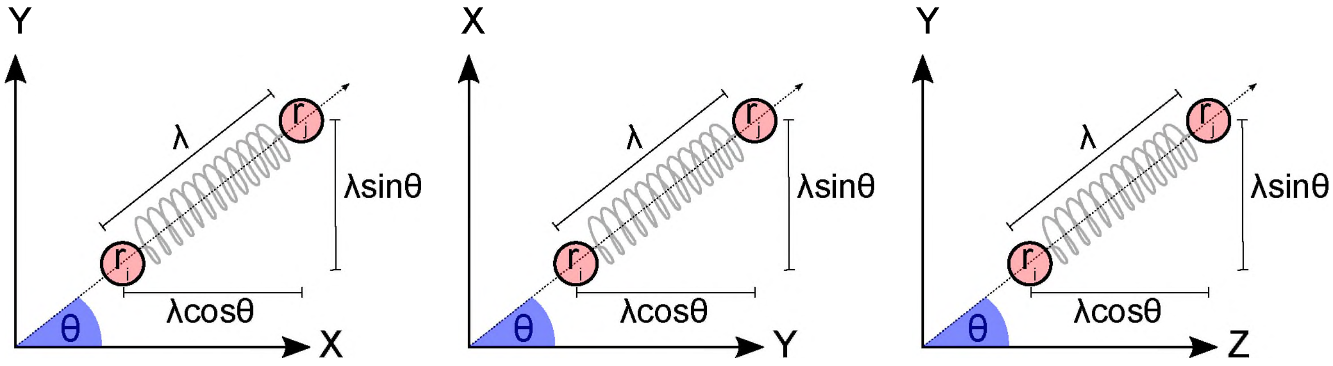


Figure 3.1: Components of the spring constant λ along 3 pairs of axes from 3D Cartesian space for residues pair r_i and r_j, connected by a spring. The constant λ is usually set to one.

values from the EVD diagonal matrix [217]. Coincidentally, in principal components analysis, a covariance matrix (C) is decomposed [218] instead of the Hessian, nevertheless the matrix C can also be obtained by inverting the Hessian [219]. The use of NMA relies on a main assumption, which is that the equilibrium conformations oscillate around a single well-defined minimum energy conformation [220, 219]. In other words, NMA results are only valid at the immediate neighbourhood of this potential energy minimum, as large displacements from it may visit new minima [219], for which the modelled Hessian will differ. The extraction of orthonormal modes only hint at all the mathematically-possible ways a protein can move around a defined equilibrium and do not directly show how the actual motion happens, even though one or more of the low-frequency modes are functionally-relevant in most cases [221]. These modes correspond to the global motions, which are insensitive to local interactions [219]. This fact was exploited by Tirion's pioneering work in which the usual multi-term molecular mechanical potentials were replaced by the simpler Hookean potential to successfully reproduce the slow dynamics of multiple large globular proteins [216]. Following this work, Bahar and colleagues further propose the concept of elastic network models (ENM) by introducing a cut-off distance ($\leq 7\mathring{A}$) for the calculation of a Kirchhoff (or valency-adjacency [222]) matrix Γ from proteins coarse-grained by C_α atoms to be used by two main models, namely the Gaussian Network Model (GNM) [223] and the Anisotropic Network Model (ANM) [224]. In both these models, C_α atoms serve as nodes, which are connected by springs of uniform stiffness γ, fixed by a cut-off radius r_c [213]. The Γ matrix for GNM is of size $N \times N$ [223] and thus loses mode directionality, while that of the ANM is of size $3N \times 3N$ [224] which caters for displacement along each axis. While the uniform spring constant λ can be varied, the Kirchhoff matrix is defined as such:

$$\Gamma_{ij} = \begin{cases} -1, & i{\neq}j,\ R_{ij}{\leq}r_c \\ 0, & i{\neq}j,\ R_{ij} > r_c \\ -\sum_{i,i\neq j} r_{ij} & i = j \end{cases} \tag{3.13}$$

In order to determine the contribution for each force coefficients along each component axis, the second order partial derivative λ is scaled by the angle cosine formed between any pair of axes at each residue pair, for the off-diagonal residue entries from the Hessian. Figure 3.1 shows components of the spring constant along some example axes. The cosine is calculated from the dot product of the unit vectors. Each diagonal value is however evaluated by the row sum from the matrix, as shown in equation 3.13.

3.1.5 Dynamic Cross Correlation (DCC)

Dynamic cross-correlation is a popular method for MD analysis used in the study of correlative motions between atoms of molecular systems [225, 226]. One of its implementations is defined by the following equation, which we use in this section:

$$C_{ij} = \frac{\langle \Delta r_i . \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle \langle \Delta r_j^2 \rangle}} \tag{3.14}$$

where Δr_i or $r_i - \langle r_i \rangle$ is the displacement of the i^{th} residue from its time-averaged position [227]. It also corresponds to the normalized covariance matrix, also known as the second moment matrix $\mathbf{A}$, where a_{ij} is defined by $\langle \Delta r_i . \Delta r_j \rangle$, which is conceptually the mean squared fluctuation when i equals j [228]. DCC values are generally plotted as a 2D map, which displays the correlations between every residue pair as determined over a defined amount of time, whereby values in the range [-1,1] show negative and positive correlations respectively. A value of zero corresponds to uncorrelated motion. The method cannot detect atomic motions that are in the same phase and period but at perpendicular angles; a major caveat found at the sub-nanosecond time scales is that perceived correlations heavily depend on the time-scale over which conformations are sampled. Very early work by Hunenberger and co-workers [229] demonstrate that cross-correlation values may not converge when evaluated over small sections (picoseconds) from an MD trajectory. Nevertheless the method is useful and can lead to interesting insights, such as the identification of collective motions and the characterization of protein domains [227] and has been used in several applications investigating differential dynamics happening in protein systems, such as those of (1) wild type and mutant coronavirus proteins upon RNA binding [230], (2) allosteric mechanisms in human Ras protein onco-mutations [231], (3) binding of the cyclic peptide DC3 to the androgen receptor involved in prostate cancer [232] and many more. Here DCC is used to investigate whether correlations prevailing in drug-resistant and drug-susceptible ensembles are differential. If successful, the method could be used to further focus on early resistance-associated movements that are most differential, specific enough to function in non-subtype B sequences. Instead of using a DCC map, aggregates of the cross correlation values are compared across ensembles, as explained in the methods section.

3.1.6 Application of Perturbation Response Scanning to search for trigger residue positions correlated with the resistance state

Originally described in 2009 by Atilgan and Atilgan in the context of Fe^{3+} shuttling within a ferric-binding protein [233], the principle has been slowly gaining popularity [234, 235, 236, 237, 238, 239] in the analysis of allostery in proteins. The principle relies on the application of uniformly-distributed random forces at a single residue C_α atoms, centered around zero, over their three coordinate axes and recording the resultant displacement of all other residues. The force applied, or the perturbation, is isotropic, meaning that its uniformly distributed spherically [233]. There are three main assumptions for setting up the main equation behind the PRS algorithm. Firstly, for a net force of zero at equilibrium, the sum of forces resulting from internal (residue-residue interactions) and externally-applied perturbations should be zero along every direction cosine for each residue.

$$B_{3N \times M} \times \Delta f_{M \times 1} = \Delta F_{3N \times 1} \tag{3.15}$$

where B if the matrix of direction cosines, Δf is the vector of internal forces and ΔF is the matrix of external perturbations, for a system of C_α atoms connected by springs. Secondly, the displacements along each of the bond vectors (Δr) are matched to those of the position vectors (ΔR), given the direction cosines.

$$B^T_{M \times 3N} \times \Delta R_{3N \times 1} = \Delta r_{M \times 1} \tag{3.16}$$

Lastly, Hooke's law is formulated as such, using uniform spring constants, which are defined only for residues within a specified cut-off radius r_c:

$$K_{M \times M} \times \Delta r_{M \times 1} = \Delta f_{M \times 1} \tag{3.17}$$

By replacing the internal force Δf and subsequently the bond vectors Δr in equation 3.15, a relation involving the Hessian matrix $(BKB)^T$ is retrieved, in an expression representing the external force as a linear product of the position vector and the Hessian:

$$(BKB)^T \Delta R = \Delta F \tag{3.18}$$

By changing the subject of formula to determine the displacement ΔR from ΔF, the main perturbation equation for the whole system is obtained by inverting the force constants of the Hessian into a matrix of displacements per unit force, along the direction cosines:

$$(BKB^T)^{-1} \Delta F = \Delta R \tag{3.19}$$

Equation 3.19 is applied system-wide, using a force vector, comprising of zeros except for the residue to be perturbed, i.e. $\{0, 0, 0 ... \Delta F^i_x, \Delta F^i_y, \Delta F^i_z ... 0, 0, 0\}$ where the force is uniformly applied to residue i along each of its units of displacement. Perturbation components propagate through the entire molecule to result in a global displacement from an initial molecular configuration, which is recorded by subtracting the displacement. The perturbation is repeated for the same residue (for example 100 times), and average displacements are obtained before calculating the overlap (or correlation) against a target conformational change. The same experiment is repeated for each residue to determine the regions yielding highest similarities to the target conformation. The Hessian matrix can also be obtained from a trajectory, as explained for the Normal Mode Analysis in subsection 3.1.4, and is the one used in this analysis.

3.1.7 Residue network analysis for identifying differential network behaviours associated with resistance

A network graph is a mathematical representation of variables with underlying relational information. Being composed of nodes (also known as actors or vertices depending on the field of application) and edges (also known as arcs or ties) which link the former, they can represent very complex data and lend themselves to different types of analyses. There are two main ways of representing networks, firstly as edge lists, which are more compact and secondly as the more computationally-friendly adjacency matrices, which allow the use of linear algebra operations. Edges can be binary with values $\{0, 1\}$ or continuous, in which case they are weighted. Depending on types of relations supported by the edges, networks are classified as undirected or directed. In the former case, the adjacency matrix is symmetrical while in the latter case it is asymmetric (also termed a digraph), reflecting the non-reciprocity of node relationships [240]. In addition, networks can be cyclic or acyclic, in which case nodes are allowed to have self-connections or not, respectively.

Several network concepts are defined to summarize and reveal underlying engrained patterns from relational data. These include, but are not limited to the degree, betweenness, closeness, shortest paths and clustering coefficient. These and additional metrics are described herein, with social network interpretations (mainly adapted from [241]). The **degree** centrality, also known as connectivity is defined by the following:

$$k_i = \sum_{i \neq j} A_{ij}, \tag{3.20}$$

where A is the adjacency matrix. Node degree defines the sum of neighbours for any given node. It can be calculated as the row or column sum from the adjacency matrix. Scaling by the maximum degree adjusts its values to the range [0,1], and can be defined as:

$$K_i = \frac{k_i}{k_{max}}. \tag{3.21}$$

Heterogeneity is a measure of degree variability, normalised by its average. As opposed to regular graphs, where every node has the same degree, complex networks tend to be approximately scale-free and tend to display a higher degree heterogeneity [241]. A scale-free topology is a network property in which a straight line displays a high fit if the log degree centrality is plotted against its log probability density [241]. Such networks generally comprise a few highly-connected nodes and a large number of low-degree nodes, making such networks resilient to accidental errors [242]. The heterogeneity metric is defined by the following:

$$\frac{\sqrt{var(k)}}{mean(k)} = \sqrt{\frac{n \times sum(k^2)}{sum(k)^2} - 1}. \tag{3.22}$$

The **maximum adjacency ratio** (MAR) for a given hub node shows whether it connects strongly to a few neighbouring nodes or moderately to many more nodes, and is defined as:

$$MAR(i) = \frac{\sum_{i \neq j} (A_{ij})^2}{\sum_{i \neq j} A_{ij}}. \tag{3.23}$$

The **density** metric describes the average degree centrality prevailing within a given network. A value of 1 corresponds to fully-connected nodes, whereas lower values point to more heterogeneous relationships. Network density is defined by the following:

$$Density = \frac{mean(k)}{n-1}. \tag{3.24}$$

The **clustering coefficient** assesses the degree of interconnectedness around a given node, the social interpretation being whether somebody's friends are also friends with each other thus forming a clique. The highest value would thus occur when all the nodes are connected to each other around a given node. It is defined likewise:

$$Clustering\ coefficient(i) = \frac{\sum_j \sum_k A_{ij} A_{jk} A_{ki} - \sum_j A_{ij}^2 A_{ij}}{\left(\sum A_{ij}\right)^2 - \sum_j A_{ij}^2}. \tag{3.25}$$

Topological overlap (TO) measures the extent of shared neighbours between two nodes. This metric adds to clustering coefficient by incorporating shared presence and absence of contacts (or "friendships" in the social network jargon) between two nodes, and can thus be used as a way to decrease the impact of erroneous adjacencies in a network thus yielding a more robust measure of interconnectivity [241]. TO is defined by the following:

$$TO(i,j) = \frac{\sum_{u \neq i,j} A_{iu} A_{uj} + A_{ij}}{\min\left(\sum_{u \neq i} A_{iu}, \sum_{u \neq j} A_{ju}\right) + 1 - A_{ij}}. \tag{3.26}$$

Betweenness centrality (BC) measures the number geodesics going through an intermediate node [243]. This therefore reflects the importance of high-betweenness nodes as information control points within a network [244]. For this reason, their removal will tend to cause the most disruption within the network [245]. BC is defined as:

$$BC(i) = \sum_{j \neq i \neq k \in V} \frac{\sigma_{jk}(i)}{\sigma_{jk}}, \tag{3.27}$$

where $\sigma_{jk}(i)$ denotes the number of geodesics for each of the node pairs j and k that go through node i. The term geodesic distance is generally used to refer to the shortest path connecting two given points along a curved surface [246] and is borrowed as a network terminology to describe the shortest path (or degree of separation [240]) connecting two nodes in a network graph [247]. Some examples of algorithms used for determining the geodesic distances are Dijkstra's [248], Bellman-Ford's [249], the A^* [250] and Johnson's [251] algorithms, among a list of many more [252]. The **average shortest path** (L), also known as farness, unlike betweenness, gives the averaged shortest paths leading to instead of through a node. The computation of L is essentially performed for each node by determining the shortest distance to every other node before obtaining an average for each residue, defined likewise:

$$L(i) = \frac{\sum_{j=1}^{n-1} distance(i, j)}{n-1} \tag{3.28}$$

Closeness centrality is the inverse of L, defined such that a larger value denotes a node that is closer to every other node in the network. In other words, a node of high closeness centrality has the shortest average path length to every other node in a network graph, and the equation for such a property is generally represented likewise:

$$closeness(i) = \frac{n-1}{\sum_{j=1}^{n-1} distance(i, j)} \tag{3.29}$$

There is a wide array of applications for the use of networks in representing and analysing complex biological processes. We mention a few of the exciting applications enabled by the network analysis framework in biology. Protein-protein interaction data and various annotations by the STRING database have been put together to assist in the inference of function for orthologous proteins by association of relational data [253]. The GeneMANIA database similarly integrates various annotations such as co-expression, pathway and literature to assist in predicting gene function from contextual gene lists [254]. They have also been used to investigate gene co-expression modules in a complex diseases such as Type 1 diabetes [255] and in investigating the epigenetics of Parkinson's disease, which is the second most prevalent neurodegenerative disorder worldwide [256, 257]. More generally, network analysis is also used for mapping, tracking and predicting pathology-associated neural patterns involved in the impairment of the connectome, which is a mapping of brain connectivity topology [258]. There is growing interest and great research impetus in the application of network analysis for the study of intra-protein residue behaviour termed by various names (e.g. RIN, DRIN, DRN, RCN) depending on implementation details. In this work we used some network metrics over the course of MD simulations to investigate network behaviour associated with drug-resistance or drug-susceptibility in HIV protease complexes. In each case, networks were built from subsets of sampled time frames before averaging each of the network metrics and comparing them across collections of the resistant and susceptible complexes for each PI drug. This differs from previous methods which mainly infer residue contact based on static structures, such as in [259, 260] or in [261] where energy minimization was used. Our method was developed independently but is similar to that described by Doshi and co-workers, without the use of upper and lower bounds for filtering contacts [262].

3.2 Methods

Dataset preparation: Sequences labelled with drug fold resistance ratios were obtained from previous filtering, as explained in Chapter 2, subsection 2.2.1. The aim here is to use a battery of different approaches to highlight any type of difference occurring at the sequence or structural level able to differentiate the resistance state from the susceptible state for all of the eight FDA-approved PIs. Stanford HIVdb cut-off values defined previously in Table 2.2 were used for each of the protease inhibitors ATV, DRV, FPV, IDV, LPV, NFV, SQV and TPV. In order to extract conserved differences, multiple sequences were used, in this case 100 highest and the 100 lowest fold resistance ratios were selected using the individual cut-off values. Mutations (with respect to the subtype B consensus protease) present in the resistant sequences were subsequently collected.

3.2.1 Homology modelling and energy-minimization

We begin by evaluating the simplest approach involving homology-modelling of protease receptors together with a PI drug and correlating the estimated drug binding energy against actual fold resistance values to see whether these can be used to infer resistance later in non-B subtypes. HIV subtype B protease variants were modelled together with a copy of the co-crystallized ATV ligand from the template structure (PDB ID: 3EL9) using MODELLER (version 9.18). As the software mainly caters for protein backbone and side chain placement from template-derived restraints, minimization of the complex was subsequently performed, using the method of steepest descent (SD) in GROMACS (version 2016.1) [263] in order to improve receptor-ligand contacts. Homology modelling was mainly achieved by preparing a template-target PIR alignment file, which also included the ligand residue from the crystal structure. As a preparation for minimization, the modelled receptor and ligand were extracted for topology generation in each case. Due to the sheer number of minimization experiments, the co-crystallized ligand was directly used to prepare a single "itp" topology file using bond charge correction (BCC) algorithm from the ACPYPE tool [264] and simply copied to each model directory. Fully-protonated "gro" coordinate files were generated for each variant from their extracted modelled ligand file using Open Babel (version 2.3.1). This short cut made the method feasible and was possible due to the fact that MODELLER does not alter ligand conformation during modelling, thus retaining atomic partial charges for all variants. To determine the stopping criterion for energy-minimization, minimum gradient components of the potential energy (namely 5 and 10 kJmol^{-1}nm^{-1}) referred to as ϵ values, were evaluated with an initial step size of 0.01nm using a Python script wrapping the GROMACS commands. An automatic resolution of minimization failures due to bad contacts was done by restarting the minimization with a halved initial step size. This mechanism would try the SD algorithm up to four times, failing which the procedure would be deemed a failure. The closed conformation high resolution (1.6 Å) template crystal structure containing the drug of interest was retrieved from PDB to build high quality models for every resistance-labelled sequence retrieved from Stanford HIVdb. Due to the number of drugs and the possibility of replication, the optimizations were limited to ATV-related sequences for preliminary investigations. The top 10000 variants were selected from the filtered data that were previously used to train and test the ANN model for predicting ATV resistance from protease sequences (in Chapter 2) by ranking them in increasing order of similarity. The last frame from the each trajectory was used to extract the protein and ligand in each case using the *trjconv* command. Open Babel and AutoDockTools were used to generate ligand MOL2 files for X-Score and PDBQT files for Vina to score the poses in-place. Correlations were calculated to determine the trend between actual drug fold resistance and the determined binding energies. Individual energy terms from Vina were also evaluated for correlations.

3.2.2 Molecular docking: Investigating binding energy relationship with drug resistance by improving ligand poses

The ligand conformational search is made more exhaustive by using molecular docking to allow for improved ligand movement and positioning within the modelled receptor variants. Vina is used for flexible ligand docking and X-score is used as an independent scoring tool for the top-scoring pose obtained from Vina. Due to the decreased requirement for computational resources, the experiment is performed for all 8 PI drugs with 10000 models in each case. As there would be no further energy-minimization of the complex after ligand placement in the binding site, models were built with very slow refinement for a more thorough receptor structure optimization via an increased number of MD and simulated annealing steps, from within MODELLER [185]. A random seed (-10000) was fixed for modelling following which control docking was performed using the templates with PDB accessions 3EL9 [265], 2HS1 [266], 3NU3 [267], 2AVO [268], 2O4S [269], 3EL5 [265], 2NMZ [270] and 3SPK [20] in which the inhibitors ATV, DRV, FPV, IDV, LPV, NFV, SQV and TPV had been obtained co-crystallized within the receptor proteases. All templates were pre-processed to only keep higher-occupancy side chain rotamers using an in-house Python script. In case of equal occupancy values, the last rotamer was kept. Based on ligand RMSD values, a co-crystallized flap water was retained only for non-TPV related targets. The algorithm for picking interfacial water selects only those water molecules that are found between the ILE50 (from both protease chains) and the ligand residue interface using an overlapping radius of 3 Å centered round any of the residue atoms. Modelled receptors were aligned to the SQV-containing template to get a common docking center (20.147, 29.716, 16.093) using ProDy [271]. These were then protonated to pH7 using PDB2PQR (version 2.1.0) [272] with the PROPKA algorithm prior to merging non-polar hydrogen atoms and setting Gasteiger partial charges with AutoDockTools [273], whilst keeping the modelled water where present. The grid box size was defined with sides (20 x 26 x 20 Å^3). Each Vina run was seeded with the number 10000, setting the exhaustiveness at 16 to run with 12 cores per job. Lowest energy poses were harvested from each complex to retrieve their binding affinities and values for each of the energy terms. Thereafter, X-Score was used to cross-evaluate the docked ligands (after converting the docked PDBQT files in MOL2 format) within the aligned protonated target receptors. It should be noted that the X-Score tool does not include the explicit placed water molecule in its calculation. Multiple sections of the experiment were run concurrently using GNU Parallel (version 20160422) [274] over the large queue at the Centre for High Performance Computing (CHPC). Finally, all binding energies and available terms were correlated against their actual fold resistance ratios to search for any trend that could be later used for inferring drug resistance in subtype B HIV proteases, and eventually repeat the same experiment in non-B subtypes if large enough sample sizes are available.

3.2.3 Molecular dynamics: Searching for trends in drug binding energies and resistance at several time points in phase space

The aim of this experiment is to improve on previous work reporting that *in silico* drug binding affinities can be utilized to predict ARV resistance in HIV [275, 180] and suggesting that increased simulation time [179] may improve predictability. We design the experiment to determine if any of the binding energies will correlate favourably with actual drug resistance values available from the raw dataset, pre-filtered by our own means. For each of the 8 PI drugs, the top and bottom 100 complexes were chosen from the 10000 docked proteases (from subsection 3.2.2), after ranking them by their PhenoSense Assay fold ratio label. This yielded resistant and susceptible data subsets, which we refer as ensembles for each PI drug. The cut-off values used by Hedlin in [165] were applied to determine the resistance statuses based on these available labels. Thus the MD simulations were started with the protonated receptors (pH7) with the AMBER residue type and docked ligands. These ligands were fully-protonated and converted to PDB format using Open Babel before preparing the topologies for each of the poses (as the partial charges changed due to altered ligand pose). Correct receptor protonation states were restored by GROMACS during topology generation based on the AMBER residue types created by PDB2PQR. After docking, the VEGA software (version 3.1.1) [276] was utilised to restore the fully-protonated state of the ligands before preparing the atom topologies using ACPYPE. Receptor topology files were prepared using the *pdb2gmx* command from within GROMACS. The all-atom AMBER03 force field was used for the system. Non-bonded short-range interactions (van der Waals and Coulombic) cut-off values were set at a maximum of 1.2 nm while long range charged interactions were handled by the smooth Particle Mesh Ewald algorithm. The energy of the system was minimized using the method of steepest descent, after adding SPC-modelled water and neutralizing charges with 0.15 M of sodium chloride in a triclinic periodic box, with a minimum image distance of 1 nm. The same target criteria used earlier for minimization stop were used - a potential energy gradient minimum of 10 $\text{kJmol}^{-1}\text{nm}^{-1}$ for an initial step size of 0.01 nm. As before, the step sizes were automatically halved in case of bad contacts for a maximum of 4 attempts. Temperature (310 K) and pressure (1 atm) were subsequently equilibrated over a period of 50 ps in each case, using the Berendsen [207] and Parrinello-Rahman [208] algorithms respectively. A seed of 10000 was used for velocity generation in the temperature equilibration step. Production MD simulations were then run for a total of 2 ns. For both equilibration and final MD runs, the LINCS constraint was applied to all atoms and a 2 fs time step was used. All simulations were run in parallel using GNU Parallel (version 20160422) [274] over the large queue (2400 cores) at the CHPC, with 24 cores per job. Proteins complexes were finally centered, removing rotational and translational movements using GROMACS's *trjconv* command. C_α RMSD values were then calculated to detect any obvious failures in the removal of periodic boundary conditions before proceeding to further analysis. Finally drug binding energies were scored in-place from each complex at 12 time points (10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000 ps) by extracting the receptors and bound drugs before evaluating them with Vina and idock. As explained in subsection 3.2.2, receptors and ligands were prepared with AutoDockTools prior to calculating correlations with actual fold resistance scores.

3.2.4 Elastic Network Model: Mining for a resistance-related motions

Using the anisotropic network model (ANM), we collect all the modes and search for any clustering pattern that can hopefully differentiate resistant and susceptible protease models. Homology models prepared in subsection 3.2.2 were used as input for analysis using ProDy. The ANM model was thus applied to each of the 10000 protease variants for each of the 8 drugs, decomposing the Hessian matrices to their respective eigenvectors and eigenvalues. Only the first 297 non-zero modes out of 594 modes (corresponding to the 3N degrees of freedom for the HIV proteases, where N is 198 residues) were retained for each of the structures that were coarse-grained by C_α atoms. The default spring constant (γ=1) and cut-off distance of 15 Å were used. Only the first 2 non-trivial modes are reported. Each mode was compared across proteins by stacking them into individual matrices, with each row being the mode from a protease variant for a given drug. The first two principal components of the decomposed matrix were then represented as scatter plots (Figures 3.6a and 3.6b), colouring the representative proteases by their resistance status, with the aim of identifying differential clustering patterns.

3.2.5 Dynamic Cross Correlation (DCC): Searching for correlation patterns associated with resistance

For calculating DCC, equation 3.14 is applied over the MD trajectories for each of the resistant and susceptible protease complexes for each of the 8 PIs. DCC is typically represented as a square matrix, representing pairwise correlations. However, in this experiment 100 matrices are computed for each of the 2 resistance classes and represented in a concise manner to facilitate interpretation. Each DCC matrix is simply linearised by taking values from its upper triangle and stacking them into rectangular matrices to be visualized as two heat maps for each drug. The aim was to investigate for any differentiation, before proceeding further should any consistent pattern be obtained.

3.2.6 Perturbation Response Scanning: Searching for trigger points correlated with drug resistance

Logsdon and co-workers [277] determined that an expanded active site cavity is correlated with reduced binding to protease inhibitors that resistant proteases. Based on this observation, we further investigate the hypothesis that resistant proteases would have a higher likelihood of reaching the opened conformation, compared to the WT protease, for all PIs. We thus model this expected "drug resistant conformation" for each of the 100 resistant and 100 susceptible sequences using a wide-open multi-drug-resistant protease crystal structure (PDB ID: 1TW7) as target structure. MODELLER was used with very slow refinement and a random seed of -10000 to model the target conformation for each sequence. For each protease complex, the corresponding MD topology file was used as the starting protease conformation. For both the initial and final protein conformations, Open Babel (version 2.3.2) was used to convert the PDB files to XYZ format, after the C_α atoms were selected using the *grep* command. PRS (as implemented in MD-TASK) was used for each protease to successively apply the default 250 random perturbations on each residue of the dimer. Additionally, a 2ns-MD trajectory was provided for the covariance matrix calculation in each case.

3.2.7 Residue network analysis: Investigating differential network behaviours associated with resistance

A weighted contact network was constructed by considering all residue pairs (defined by C_β or glycine C_α atoms) with a euclidean distance of less than 0.67 nm. All contacts were simply aggregated along each MD frame before calculating network metrics, namely degree, betweenness, closeness and density. For all the metrics, with the exception of density, the values were stacked together into a matrix to be represented as heat maps. The MD frames were read in strides of 4, starting from 100ps up to 2ns. Using this strategy, we aimed at identifying network properties that might differentiate drug resistance from susceptibility in HIV protease.

3.3 Results and Discussions

3.3.1 Homology modelling and energy-minimization

The simplest structure-based approach is based on homology-modelling and the in-place scoring of ligand binding energy. Energy-minimization was then applied to guide the complex down a local energy minimum for each protease variant using the AMBER03 forcefield after solvation in a triclinic box with SPC water and 0.15 M of NaCl using the GROMACS tool. For each drug, we tested the hypothesis of association between fold resistance ratio and binding energies using a parametric (Pearson's) and a non-parametric (Spearman's) test. While both tests employ a normalized covariance, the Spearman correlation test uses ranks instead of the values themselves thus ignoring the effect of magnitudes for the actual differences occurring between each matched pair (fold ratio and energy score). A good performance for either test would be values approaching 1 or -1 for positive or negative correlations respectively, with 0 denoting the absence of correlation. We unfortunately observed low correlations (Figure 3.2), as seen from the low magnitudes of correlation between fold resistance ratios and energy scores obtained from X-Score and several metrics from AutoDock Vina (overall score and unweighted energy terms) for the drug ATV, which was used as test system for determining suitable parameters, before evaluating same for the remaining PI drugs. Binding energies had initially been calculated after minimizing the solvated complexes with an epsilon value of 10 kJmol^{-1}nm^{-1}, which is the maximum force obtained from the force vector used as the main stopping criterion. Due to the very large number of systems to be minimized, an automatic resolution of bad contacts was devised in a wrapper Python script to automatically halve the initial step size (0.01 nm) upon each minimization failure, by up to four times before reporting a minimization failure. In an attempt to improve energy correlations, a separate run was performed at a lower minimum, with epsilon halved down to 5 kJmol^{-1}nm^{-1} before re-evaluating the correlations. While hydrophobic contributions displayed a comparatively higher absolute correlation with respect to ATV fold resistance ratios, a lower energy minimum

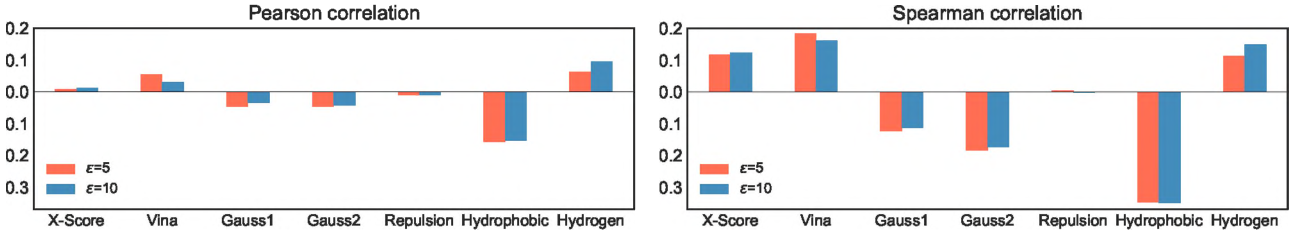


Figure 3.2: Correlations of X-Score and Vina binding energies against fold drug resistance ratio in energy-minimised ATV-bound protease models. Individual energy terms from the Vina potential are also shown.

did improve AutoDock Vina's overall score correlation together with both of its Gaussian terms but decreased the hydrogen bonding energetic contributions. Assuming the protease sequence fold ratio labels were predominantly correct, we hypothesize that the disregard of bond angles in the estimation of hydrogen bonding energies may be at cause for the observed differences happening after further minimization for that energy term. Despite a simple estimation by linear interpolation based on distance, atom typing and the application of discontinuity, the hydrophobic interaction appears to be a major contributing factor associated with drug resistance. It is possible that a highly hydrophobic interior of the binding cavity may have a higher influence on drug positioning and hence better describes the energetic effects related to resistance. X-Score's performance was generally poor. We note a potentially interesting finding from this experiment when comparing the corresponding Pearson's and Spearman correlations for the same energy terms, which is that of an improved predictive ability for each score when the energy values are swapped by their ranks. This may point to the fact that both tools will tend to have a relatively better ranking ability but a comparatively lower capacity to predict accurate energy values, for the same drug. Though not directly related to drug resistance prediction, we predict that the relationship between ligand binding energy and ranking will definitely be impacted differently as a function of properties of the receptor binding surface in high-throughput virtual drug screening experiments.

As a quality control for the 10,000 models used in this experiment, we show the distribution of z-DOPE values for the models in Figure 3.3. z-DOPE scores tending towards or lower than -1 point to native-like conformations while positive ones are likely poor models [278]. We can see that the large majority (>75%) of the models are below a score of -0.8, with very few outliers around -0.4, meaning that the models were generally conducive for biological investigations.

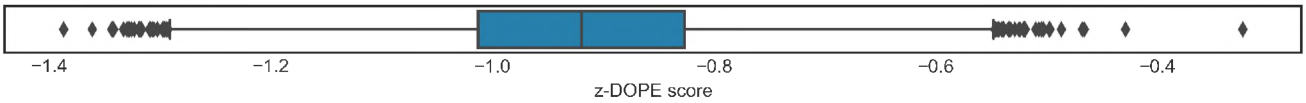


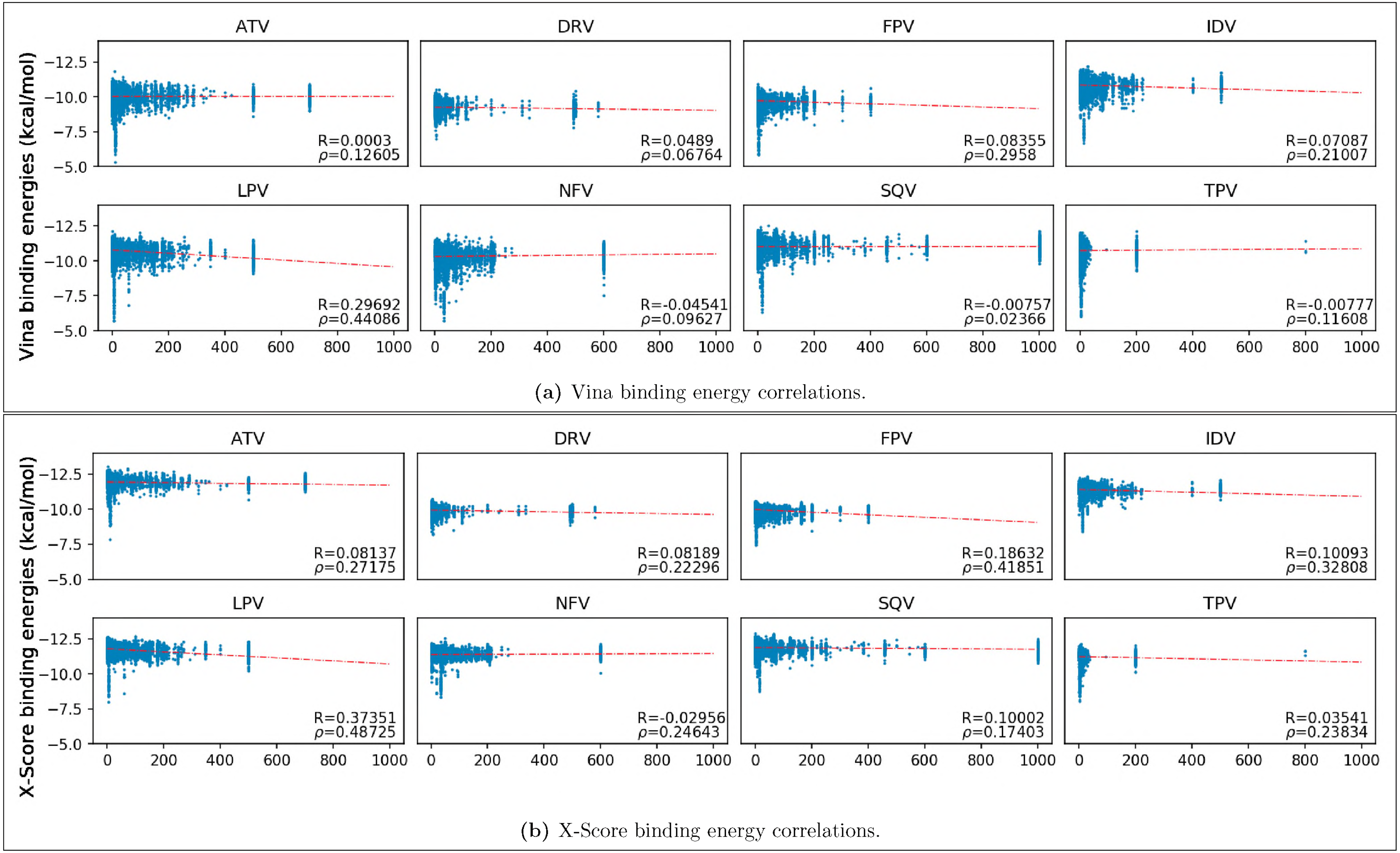
Figure 3.3: Box plot showing the z-DOPE distribution for models of the ATV complexes before energy minimization. The whiskers are placed at any value that is just below 1.5 times the interquartile range running along each direction from the box, and any value away from them is treated as an outlier.

3.3.2 The use of docking to estimate ligand affinity

Prior to performing high throughput docking using known PI drugs against the protease variants, docking controls were performed by removing and re-docking the drug co-crystallized within template protease X-ray crystal structures. Compared to in-place scoring where ligand flexibility was partly accounted for by carrying out energy minimization after protein homology-modelling, flexible ligand docking allows for a widened conformational search space and an improved ligand positioning within the binding site. X-Score was used to re-score the docked drug poses produced using Vina, to look for any improvement in the correlations with drug fold resistance. For the positive controls, we investigated the presence and absence of a water molecule at the flap-ligand interface, using docking parameters for exhaustiveness, grid box specifications as described in subsection 3.2.2. This water was automatically extracted from the crystal structure using an in-house Python script implementing an algorithm that searches for any water shared between the crystallized ligand and the protease ILE50 residue (from both chains A and B) at a cut-off distance of 3 Å. It can be seen from Table 3.1 that the interfacial water largely improves the RMSD of the docked drug from its originally crystallized pose for all cases, except for the drug ATV where the RMSD was slightly higher. As no water was found crystallized at the flap/ligand interface for the modelling template containing the drug TPV, it was omitted for the high-throughput docking for this drug, as it is not needed for ligand stabilization within the protease active site. As our experimental set-up agreed with literature, which describes the requirement of this interfacial water for ligand stabilization, with the exception of TPV, we proceeded with the docking experiment targeting the different sets of HIV protease variants. X-Score was used to re-score the docked poses in all cases from the poses generated by Vina, before estimating the correlations of their binding energies against actual fold resistance ratios. As seen in Figures 3.4a and 3.4b, which show the performances of both tools for all 8 PI drugs, poor correlations were generally obtained as they were all closer to zero than one. As observed previously in the case of modelling the ligand along with the protein followed by energy-minimization step, the Pearson correlation is generally low but improves when the ranks are used for the Spearman's test. It would appear that X-Score generally slightly improves the binding energies in each case, even though the final outcomes do not show strong correlations. The highest correlation is obtained with the drug LPV, using X-Score with Spearman's test. Based on the distribution of samples on the scatter plot, it is also possible that having a more homogeneous spread of actual fold scores along the x-axis would improve on the estimations of correlation in each case. As the estimation of binding energies were from a single time point, the next experiment was designed to factor in the time evolution of binding energies in an attempt to investigate any improvement in correlation with drug resistance.

Table 3.1: Comparison of RMSD values from ligand control docking in the presence and absence of crystallized water for each PI drug.

Docking condition	ATV	DRV	FPV	IDV	LPV	NFV	SQV	TPV
Without flap water	1.103	1.404	0.834	0.250	1.654	1.079	0.891	1.024
With flap water	1.243	0.972	0.239	0.245	0.819	0.290	0.892	NA



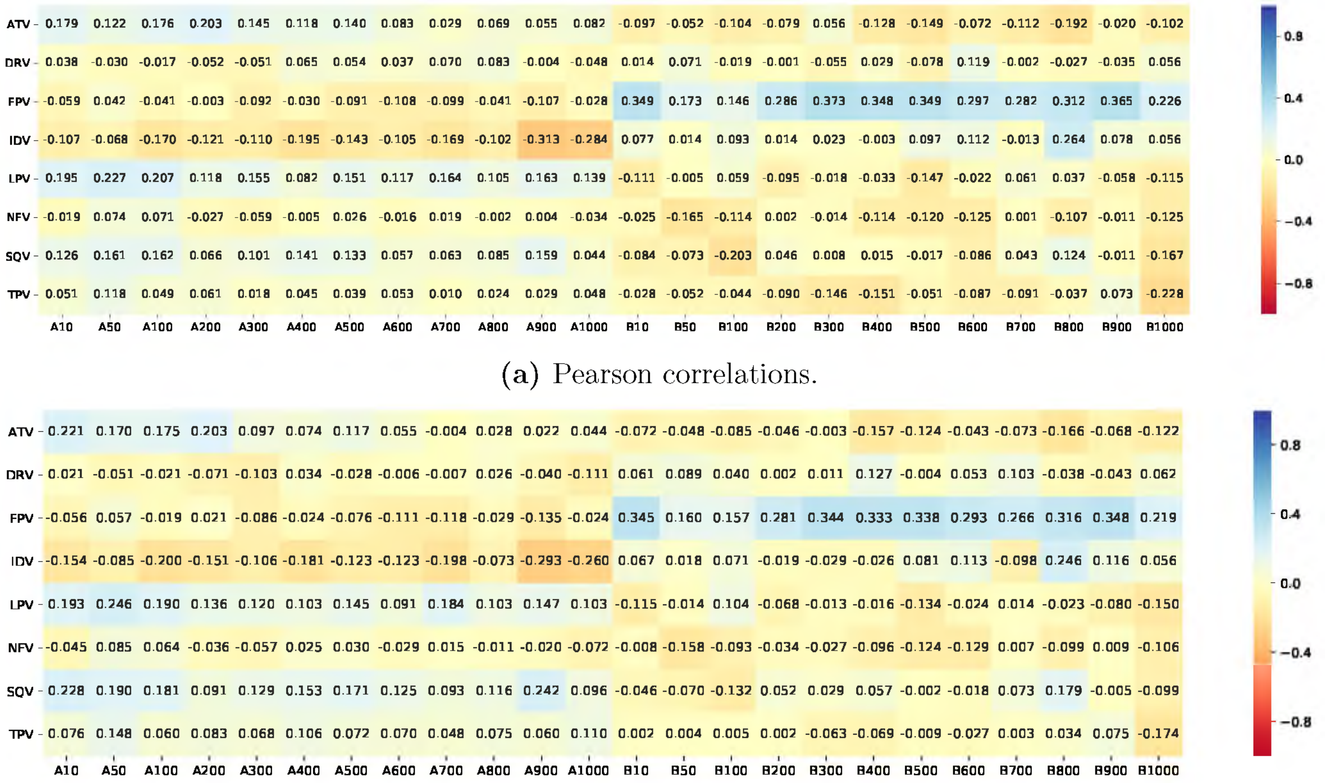
(a) Vina binding energy correlations.

(b) X-Score binding energy correlations.

Figure 3.4: Pearson and Spearman correlations of binding energies from (a) Vina and (b) X-Score against actual drug fold resistance ratios (x-axis) for the 8 PI drugs.

3.3.3 Investigating energy patterns during molecular dynamics

Performing molecular dynamics allows for receptor mobility, in an attempt to improve residue positioning with respect to the complexed ligand, within a system containing explicitly-modelled water and ions. Due to these added components, the systems containing the docked complexes had to be energy-minimized before simulating motion under the assumptions of Newtonian mechanics at physiological conditions under the energetic constraints of the AMBER03 forcefield. Temperature and pressure equilibration were also performed prior to the production run to maintain a temperature of 310 K and pressure of 1 bar. Rotational bond lengthening was corrected by the LINCS algorithm applied over all atoms. With this set-up, ligand binding energies were estimated from frames sampled at 12 time points, using AutoDock Vina and idock [279]. Both programs share the exact same scoring functions, but have some minor implementation details, especially in the optimization of conformational search algorithms. Correlations were evaluated at each of the time points to firstly determine if the predictability of drug fold score ratios was improved when the protein was allowed to move, and secondly whether performance improved over time, as observed by [179] for run times of 0.1-1ps of MD simulation. As seen in Figures 3.5a and 3.5b for the Pearson and Spearman tests of correlation, correlations remain globally of low magnitudes, and do not seem to generally improve on the results obtained from docking, but rather seem to oscillate around their initially recorded energies at 10ps, if not worsening in several cases. Despite low absolute values of correlation, the best performances are observed after 0.9 ns for IDV and at



(a) Pearson correlations.

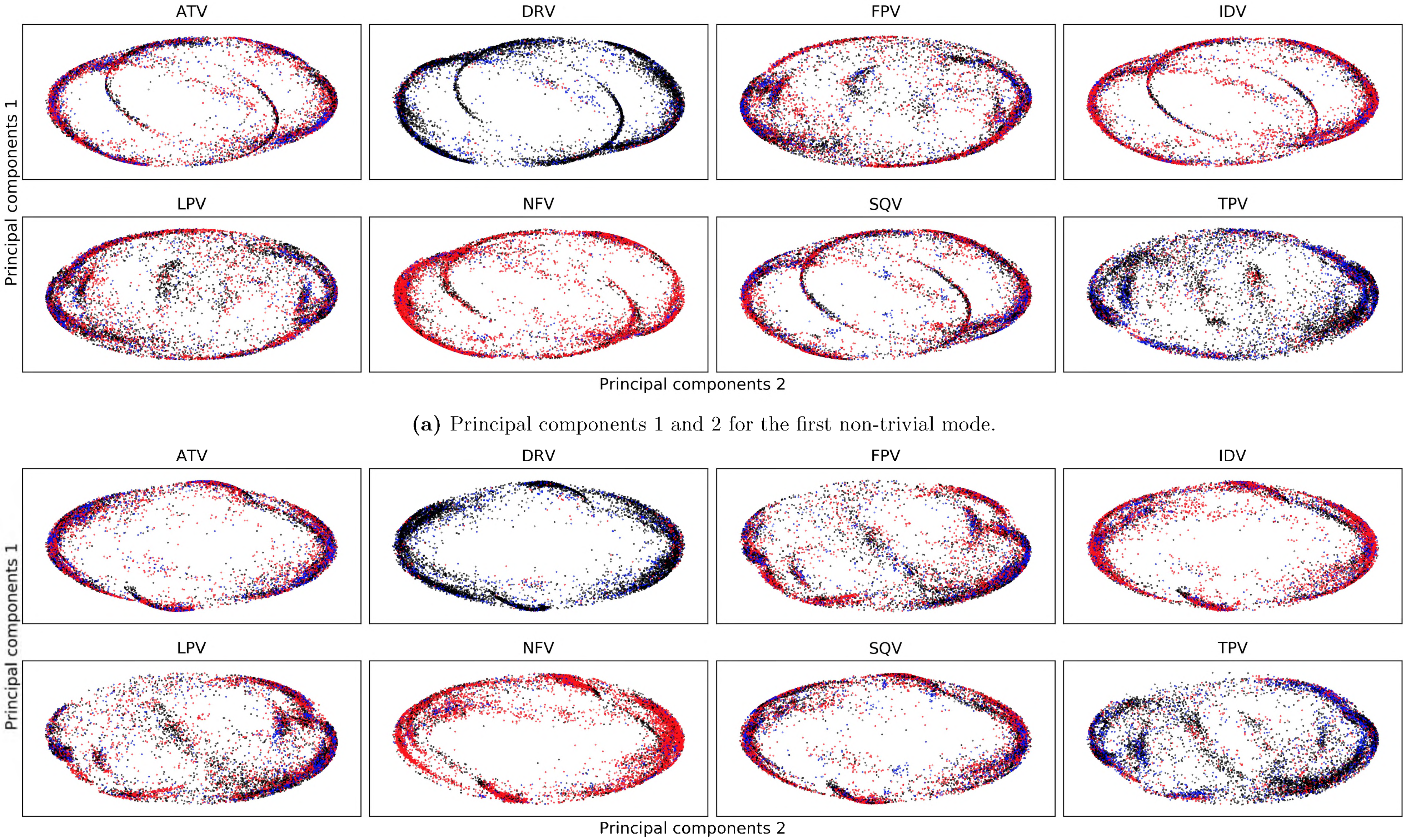
(b) Spearman correlations.

Figure 3.5: Correlations of drug binding energies scored in-place at different intervals along MD trajectories using the ligand docking tools Vina (A10-A1000) and idock (B10-B1000), for each of the 8 PI drugs. The numbers represent the time (in picoseconds) at which energy-scoring was performed.

10 ps for idock, using Pearson's correlation coefficient. As samples have been taken at different time points without seeing any consistent pattern in favour of drug resistance for any of the PI drugs despite the large number of labelled complexes, we suspect the errors to mainly come from the insufficiency of non-equilibrium energy-scoring functions to properly describe the energetics of receptor-ligand interactions. While more involved methods, such as the MM-PBSA might have been attempted, these are too computationally-expensive for the number of complexes. Expert opinion by Genheden and Ryde raised many concerns about the reproducibility and accuracy of the MM-PBSA method, for instance due to the use of continuum-solvation (implicit) models, sampling method and treatment of entropy, which can overestimate the ligand effects, in addition to large ranges of uncertainty in energy values that can make comparisons impractical [280, 172].

3.3.4 Mining for a resistance-related motion using Normal Modes

The calculation of normal modes was used here to investigate its potential use in highlighting resistance-related differences. By comparing similarly-ranked modes prevailing across resistance states, we hoped to either use modes to firstly predict drug-resistance and secondly map these associated motions, on a 3D structure. The ordering of modes were assumed to be analogous for complexes modelled from the same template. For instance, in the case of ATV 10,000 models comprising of resistant, susceptible and intermediate resistance sequences were assumed to contain comparable modes, which are typically arranged in descending order of eigenvalue magnitude. For a given drug, the first two non-trivial modes were gathered from each modelled protease. Matrix decomposition was separately performed on each mode across the complexes and plotting the first and second principal components as a scatter plot, one for first and another for the second non-trivial mode. Each mode consisted of 594 (198 $\times 3$) components, for any given protease structure. Each sample was then coloured according to its level of resistance (resistant, intermediate and susceptible) against the given drug. Only the non-trivial modes one and two are shown in Figure 3.6. As can be seen, no clear clustering pattern delineating resistant sequences from the susceptible ones can be obtained from the 2 modes for any of the protease inhibitors. Even though each set of protein variant conformations for each drug should be unrelated, variations of elliptical silhouettes were obtained for each of the 8 drugs, however we do not have a clear explanation for this behaviour. Also visible are the partial symmetries occurring mainly from the first non-trivial modes, which we posit may be related to the two-fold symmetry of the dimers. No meaningful clustering pattern was obtained when more modes were explored using the same approach for the drug ATV. Due to this behaviour, we hypothesised that resistance-related activity may not be a result of sampling from mutually-exclusive sets of receptor conformations, but may be a result of their altered probability distributions instead, such that both states can elicit the same receptor shapes, but favour some more than others over time. However, due to the static input for the ANM calculation, such may not be visible. Another possibility may be linked to the loss of side chain information as a result of coarse-graining, such that the perceived vibrations are mainly a result exploring the system of connected nodes constrained only by the observed carbon atom positions along the backbone, with Hookean spring properties defined by the cut-off radius.



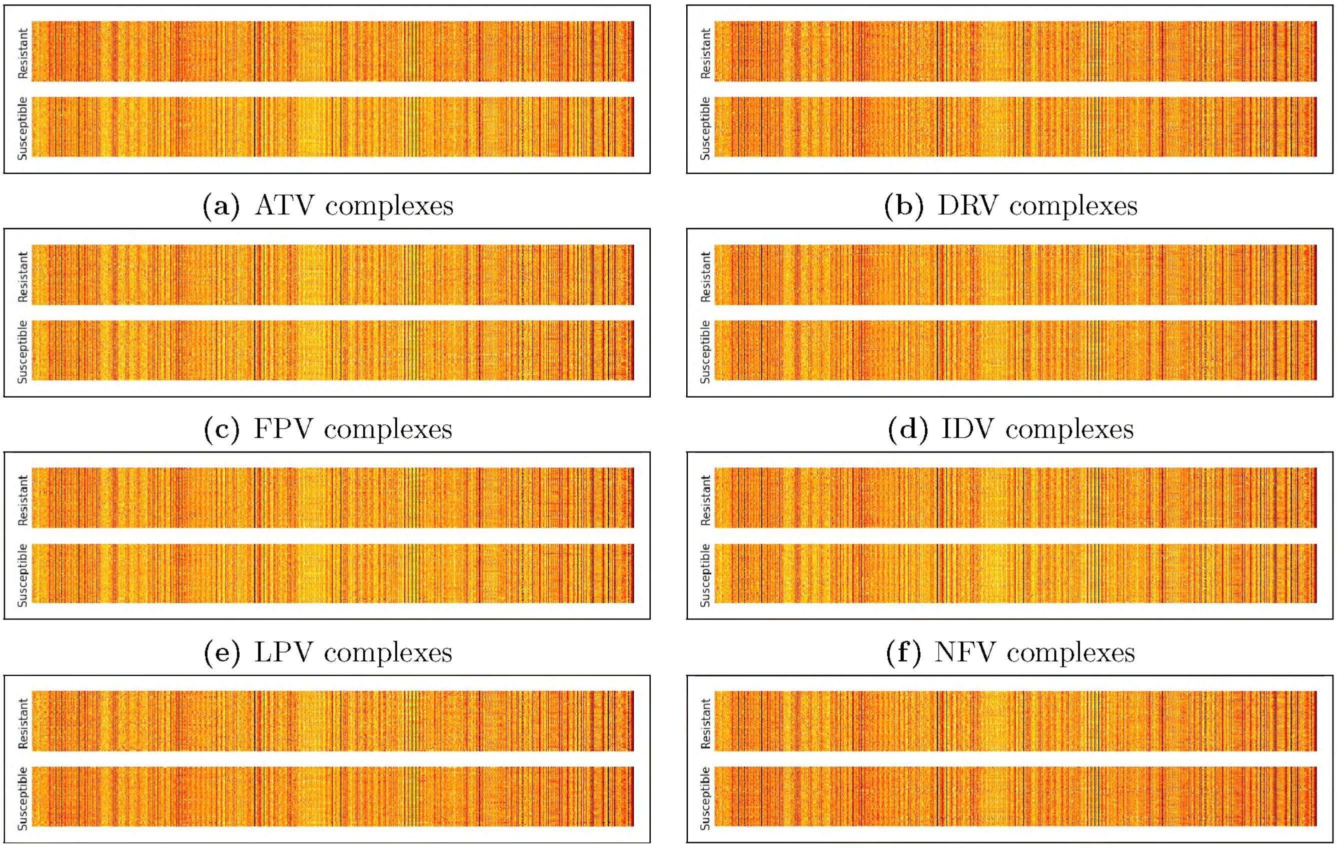
(a) Principal components 1 and 2 for the first non-trivial mode.

(b) Principal components 1 and 2 for the second non-trivial mode.

Figure 3.6: Principal components analysis for the first and second slowest ANM modes gathered from each of 10000 modelled proteases. Each coloured dot represents a protease. Drug resistant, intermediate and susceptible samples are coloured in red, blue and black respectively for each PI drug.

3.3.5 Searching for correlated residue positions associated with resistance using Dynamic Cross Correlation (DCC)

DCC results are shown in Figure 3.7, where the resistant and susceptible sequence sets are compared for each PI drug. The cross-correlation values corresponding to the amino acid residue pairs are laid out into an ascending ordered pair of the matrix indices. While not very meaningful by themselves, they can be compared across sequences as the order of residue pairs is identical in each case. We conclude from this "profiling" experiment that this DCC implementation cannot reveal any useful differences coinciding with drug resistance for all of the PI drugs, inferred from α carbon atom movement recorded over a period of 2ns. Non-specific differences were however observed mainly within each of the drug complexes. Variations between drugs were present, though very minute and may be attributable to either (1) the modelling templates used in each case which may have created slightly different energy surfaces or (2) be a consequence of the different ligand in each case, propagating drug-specific signals from their point of contact within the active site. It is remarkable that such conservation prevails in two very different states of ARV drug resistance, as it suggests a level drug resistance-agnostic order in residue movement. It is however also possible that the differential changes occur over longer time scales. However such an approach would be very lengthy, since same would need be replication as well.



(a) ATV complexes

(b) DRV complexes

(c) FPV complexes

(d) IDV complexes

(e) LPV complexes

(f) NFV complexes

(g) SQV complexes

(h) TPV complexes

Figure 3.7: Dynamic Cross Correlation for protease inhibitor complexes. Linearised DCC values are shown for 100 resistant and 100 susceptible complexes in each case. Protease sequences are displayed along the y-axes while homologous residue pairs (in no biologically-meaningful order) are shown on the x-axes

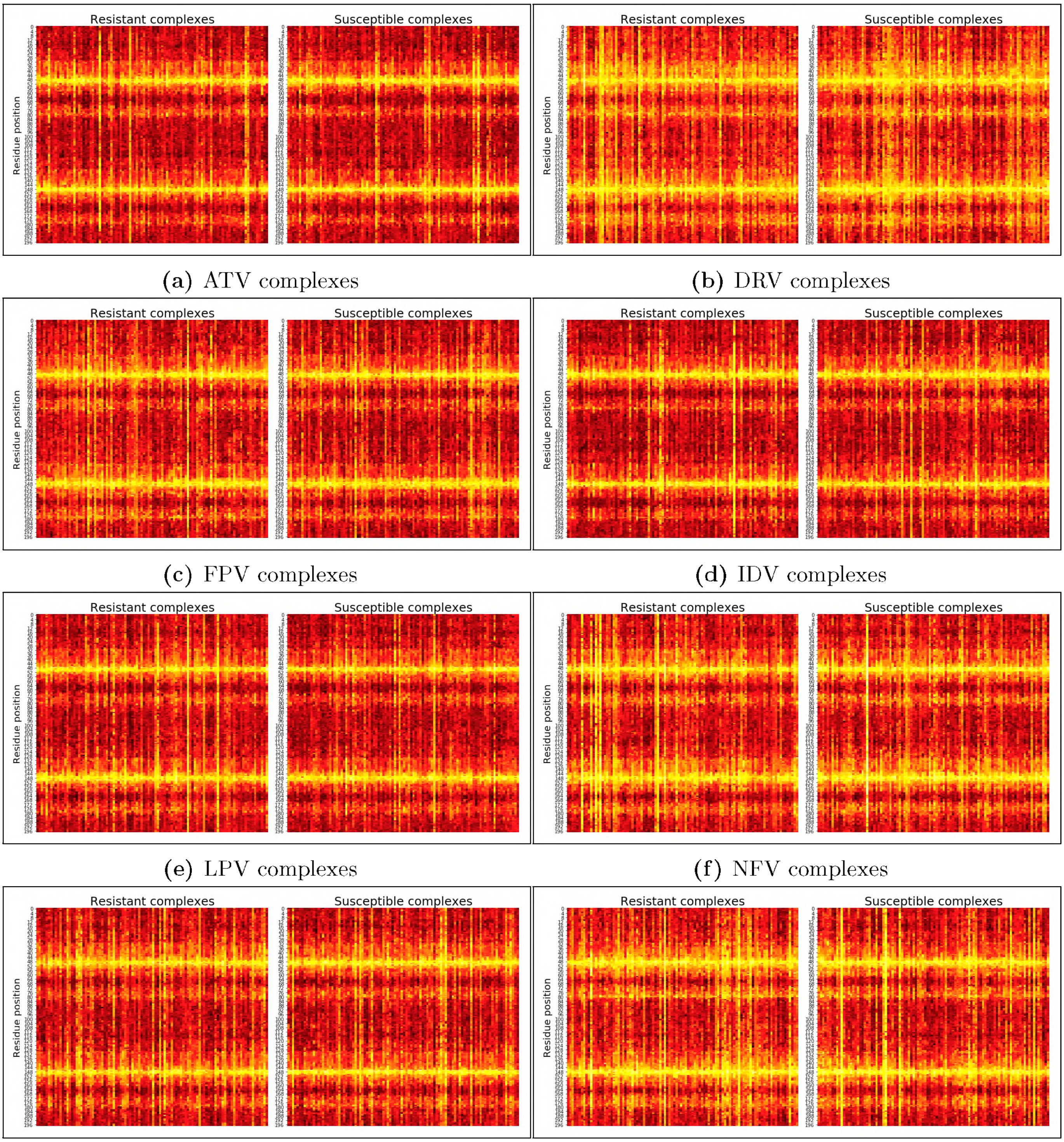
3.3.6 Application of Perturbation Response Scanning to search for trigger residue positions correlated with the resistance state

The main aim of the analysis is to compare the trigger residues prevailing in the different resistant states for each drug, with respect to a wide-opened, multi-drug resistant receptor conformation. Our assumption is that the susceptible state should theoretically have a lesser tendency of having a wide-opened active site, which is otherwise a mechanism that leads to drug exit from the binding cavity. According to Logsdon and co-workers [277] an expanded active site cavity is correlated with a reduced binding to protease inhibitors. This is in agreement with both the substrate envelope hypothesis and ligand-binding thermodynamics experiments [281, 109]. PRS applies forces of random magnitude and direction, in sequence to all residues (represented by C_α atoms) onto a starting conformation of protein, using the covariance matrix obtained from an MD simulation. The resulting coordinates are then correlated to a desired final structure, in this case a wide open conformation multi-drug resistant (MDR) protease (1TW7 [109]). From Figure 3.8, we observe some form of symmetry in the distribution of trigger residue correlations, which stems from the fact that the top 99 residues are from chain A and the rest are from chain B. In each case, the areas of highest correlation to the MDR opened-conformation point to regions surrounding indices 49 and 148, which are flap residues flanking the ILE50 residue from both chains of the homodimer. Also visible, but to a lesser degree are positive correlations occurring at indices 80 and 179, which correspond to residue 79 from the 80's loop - an area forming part of the binding cavity. From this perturbation response analysis, it would seem that both resistant and susceptible HIV proteases apply similar modes of motion leading to flap opening, irrespective of drug resistance mutations or PI drug present in the active site. Unfortunately, from the results of this experiment, a differentiating response was not obtained, which may suggest some more complex motions, for instance manifesting as a differential sampling of conformations happening over time. The PRS analysis inherits from the same limitations coming from Normal Mode Analysis, which due to the reliance on an energy minimum are only applicable in the immediate vicinity around an equilibrium from the chosen protein conformation. Additionally, while practical for enabling comparisons of proteins with different atomic compositions, protein coarse-graining may be reducing the information content, thus decreasing its sensitivity to side chain effects.

3.3.7 Residue network analysis for identifying differential network behaviours associated with resistance

As all of the energy-based and correlation-based methods did not show meaningful resistance-associated patterns that could be utilized for identifying any clear and consistent differential signal, the simpler concept of networks was employed in an attempt to extract several aspects from the pairwise relationships prevailing between residue pairs over the course of dynamics. While energy minimization as used by Ozbaykal was an attractive idea due to the reduction of computational demands to construct a single network [261], we found that minor changes in a minimization criterion (the number of steps, which we will refer to as EM steps) could alter the behaviour of network metrics for the same starting receptor topology. By varying the number of

EM steps, we found significant changes in the difference of averaged geodesics (wild type versus 20 variants, shown in Figure 3.9), which upon hierarchical clustering, displayed a very high degree of clustering based mainly on the number of minimization steps used (Figure 3.10) instead of the sequence variation itself, which points to the fact that using intermediate number (100 - 5000) of EM steps would bias and affect reproducibility of contact inference in our case.



(a) ATV complexes

(b) DRV complexes

(c) FPV complexes

(d) IDV complexes

(e) LPV complexes

(f) NFV complexes

(g) SQV complexes

(h) TPV complexes

Figure 3.8: Perturbation Response Scanning of protease inhibitor complexes. For each PI drug, resistant complexes are shown on the left while susceptible ones are on the right of each figure frame. The colour scale ranges from yellow through red to black, representing the correlation values from 1 to 0 against models built from an opened-conformation multi-drug resistant target protease.

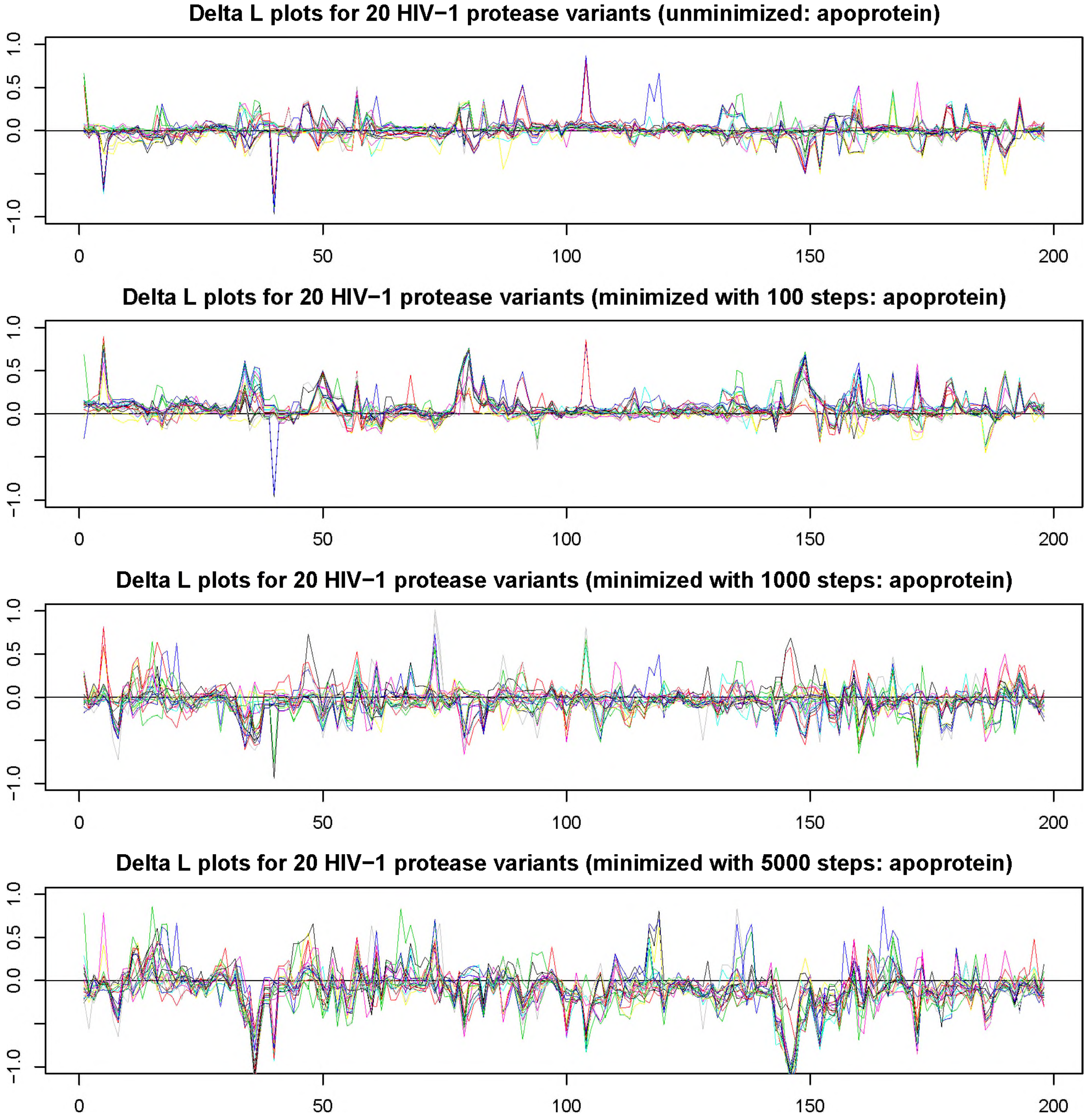


Figure 3.9: Assessment of the impact of the number of energy minimization steps on network behaviour using the difference of average reachability ΔL calculated between a consensus protease and 20 variants. The different coloured lines denote the individual protease variants and are matching across the different sub-figures. The x-axis represents the protease residue positions for chains A (1-99) and B (100-198).

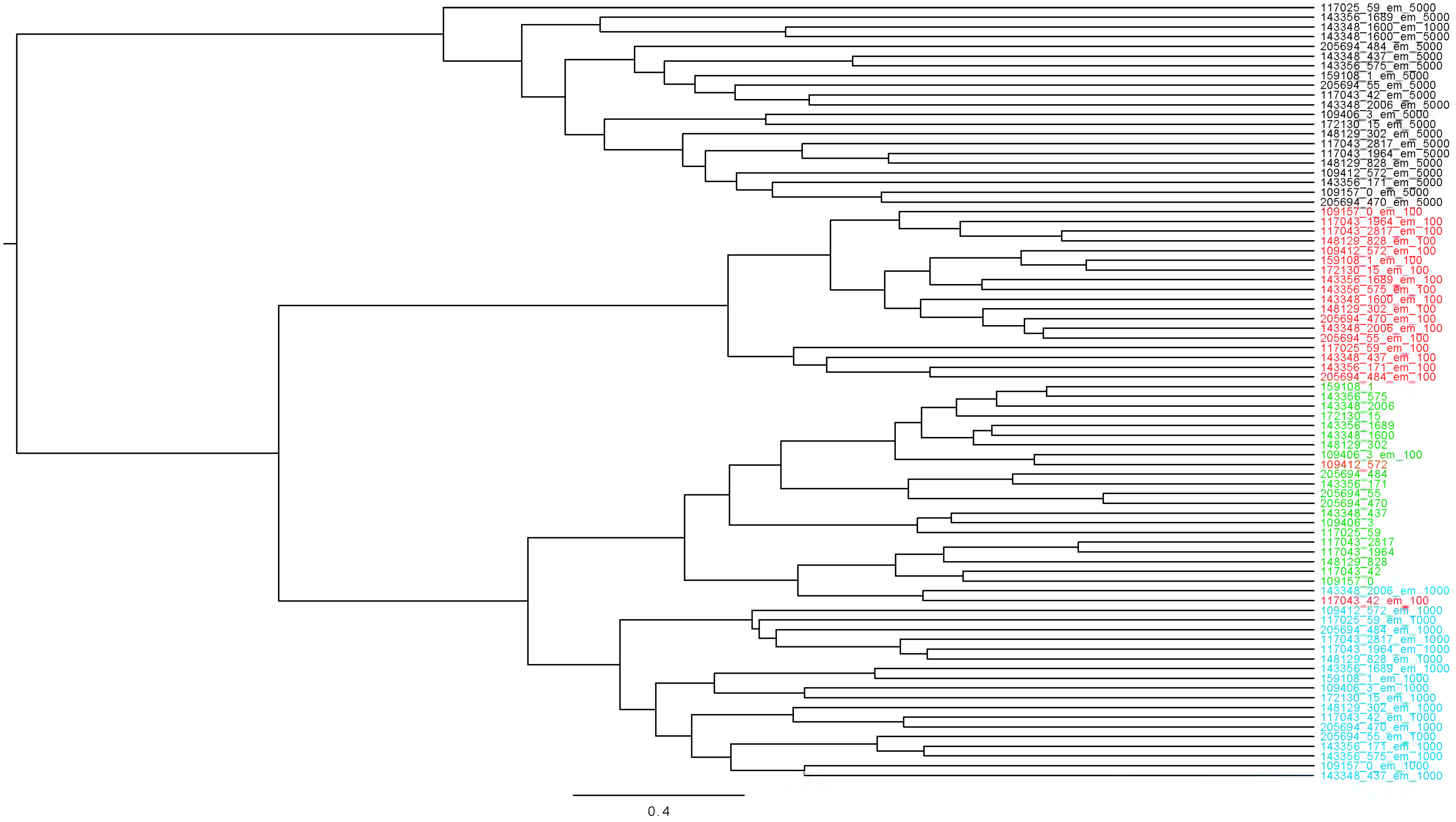


Figure 3.10: Comparing the impact of the number of energy minimization steps on network behaviour using hierarchical clustering with Euclidean distance ΔL for the 20 protease variants. Un-minimized proteins structures are in green while minimized ones are in red (100 steps), cyan (1000 steps) and black (5000 steps). The first section of the node labels comprise the Stanford HIVdb sequence ID, followed by a unique number corresponding to the variant (after sequence expansion and filtering), and the number of minimization steps following the keyword “em”.

At the other extreme, a high number of EM steps was avoided as several researchers mention the effect of over-minimization as a problem related to the fact that current force-fields are not accurate enough representations of atomic models [282, 283, 284, 285, 286] to proceed far down in the energy landscape. Quantum effects become more important as inter-atomic distances converge, for instance as demonstrated in the photosynthetic bacterium *Chromatium vinsosum* where electron-transfer is observed when a cytochrome molecule and bacteriochlorophyll are within the van der Waals contact distance, with insulation happening otherwise [287]. The Coulomb potential coupled with Newtonian-mechanics cannot model such phenomena as they have very limited treatment for handling the effects of electrons, such as the fixing of atomic partial charges, which thus ignores polarisation from electron clouds. Our solution was therefore to calculate networks over the course of MD simulation, in a manner similar but not identical to work done by Doshi and co-workers, in that all contacts are retained [262].

For analysing the time series data sets (MD), we have used the degree, betweenness, closeness and density to expose and compare the different facets of pairwise residue contact behaviour happening over time. While the first three metrics were averaged by calculating the centrality at several time points before averaging them, box plots were used to represent the distributions of averaged network densities from each complex composing the individual drug resistance ensembles.

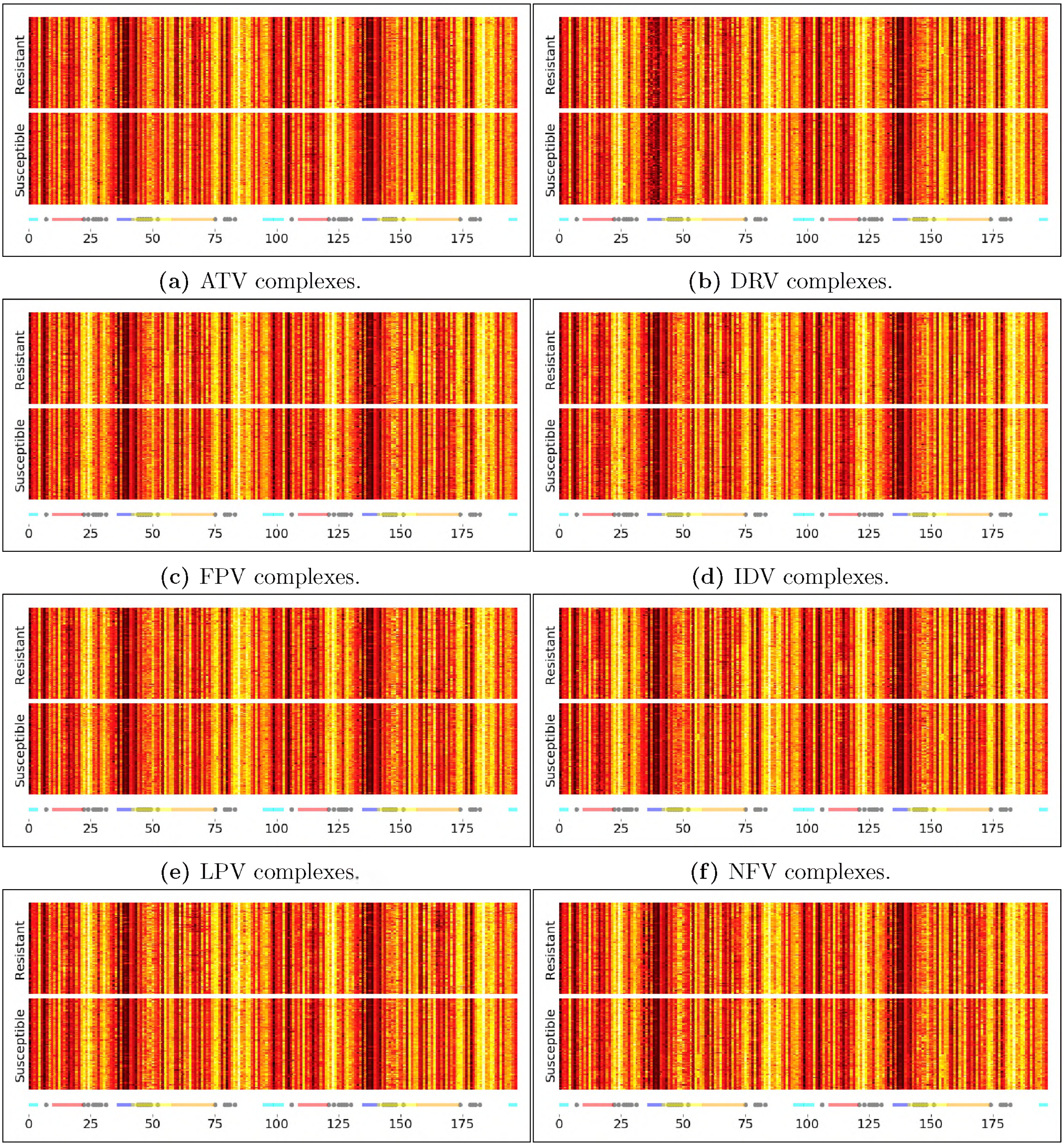
From the **degree centrality** plots in Figure 3.11, no consistent difference between the classes of drug resistance were observed whether within or between drugs. High degree centrality points to residues that have a large number of neighbours and thus give an idea of local density around each residue, and thus may be partly related to the radius of gyration, which is instead generally computed for the whole molecule. From the mapping of protease functional residues, we see that the flaps have low connectivities in each chain, and is related to the fact that they are composed of a surface-exposed mobile loop, which relieves some of the mechanical constraints that would be present if they were buried or formed part of a regular secondary structure, such as β-sheets and α-helices. Residues close to and including the catalytic ASP25 from both chains have very high connectivities, most likely due to stabilisations conferred by the bound ligand in the active site and strong stabilisations by hydrogen bonding networks [288] at the dimer interface, in addition to being buried, despite being part of a loop. Also of consistently high degree centrality is residue 86 (from both chains of the dimer), which is found at the base of the 80's loop in close proximity to the fireman's grip and part of the TIM barrel. As it turns out, this glycine residue is highly-conserved, and mutation experiments have shown severely reduced enzymatic function, showing no evidence of substrate interaction from NMR experiments [289].

Betweenness centrality (BC) incorporates more distal effects by counting the number of paths going through each residue. In the context of social network analysis, a person of high betweenness centrality would be a key player in communicating information between two groups. In our case, the highest betweenness centrality for any given residue would be a result of highly-connected clusters found on both sides of that particular residue, as explained by Kasahara and co-workers [226] who mention that BC increases for residues bridging cliques (highly interconnected group). All complexes irrespective of resistance state, variation or PI highlighted the catalytic ASP25 for chains A and B, which unfortunately does not help characterize resistance. However this informs us that the protease monomers are held together very strongly by interactions that stabilize these two aspartates, which is an already known phenomenon that explains the name fireman's grip for the connected loops held together by a strong hydrogen-bonding network. Comparing the degree and betweenness centralities suggests that the catalytic aspartates comprise a highly-connected hub that stays together, irrespective of intra-domain motion occurring from both sides of the dimer. Also, the flaps have relatively lower BC values due to their constant movement that frequently creates and breaks residue contacts. Conversely, by observing the high proportion of binding cavity residues in or close to regions of high BC, one could further infer that relatively high values are a result of their lowered mobility with respect to all the other residues.

Closeness centrality hints at the ease of reachability for a given node to every other node based on average shortest paths leading to it. Based on its implementation in NetworkX [290], larger numbers indicate shorter average distances to every other node [291]. As the connecting edges are all same and positive, the actual geodesics are equivalent to the number of contacts, and thus larger values of closeness occur when the total geodesic is small (equation 3.29) for a given node. With this reasoning, a centrally-located residue would tend to have higher closeness to all surrounding residues while peripheral ones should be generally displaying lower closeness (higher farness or L) values. From Figure 3.13, no discernible resistance-related patterns were observed for any of the drugs. However, it can be seen that the elbow regions have lower closeness, probably because they are surface-exposed and have higher relative mobilities, which would frequently interrupt stabilizing interactions thus impacting reachability. On the other hand, areas around ASP25 (from both chains) have the highest values for the metric, followed by areas close to the catalytic wall, which may be generally stable core residues.

Finally, to compare the distributions of global connectivity between both resistance states, for each drug, **network density** was evaluated and represented as box plots. This metric is the averaged degree centrality, which we hypothesized to show some form of relationship to the radius of gyration, which estimates molecular compaction based on the root mean squared atomic deviation from the molecular center of mass. However, instead of obtaining lower network densities in the resistant-ensembles relative to the susceptible ensemble (reflecting a higher tendency towards a wider binding cavity in resistant HIV protease) we mostly obtained larger network density medians in all resistant complexes, with the exception of SQV. Upon careful comparison of the equations used for computing network density and the radius of gyration, we find that the former is evaluated between every residue pair, whereas the latter is based off evaluations between pairs comprising a common molecular center of mass and each residue. Therefore, they represent different types of information. One possible explanation for the unintuitive observations is that the protease relaxation associated with a wider binding cavity, may be generating new highly-connected hubs elsewhere within the protein, which collectively yield larger degree centralities when compared to analogous values obtainable from the drug-susceptible complexes.

As a general observation from the limited time of the MD simulations and basis for defining network edges, we hypothesise that resistance may not be a state statically defined across time, and may thus be a phenomenon that is defined with a probability of occurrence. Time may also genuinely be insufficient to observe more characteristic motions occurring over longer time scales, however based on our hypothesis suggesting the role of probabilities of sampling certain local conformations, we decided to rethink and redesign the network construction method to factor in statistical differences, as elaborated in Chapter 4.



(a) ATV complexes.

(b) DRV complexes.

(c) FPV complexes.

(d) IDV complexes.

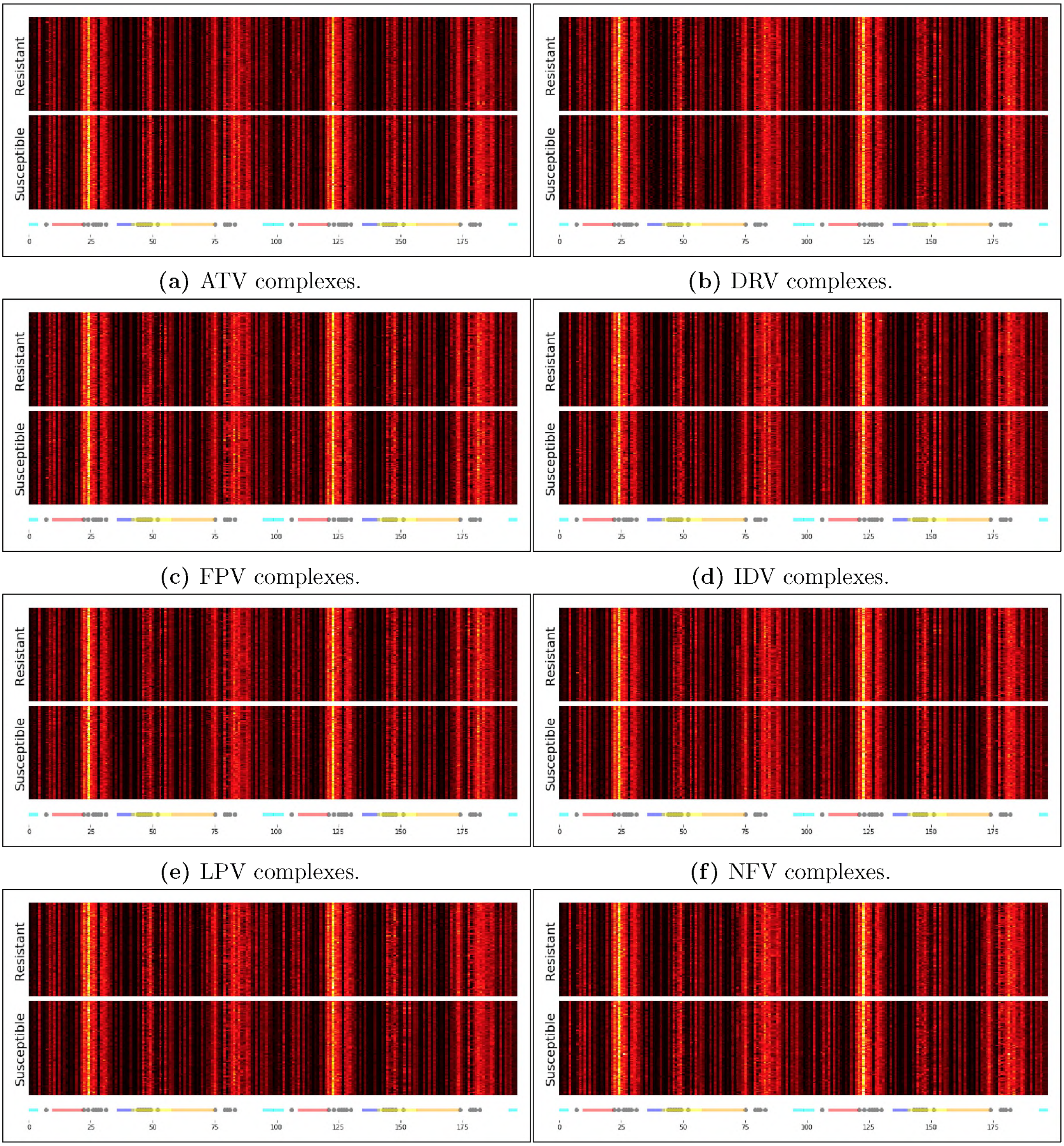
(e) LPV complexes.

(f) NFV complexes.

(g) SQV complexes.

(h) TPV complexes.

Figure 3.11: Time-averaged degree centrality from each MD trajectory from resistant and susceptible ensembles. Highest degree centralities are shown in white, going through yellow and red to black corresponding to decreasing values of centrality. The coloured strips are a mapping of functional residues from HIV protease, showing the fulcrum (red), elbow (blue), flap (yellow), cantilever (orange), interface (cyan) and binding cavity residues (grey spheres).



(a) ATV complexes.

(b) DRV complexes.

(c) FPV complexes.

(d) IDV complexes.

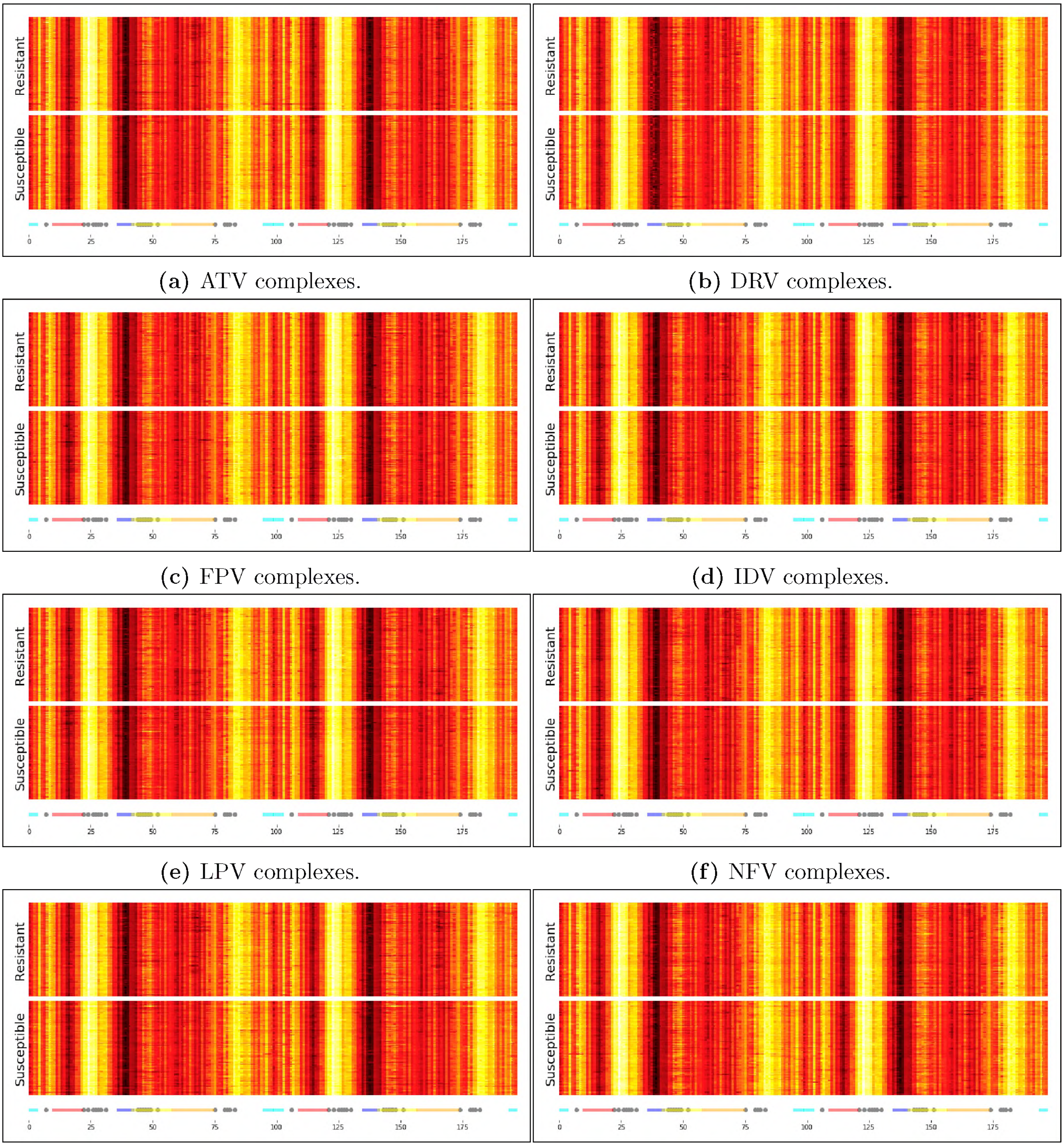
(e) LPV complexes.

(f) NFV complexes.

(g) SQV complexes.

(h) TPV complexes.

Figure 3.12: Time-averaged betweenness centrality from each MD trajectory from resistant and susceptible ensembles. Highest betweenness centralities are shown in white, going through yellow and red to black corresponding to decreasing values of centrality. The coloured strips are a mapping of functional residues from HIV protease, showing the fulcrum (red), elbow (blue), flap (yellow), cantilever (orange), interface (cyan) and binding cavity residues (grey spheres).



(a) ATV complexes.

(b) DRV complexes.

(c) FPV complexes.

(d) IDV complexes.

(e) LPV complexes.

(f) NFV complexes.

(g) SQV complexes.

(h) TPV complexes.

Figure 3.13: Time-averaged closeness centrality from each MD trajectory from resistant and susceptible ensembles. Highest closeness centralities are shown in white, going through yellow and red to black corresponding to decreasing values of centrality. The coloured strips are a mapping of functional residues from HIV protease, showing the fulcrum (red), elbow (blue), flap (yellow), cantilever (orange), interface (cyan) and binding cavity residues (grey spheres).

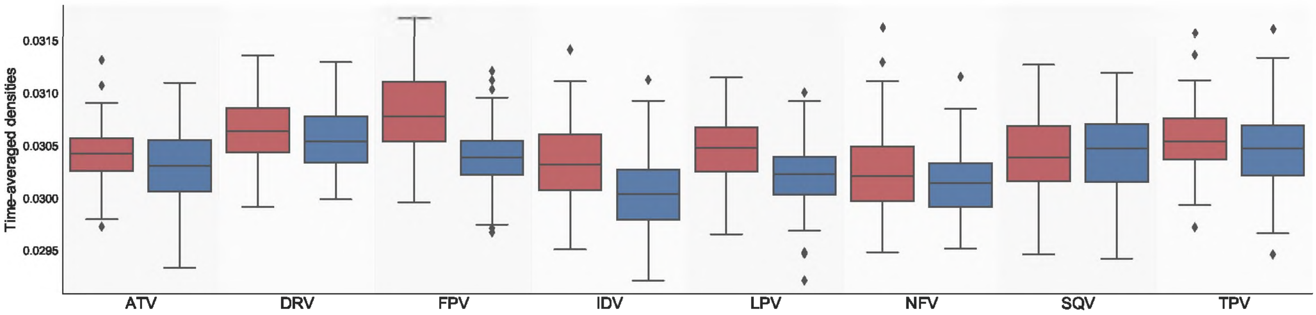


Figure 3.14: Box plot of network density values from MD for each resistance ensemble for the PI drugs. Resistant samples are in red while susceptible ones are coloured blue. Dots represent the outliers sample above 1.5 times the interquartile range.

Table 3.2: All sequence mutations present in each of the resistant subsets against each PI.

PI	**Residue mutations**
ATV	4S, 10IVF, 11LI, 12I, 13VM, 14R, 15V, 16A, 19IV, 20RIVTM, 21D, 22V, 24IM, 32I, 33FV, 34QF, 35DNG, 36IV, 37DS, 41K, 43IT, 46LI, 47V, 48QV, 51A, 53L, 54LVM, 55RN, 57K, 58E, 60E, 61HDN, 62V, 63P, 64VM, 66V, 67YFE, 69RYQ, 70E, 71IV, 72LTVMF, 73STA, 74PS, 75I, 76V, 77I, 79SA, 82LSTA, 83D, 84VA, 89V, 90M, 91S, 92K, 93LM, 95F, 96S
DRV	10IFV, 11LI, 12KPA, 13V, 14R, 15V, 16A, 19QV, 20RTVM, 24M, 30N, 32I, 33F, 34Q, 35DG, 36IVT, 37QDT, 43IQT, 46LI, 47V, 51A, 53L, 54LM, 55RN, 57KG, 58E, 60E, 62V, 63P, 64VM, 66V, 67WY, 69Q, 71IV, 72LVM, 73STA, 77IT, 79S, 82LA, 84V, 85L, 88D, 89V, 90M, 91S, 92K, 93L
FPV	10IVF, 11I, 12KDPV, 13V, 14RT, 15V, 16A, 19QTI, 20RITVM, 24M, 30N, 32I, 33F, 35KDNG, 36LI, 37HDQTP, 41K, 43IT, 46LI, 47V, 48Q, 53L, 54LVM, 55RN, 57KG, 58E, 60E, 61DN, 62V, 63P, 64LV, 66FV, 67YGFE, 68E, 69KLQ, 70TE, 71LIVT, 72LTM, 73ST, 74PS, 76V, 77IT, 79SA, 82LTA, 84VA, 87G, 88D, 89VM, 90M, 91S, 92KR, 93L, 95F
IDV	4A, 7R, 10IFV, 11I, 12KP, 13VM, 15V, 19QIPT, 20RIT, 21D, 22V, 24M, 32I, 33IF, 34TQ, 35DNG, 36LI, 37SDE, 41K, 43TE, 46LI, 47V, 48LVM, 50V, 53LY, 54TVSACM, 57K, 58E, 60NE, 61HNE, 62V, 63P, 64VM, 66VF, 67FE, 68E, 69RYQ, 71TV, 72LTV, 73CSTA, 74DS, 75I, 76V, 77I, 82SFTA, 83D, 84VA, 85V, 89V, 90M, 92K, 93L
LPV	10IFV, 11LI, 12KP, 13V, 14R, 15V, 16A, 19IPQV, 20RTVM, 24I, 30N, 32I, 33FV, 34TQ, 35DAG, 36I, 37EDQTS, 41K, 43QITE, 45R, 46LI, 47VA, 48V, 50V, 51A, 53L, 54LTVMS, 55RN, 57K, 58E, 60E, 61HE, 62V, 63P, 64VM, 65D, 66VF, 67Y, 68E, 69Y, 70E, 71IVT, 72TVM, 73CSTA, 74PAS, 75I, 76V, 77IT, 79S, 82SIFTA, 84V, 85LV, 87G, 88D, 89TV, 90M, 91S, 92K, 93L, 95F
NFV	7R, 10IVF, 11LI, 12IA, 13V, 15V, 16E, 18H, 19I, 20RIVT, 21D, 22V, 24M, 30N, 33IVFM, 35DNG, 36I, 37DSTE, 38F, 41K, 43TN, 45R, 46LI, 48M, 53L, 54VM, 55R, 57K, 58E, 60E, 61HNE, 62V, 63PT, 64V, 65D, 66VF, 67W, 68E, 69R, 71TVI, 72LTV, 73STA, 77I, 82FA, 83D, 84VA, 85V, 88D, 90M, 93L
SQV	10IF, 11LI, 12IP, 13VM, 15V, 18H, 19I, 20RIVT, 21D, 22V, 24M, 32I, 33IVF, 34D, 35DNG, 36IV, 37DS, 41K, 43T, 46LI, 48QV, 53L, 54LVM, 55R, 57K, 58E, 60E, 61HN, 62V, 63PE, 64V, 66FV, 67WFE, 69Q, 71TVI, 72LTVM, 73ST, 74SP, 75I, 76V, 77I, 79A, 82A, 83D, 84VA, 85V, 89V, 90M, 91S, 92K, 93LM, 95V
TPV	10IV, 11LI, 12K, 13V, 14R, 15V, 16A, 19V, 20RTV, 21D, 22V, 24IM, 32I, 33F, 34QDF, 35DG, 36IV, 37D, 43T, 45R, 46LI, 47V, 53L, 54LVM, 55R, 57K, 58E, 60E, 61HN, 62V, 63P, 64V, 66V, 70T, 71IVM, 72TVM, 73ST, 74P, 77I, 82LTA, 83D, 84V, 85L, 87G, 89VM, 90M, 91S, 93LM

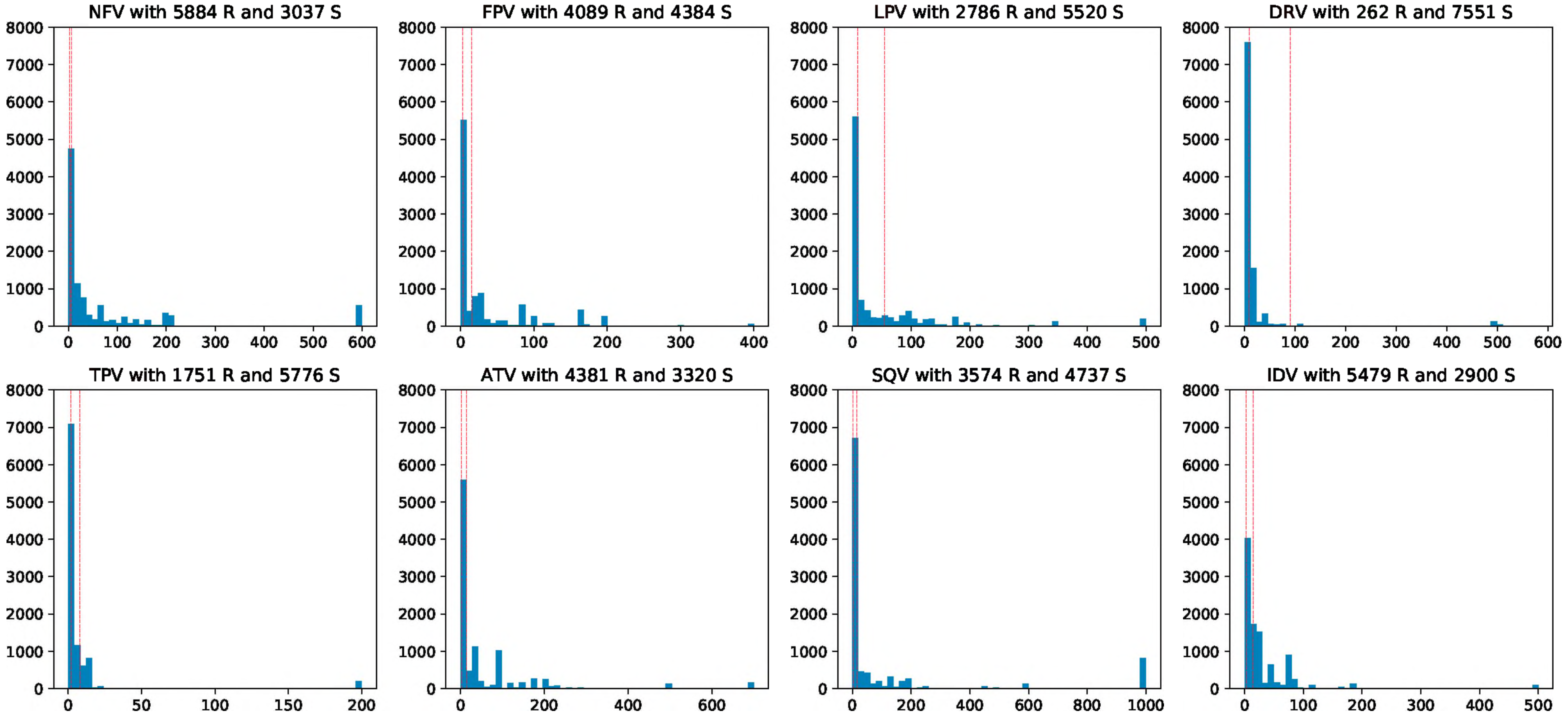


Figure 3.15: Distributions of fold resistance values of filtered datasets for each FDA-approved PI. Red lines demarcate the lower and upper bounds for classifying resistance against each drug. At the top of each plot is the number of sequences predicted to be resistant (R) and susceptible (S) to each antiretroviral.

3.4 Conclusions

This series of experiments consisted in the application of different techniques based on structural approaches for the search of conserved drug resistance-associated patterns by comparing information derived from sets of drug-resistant and drug-susceptible sequences. None of the techniques were observed to possess such discriminative ability, but rather helped in refining the initial research idea and method. To be more descriptive, non-equilibrium potential energies evaluated from (1) homology modelling and energy minimisation, (2) ligand docking and (3) molecular dynamics cannot be directly used to infer PI drug performance as they all correlated poorly with actual drug fold resistance ratios. Insufficiently accurate scoring functions may also be responsible but were chosen for computational efficiency. Nevertheless, the hydrophobic potential from Vina displayed some relatively stronger relationship with drug resistance. Because of the requirement for a single potential energy minimum for constructing the Hessian for both normal mode estimations and perturbation response scanning, the results may be limited to observing differences that are more stationary in time, thus being unsuited to comparing events that are might be of a probabilistic nature. No differential trend in the directionality of pairwise residue movements could be found by the implemented DCC algorithm. The aggregation of residue interaction networks also failed to give any meaningful differential network metric associated with resistance. It is possible that the amount and diversity of sampled conformations for each collection of the receptor/drug complexes for each of the drug resistance states was insufficient, thus missing any differentiating signal(s) entirely, especially if they are rare or manifest over relatively long time scales. Increasing the extent of conformational sampling however, would require a tremendous amount of compute time, such that the number of sequences to be examined would have to be decreased, thus reducing the number of considered drug resistance variations to ultimately affect the generalizability of the experiment. We therefore retained the hypothesis that a differential signal may still exist - not as a distinct state across time but buried as probability distributions, and thus proceeded by redefining network construction using statistically-driven network construction in chapter 4. Under this hypothesis, we suspect that both resistant and susceptible variants can adopt very similar receptor conformations, but do so at different rates.

Chapter 4

Ensemble analysis of protein geometries

This chapter draws from and reproduces certain figures and tables used in the publication listed below. Credit for the reproduced material is given as citations in the respective figure and table captions.

1. **Sheik Amamuddy O**, Bishop Nigel T and Tastan Bishop Ö. "Characterizing early drug resistance-related events using geometric ensembles from HIV protease dynamics." *Scientific Reports*, 2018 December 18. doi: 10.1038/s41598-018-36041-8.

4.1 Introduction

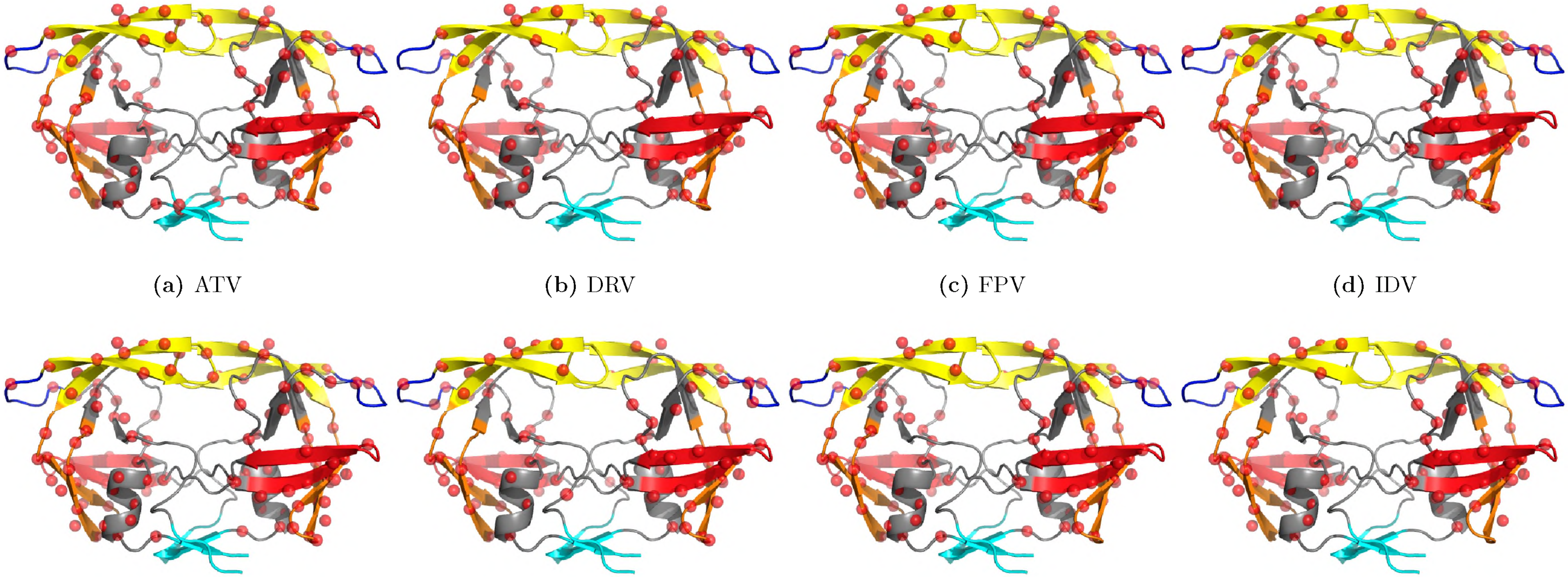
After failing with almost every approach to obtain a resistance signal in Chapter 3, we choose to increase the sensitivity of our analysis with a simple and novel approach based on statistical and network analysis from protein dynamics. Proteins show different modes of motion, some of which can be local, happening over short time-scales, while others are global, potentially adding noise to conserved motions. Depending on where the protein lies in phase space (the description of a system by its set of atomic positions and corresponding momenta), it may allow for the sampling of certain types of motions by going along potential energy surfaces located at the vicinity at that point in time. Global or coarse methods of analysis have failed to detect any signal, which we believed might still be buried in conformational samples we obtained from dynamics of carefully-separated sequences of differing resistance states. This experimental method was designed to pick up any local differences happening between the resistance states, despite the inherent noise originating from different sources of variation. Instead of relying on energy-based methods, residue pairwise distances are computed from molecular dynamics simulations for docked ARV-protease complexes. By comparing analogous residue pairs, it is hoped that a differential behaviour would be observed between groups of drug-resistant and drug-susceptible sequences. For each protein, every pairwise distance is time-averaged. Same is done for all protease dynamics data before using moderated t-tests on each pair of protease ensembles. The meaning of an ensemble here is simply the collection of sequences from one drug resistance class - in this case a drug-resistant versus a drug-susceptible group of complexes. Because of the complex dynamics of biological macromolecules, we use the network concept of preferential attachment to detect isolated residues that have a higher probability of being further away or closer with respect to a given group of residues shown to display significantly different pairwise distances to that lone residue. Degree centrality is used to rank and determine the identity of these characterizing residues. It is a fact that drug resistance is associated with a wider binding cavity [277, 292] that facilitates ligand exit. Several key factors have been tied to this phenomenon over the years. Most prominently, individual mutations that occur at the ligand-binding cavity (also known as primary mutations) can lead to the direct loss of receptor-drug interactions [293] and lowered drug sensitivity [294]. In the absence of drug exposure, viral fitness is generally reduced by such mutations, but can be subsequently offset by the selection of additional mutations at sites distal to the active site (termed accessory mutations) that enhance catalytic activity [295, 296]. These distal residue adaptations have been shown to act in a co-operative fashion in a study done by Ohtaka and co-workers where drug-binding kinetics of multi-drug resistant proteases were assayed [293]. Further, Weber and Agniswamy show reduced dimer stability in proteases bearing the drug resistance mutations L24I, I50V and F53L [23]. Same was observed by Louis and co-workers [297] in multi-drug resistant sequences bearing the L76V mutation. Recently, Goldfarb and co-workers proposed a mechanism of defective hydrophobic sliding occurring in an SQV-resistant mutant containing mutations G48T and L89M [298]. Findings from this experiment attribute the defect to changes in the distribution of van der Waals interactions in the hydrophobic core, resulting in altered protease dynamics. In this study, we further report through high-throughput molecular dynamics simulations that short movements associated with widening of the drug-binding cavity are conserved across multiple highly drug-resistant mutants for all current FDA-approved protease inhibitors. The movements encompass a variety of DRMs for each ARV tested and as such are not residue-specific but position-dependent. The methods employed for molecular modelling are partly stochastic and therefore we applied stringent filtering criteria for network construction using statistically-moderated edges. Additionally the entire experiment is repeated with a different random seed for verifying the reproducibility of our findings. Hornak and co-workers [299] showed that the HIV protease opens via an external downward rotation of the monomers by observing three conformations sampled in molecular dynamics simulations correlating with NMR data. This mechanism involves a co-operative downward movement consisting of the cantilever, fulcrum and flap elbows to result in upward motion of the flaps and catalytic aspartates via rotation about the dimer interface, thought to ease ligand binding. In our experiment, we additionally find a conserved lateral expansion at the flap elbows and contraction situated at the base of the dimerization domain towards the floor of the catalytic site, specifically attributed to drug resistance. Additionally, we find that even though these motions are well-conserved within and between different drugs, they are not accompanied by similarly-preserved angular behaviours, suggesting that the associated angular patterns may be degenerate.

4.2 Methods

4.2.1 Preparing the drug resistance ensemble

As explained in Chapter 3, sequences were filtered and grouped into separate drug resistance groups, here termed ensembles. We show the extent of variations in Figure 4.1.

In order to give clearer context to this section, previous methods described in Chapter 3 are briefly re-introduced in this section. For each of the 8 FDA-approved PIs (ATV, DRV, FPV, IDV, LPV, NFV, SQV,and TPV), 100 high-quality models were built per resistance ensemble, docked against their respective ARV and sampled for conformations by MD. Separate high-resolution (<1.55 Å) drug-bound HIV protease crystal structures were used as template for each drug ensemble. The resulting 1600 HIV protease models were protonated to pH7 using PDB2PQR [272], docked using AutoDock Vina [200] before performing short (2ns) all-atom MD simulations on solvated complexes with set seeds for reproducibility using GROMACS [263]. In all, 3200 independent MD simulations were done. Due to the enormous number of runs, quality control was done by robust removal of periodic boundary conditions (PBC) and using C_α RMSD to check for the absence of sudden jumps. Rotational and translational motions were removed from the trajectories and initially-unstable regions were disregarded in further analysis. Thereafter, we investigated the refinement of network analysis in order to increase the sensitivity of detection to pick up more local characteristic motions.



(a) ATV

(b) DRV

(c) FPV

(d) IDV

(e) LPV

(f) NFV

(g) SQV

(h) TPV

Figure 4.1: Mapping of the variation positions onto 3D structures, for each of the FDA-approved HIV protease inhibitors ATV, DRV, FPV, IDV, LPV, NFV, SQV and TPV used for the drug resistance ensembles. The coloured cartoon representations depict the fulcrum, elbow, flap, cantilever and interface, while the variation positions are depicted as red spheres. Note that even though single spheres are shown, some positions comprise multiple residue variations, some of which are validated drug resistance mutations (major and accessory DRMs) from the 2017 update. Variations from the susceptible ensembles are not shown, for clarity.

4.2.2 Assessing global compaction

Before proceeding to the more sensitive local approaches, global protein compaction was evaluated. The radius of gyration (R_g) was calculated for each protein composing the entire ensemble and summarized as distributions for each PI. Same was evaluated from a replicate run for each ensemble. The algorithm used for calculating the radius of gyration by GROMACS is defined by the following formula:

$$R_g = \left(\frac{\sum_i ||\mathbf{r}_i||^2 m_i}{\sum_i m_i} \right)^{\frac{1}{2}} \tag{4.1}$$

where m_i is the mass of particle i and $\mathbf{r}_i$ is the displacement vector of particle i from the molecular centre of mass. In other words, R_g is the root mean square distance between particles and an axis of rotation, in this case evaluated at each time point to be represented as density distributions, shown in Figure 4.3.

4.2.3 Network construction by statistical tests

Network graphs are well-suited for representing and analysing complex node relationships, however noise coming from temporary residue contacts during dynamics can be overwhelming. We therefore proceeded with a method to filter out edges that are most likely to be fortuitous. Edge refinement was performed by using independent t-tests on time-averaged pairwise distances between residues defined by C_β or glycine C_α atoms. For a given pairwise distance D_{ij} between a residue pair, the time averaged distance $\langle D_{ij} \rangle$ is calculated for every protein in one ensemble. These are then accumulated into a array and same is performed for the corresponding ensemble. Thereafter, for each drug, individual two sample Welsch t-tests (for unequal variance) are performed for each analogous position for the residue pairs between the drug-resistant and drug-susceptible ensembles. After initially evaluating a null hypothesis of there being no difference (two-tailed tests) and finding that it was less informative and noisier, two alternative hypotheses were investigated in the form of one-tailed t-tests, performed at a 99% significance level. The first one evaluated whether the average pairwise distances were larger in the resistant ensemble, whereas the second alternate hypothesis evaluated whether the same distances were larger in the susceptible ensemble. These hypotheses are equivalent to the reverse hypotheses in the opposite ensemble. For clarity, emphasis was laid only on describing resistance-related differences only, as smaller or larger. The filtered results were then represented as an adjacency matrix by converting positions of significant difference into binary contacts comprising ones and zeros for the presence or absence of a contact. Thereafter the results were represented by scaled degree centrality plots contrasting the top 5 hub positions determined to be significantly larger and smaller for the resistant ensembles. Network graphs and calculations were performed using the NetworkX library (version 1.11) [290]. For further visualization, the edges were mapped onto protein structures using the NGLview library (version 1.0) [300]. The MDTraj library (version 1.9.1) [301] was used for reading trajectories and for computing distances and bond angles. Statistical tests were performed using SciPy 1.0.0 [302]. Additionally, bond angles between C_α atoms were compared by performing one-tailed t-tests over analogous arrays of bond angles across drug resistance ensembles and recording the p-values for larger and smaller angles. Only those bond angles with a -log(p-value) above 2.5 standard deviations were retained for both the larger and smaller angles. The negative logarithmic transformation highlighted the most significant p-values by showing the magnitude of their exponents. For both angles and distances, Bonferroni corrections were applied to correct for multiple testing by dividing critical t values by the number of tests performed. Further, in the case of angle comparisons, binary vectors were constructed for each drug to represent significant (ones) and non-significant (zeros) differences. These arrays were then represented as cluster trees using average linkage to show class-wide conservations and divergence in angular behaviour across the different PIs. Finally, the whole experiment was replicated once using a different seed.

4.3 Results and Discussions

4.3.1 Preliminary quality control of MD runs

In order to remove potential sources of additional variation, an initial 100ps of simulated data was removed from each trajectory, on the basis of higher variation in C_α-RMSD observed during that time (Figure 4.2). It is very likely a result of residual effects from temperature and pressure equilibration that were performed before the production MD runs. These summary statistics have been condensed from RMSD values plotted for each individual complex done as a quality control against the failure to remove periodic boundary conditions, which can lead to distance artefacts if uncorrected.

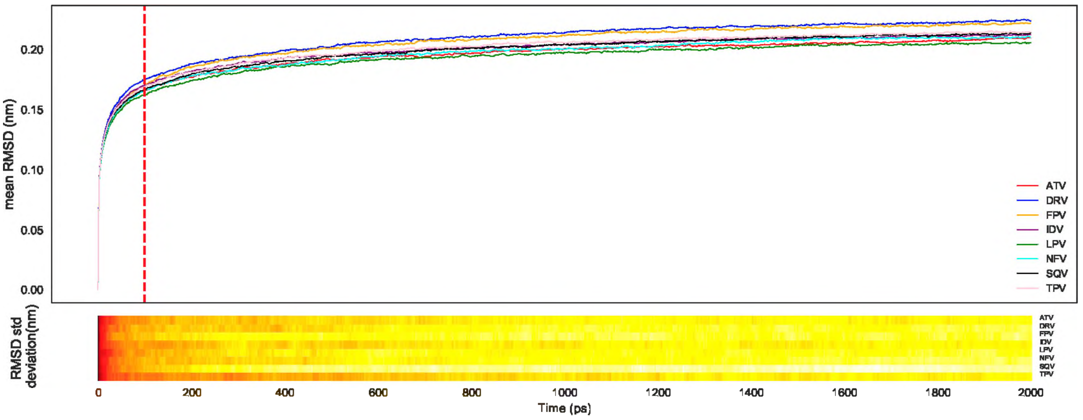
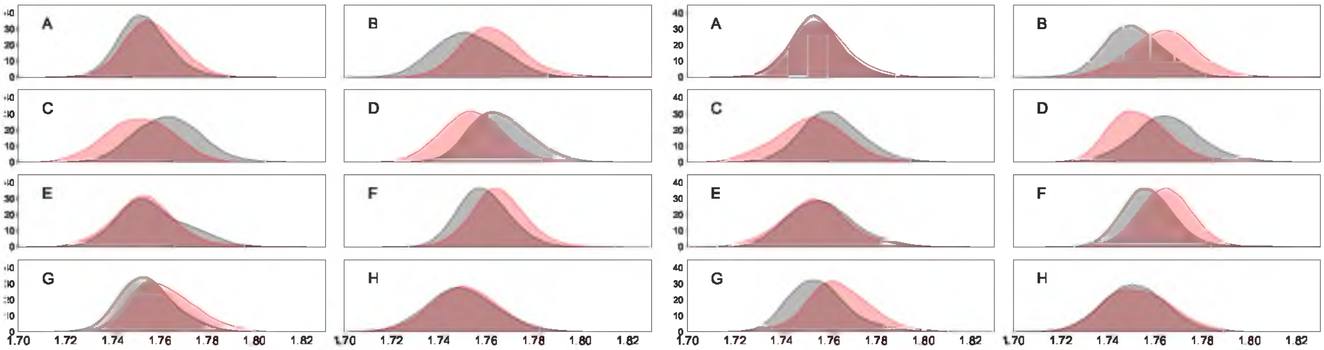


Figure 4.2: Average (top) and standard deviations (bottom) of C_α-RMSD values evaluated for all 8 FDA-approved PIs. The red line demarcates an initial 100ps equilibration region. The heat map shows the RMSD variances coloured red to white corresponding to high and low values respectively. Figure re-used from [303].

4.3.2 Using R$_g$ for global assessment of compaction

Preliminary observations were made from the analysis of global compaction by comparing distributions of R_g values obtained in each resistance ensemble for each PI sampled over MD periods of 2ns. It is known that drug resistance is associated with a wider binding cavity, however as seen in Figure 4.3a, there is no systematic skew in the distribution of Rg values generally supporting a wider cavity. More specifically, higher average R_g values (corresponding to a wider binding cavity) were clearly observed only for the drugs DRV, NFV and SQV. For the drugs ATV, LPV and TPV the distinction was subtle, while the complete opposite was observed in the case of FPV and IDV. Replication of the entire experiments with different random seeds showed similar skews and statistical modes (Figure 4.3b) in the individual distributions for each drug ensemble. The general observations obtained from analysis of this global property is that a differentiation is very weakly-detectable and appears non-conserved across PI drugs. We therefore hypothesized that a differential signal may exist at a local level but is heavily masked by more chaotic motions. For this investigation, more fine-grained analysis was done by comparing pairwise residue distances and bond angles between resistance states.



(a) Replicate 1 **(b)** Replicate 2

Figure 4.3: Distributions of R_g values for the drug-resistant (shaded in grey) and drug-susceptible ensembles (shaded in red). Each sub-figure represents the distributions for a particular protease inhibitor, namely ATV (A), DRV (B), FPV (C), IDV (D), LPV (E), NFV (F), SQV (G) and TPV (H). Figure re-used from [303].

4.3.3 Network construction using t-tests of pairwise residue distances

Initial stages of calculating pairwise distances were based on C_α atoms, but were updated to use C_β atoms (except for glycine) to represent amino acid residues with the objective of obtaining more information about side chain movement. Additionally, the MD run times were increased from 1ns to 2ns for an improved conformational sampling and to obtain more characterising information. Ideally the use of a much larger simulation time would have improved the thoroughness of sampling, but the scope of research was limited by computational time as the experiment was to be replicated for 8 drugs. On the other side, a major advantage of a short simulation time would be the observation of possible differentiating signals early, which implies being able to do reliable resistance diagnostics from infected patients. The approach described can be extremely useful in cases where subtypes are divergent from subtype B, which is commonly used in training drug resistance predictors. Such an approach also will not suffer from features of differing lengths that can be ambiguous to represent and compute using common machine learning approaches. The t-test p-values corresponding to larger and smaller distances in the resistance ensembles were used to construct adjacency matrices for each drug after performing the Bonferroni correction for family-wise error rate and is shown in Figure 4.4. While informative, the adjacency matrices are still saturated with information but already generally show some strong patterns of conservation across the ensembles, although apparently weaker in LPV and TPV. In order to examine the relationships, network graphs were constructed and degree centralities were evaluated for the larger and smaller distances separately. As a side note, a single network was initially calculated from the adjacencies for which connections were built from two-tailed tests. This method confounded information about which distances were larger and smaller, which made interpretations ambiguous, hence the calculation of one-tailed tests in each of the cases. At this stage, differential signals were still potentially noisy. Therefore we proceeded with the calculation of degree centralities hoping that the concept of preferential attachment from scale-free networks [304] would be of assistance in prioritising the differences that were consistent and most likely real. The idea relied on the fact that a residue would most probably be further away from a group of residues if the same residue was found to have statistically significant distances (greater or smaller) to multiple other residues. We find such information by ranking residues in descending order of connectivity and show our findings in Figure 4.5 where degree centralities are scaled to the range [0,1] in each case, using the following formula: $\frac{x_i - \min(x)}{\max(x) - \min(x)}$. Additionally, betweenness centralities were evaluated and overlaid on the same graph initially, but showed similar trends to the degree centrality in preliminary analyses and were therefore not considered further.

A very conserved signal was observed in all drug complexes despite the sequence variations, short simulation time and partial stochasticity of the methods used. As seen in Figures 4.5 and 4.7, the cantilever base is systematically drawn towards the catalytic core in all drug complexes in the resistant state. Not immediately apparent from the normalized degree centralities, mapping onto 3D structures clearly shows that a second conserved behaviour was detected (Figure 4.7), more specifically a lateral expansion that was also manifested early in drug resistant ensembles. It should be noted that very similar behaviours were observed from the replicated experiment, with minimal residue differences. We proceed by describing the set of observations for each PI drug.

4.3.4 Results obtained for each ARV

In the case of **ATV**, smaller distances were recorded at positions 70 and 17 on chains A and B (Figure 4.5A). Larger distances were observed at positions 36, 37 and 73 on chain A and at positions 36 and 73 on chain B. Structural mapping of these residue loci (Figure 4.7A) shows that residues displaying larger distances in the resistance ensemble favour a lateral expansion motion involving the walls of the binding cavity and the flaps elbows, which would correspond to a widened protease conformation. Residues predicted to have smaller distances display an upward motion in direction of the flaps. Additionally, those motions show a high level of symmetry with respect to each monomer. Highly similar residue behaviours were obtained upon replication, with minor differences such as residue 36 which peaked from chain A instead of B. For the resistance ensemble several mutations were present in addition to the accessory DRMs (10IVF, 32I, 33FV, 34Q, 46LI, 48V, 53L, 54LVM, 60E, 62V, 64VM, 71IV, 73STA, 90M, 93LM) and the major DRM 84V.

For **DRV** (Figures 4.5B and 4.7B), residues 71 and 72 from chain A, and residues 69, 71 and 72 from chain B move symmetrically inwards towards the active site in the resistance ensemble. On the other hand, expansion occurs at position 10 from chain A and positions 10, 21, 37 and 54 from chain B for the same ensemble. While the flap elbows did not move symmetrically, residue 10 was found to move laterally outwards from both chains A and B. Additionally, residue 54 from chain B was also found to move out perpendicularly from the bulk of the protease - a scenario that was not observed under replication (Supplementary Figure). The resistance ensemble comprised the major DRMs 47V, 54LM and 84V, while accessory DRMs consisted 11I, 32I, 33F and 89V in addition to several other variations.

In the case of **FPV** (Figures 4.5C and 4.7C), contractions were observed at residue position 71 on chain A and at positions 69-72 from chain B. On the other hand, expansions were observed along residues 15-17 on chain A, and positions 16 and 73 from chain B. Lateral expansion occurred as observed in the corresponding DRV ensemble, but here same is achieved via the 10's region and the surface-exposed residue 73, under replication. Protein contraction also occurs reproducibly in the replicate run. In addition to other mutations, the resistance ensemble comprises major DRM 84V and accessory DRMs 10IVF, 32I, 46LI, 47V, 54LVM, 73S, 76V, 82TA and 90M.

In **IDV** (Figures 4.5D and 4.7D), contractions for the resistance ensemble were observed at positions 63, 69-71 on chain A and at residue 71 on chain B. Expansions were observed at positions 16 and 73 on chain A, and at positions 16, 17 and 93 on chain B. Residues involved in expansion were identical upon replication, while inward motion was reproducible only for the cantilever loop along chain A. Cantilever residue 73 is yet again found to be implicated in lateral expansion, this time from both chains A and B. An identical response was obtained from the replicate runs for expansion along the 10's region from both chains while compaction was mediated through residues 69-71 on chain A. The resistance ensemble in this case includes, in addition to various variations, the major DRMs 46LI, 82FTA and 84V, and the accessory DRMs 10IV, 20R, 32I, 36I, 54V, 71TV, 73SA, 76V, 77I and 90M.

In the case of **LPV** (Figures 4.5E and 4.7E), contractions occurred via residue positions 70 and 71 on chain A, and residues 69-71 on chain B in the resistance ensemble. Expansion was mediated via positions 73 and 93 from chain A and residues 34, 36 and 73 from chain B. A high concordance for the residues involved in expansion was obtained upon replication, with the exception of residues 36 and 81 from chain B, which showed similar patterns despite ranking differently. Similarly compaction was mediated via residue 69 instead of 71. The resistance ensemble here comprises the major DRMs 32I, 47VA, 76V and 82SFTA while the accessory ones consist of 10IFV, 20RM, 24I, 33F, 46LI, 50V, 53L, 54LTVMS, 63P, 71VT, 73S, 84V and 90M. Additional variations are also present.

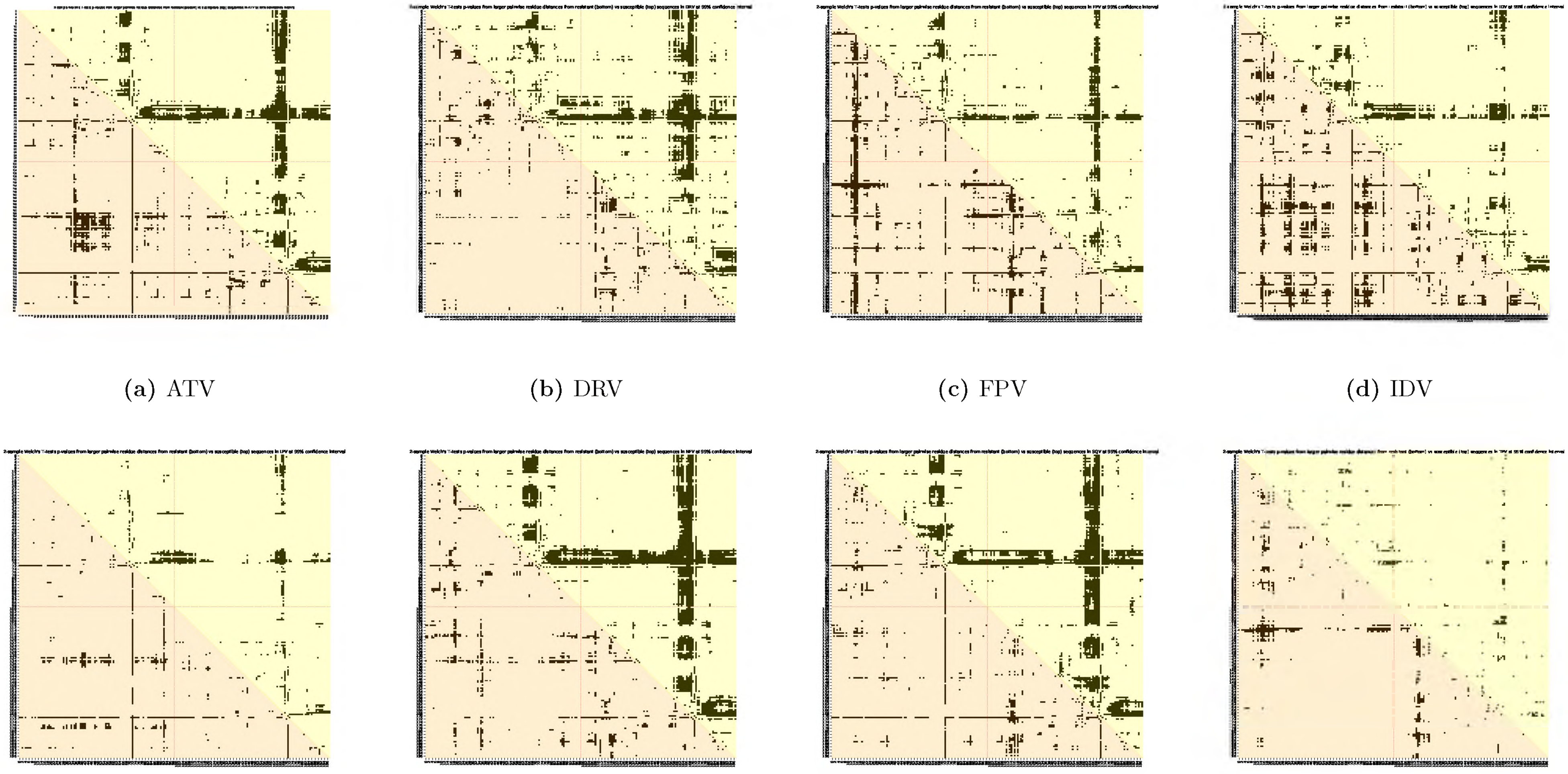
For the **NFV** resistance ensemble (Figures 4.5F and 4.7F), compaction occurred via residues 69-71 from chain A and via residues 70 and 71 from chain B. On the other hand, expansion was mediated via residues 20 and 36 from chain A and residues 20, 36 and 73 from chain B. Replication lead to very similar behaviours, however residues 20 and 36, which form part of the elbow and fulcrum respectively, appeared to move in a coordinated manner during expansion. Protease contraction is mediated similarly to previous drugs, proximal to the cantilever loop. The resistance ensemble for NFV comprised major DRMs 30N and 90M, and accessory DRMs 10IF, 36I, 46LI, 71TV, 77I, 82FA, 84V and 88D.

For **SQV** (Figures 4.5G and 4.7G), resistance-associated contractions were observed via residues 70 and 71 on chain A and residues 69, 71 and 72 from chain B. Expansion was observed via residues 73 and 89 from chain A and on chain B for residues 18, 20 and 73. Replication produced similar, though not completely identical behaviours, with a very symmetric compaction at the fulcrum area within both chains while displaying very similar expansion points. The resistance ensemble for SQV comprised major DRMs 48V and 90M, while the accessory DRMs consisted of 10I, 54LV, 62V, 71TV, 73S, 77I, 82A and 84V, amongst other mutations.

In **TPV** (Figures 4.5H and 4.7H), contractions were experienced at positions 33, 60 and 71 from chain A, and at positions 70 and 71 from chain B. Expansions occurred via residues 16 and 20 from chain A and at positions 15-17 from chain B. Upon replication a similar profile was obtained. It should be noted that contraction, while occurring proximal to the cantilever loop, mobilises the buried residue 33 from chain A - a property not observed in the other ARV complexes. The resistance ensemble for TPV comprised major DRMs 47V, 58E, 74P, 82LT, 83D, 84V and the accessory DRMs 10V, 33F, 36IV, 43T, 46L, 54VM and 89VM, amongst other mutations.

4.3.5 Monitoring angular backbone behaviour from C_α atoms

In addition to distances, it was presumed that monitoring bond rotations would further help characterise movement within the ensembles associated to resistance or susceptibility. With the aim of finding an absolutely conserved angular behaviour characteristic of drug response, $C_\alpha - C_\alpha$ angles were compared across drug resistance ensembles and represented as a cluster heat map. From Figure 4.8 and its replicate, it can be seen that position 84 is very likely larger across all drugs in the resistance state, however this is supported by only one of the replicates. TPV appears to elicit a very different response in the resistance state, very likely due to its non-reliance on interfacial water for stabilisation within the active site. Overall the angular behaviours are not highly conserved as seen from clusterings for replicate runs of both smaller (Figure 4.9) and larger (Figure 4.8) angles. The absence of such conservation may suggest that multiple angular behaviours may lead to a similar resistance effect, however an increased conformational sampling would add more evidence to this hypothesis, but would require very high computational resources unless a cheaper sampling method is available.



(a) ATV

(b) DRV

(c) FPV

(d) IDV

(e) LPV

(f) NFV

(g) SQV

(h) TPV

Figure 4.4: Overview of the adjacency matrices obtained from one-tailed t-test p-values for the distances being larger and smaller (shaded yellow and orange respectively) in the resistance ensemble for each of the 8 PI drugs. The dotted red lines represent the boundaries between chains A and B in each case.

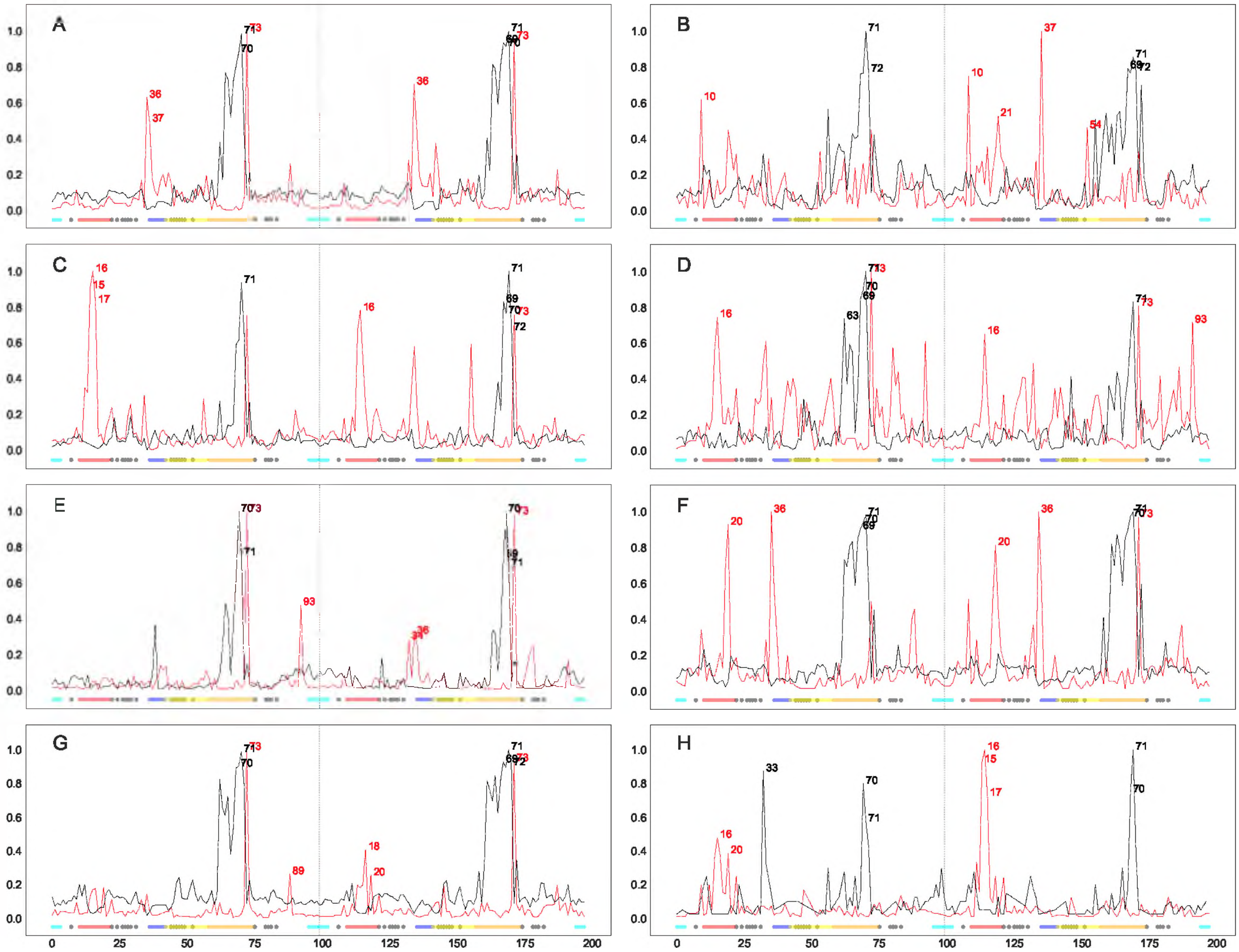


Figure 4.5: Normalized degree centralities of significantly larger **(red lines)** and smaller **(black lines)** distances observed in resistant ensembles for 8 FDA-approved protease inhibitor complexes, namely ATV **(A)**, DRV **(B)**, FPV **(C)**, IDV **(D)**, LPV **(E)**, NFV **(F)**, SQV **(G)** and TPV **(H)**. The top 5 residue positions with the highest connectivities are labelled at the peaks in each graph. Inserted underneath are the functional protease residues depicted as coloured dots, namely the fulcrum (red), the elbow (blue), the flap y (yellow), the cantilever (orange), the interface (cyan) and the binding cavity residues (grey). Figure re-used from [303].

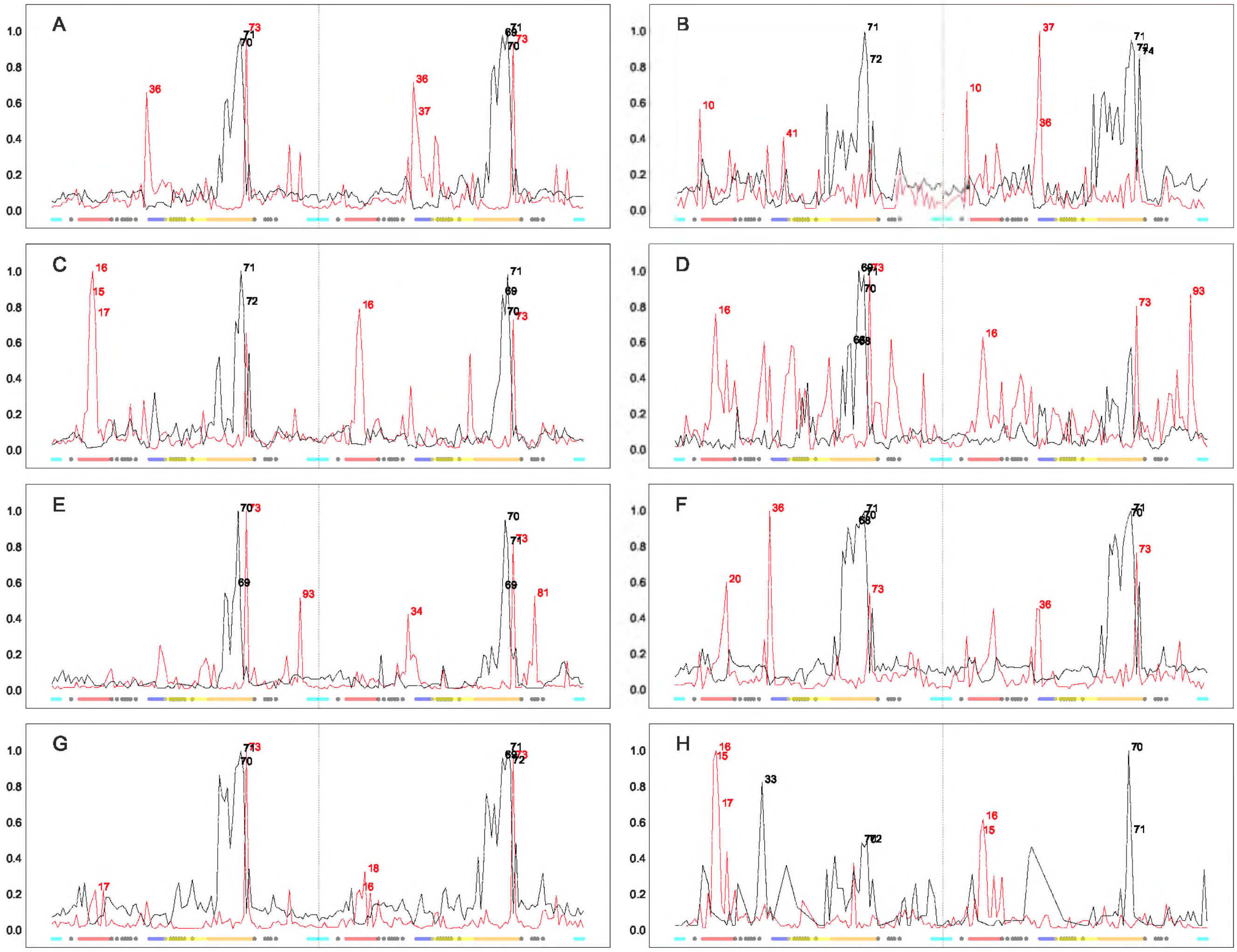


Figure 4.6: Replicate of the normalized degree centralities of significantly larger **(red lines)** and smaller **(black lines)** distances observed in resistant ensembles for 8 FDA-approved protease inhibitor complexes, namely ATV **(A)**, DRV **(B)**, FPV **(C)**, IDV **(D)**, LPV **(E)**, NFV **(F)**, SQV **(G)** and TPV **(H)**. The top 5 residue positions with the highest connectivities are labelled at the peaks in each graph. Inserted underneath are the functional protease residues depicted as coloured dots, namely the fulcrum (red), the elbow (blue), the flap y (yellow), the cantilever (orange), the interface (cyan) and the binding cavity residues (grey). Figure re-used from [303].

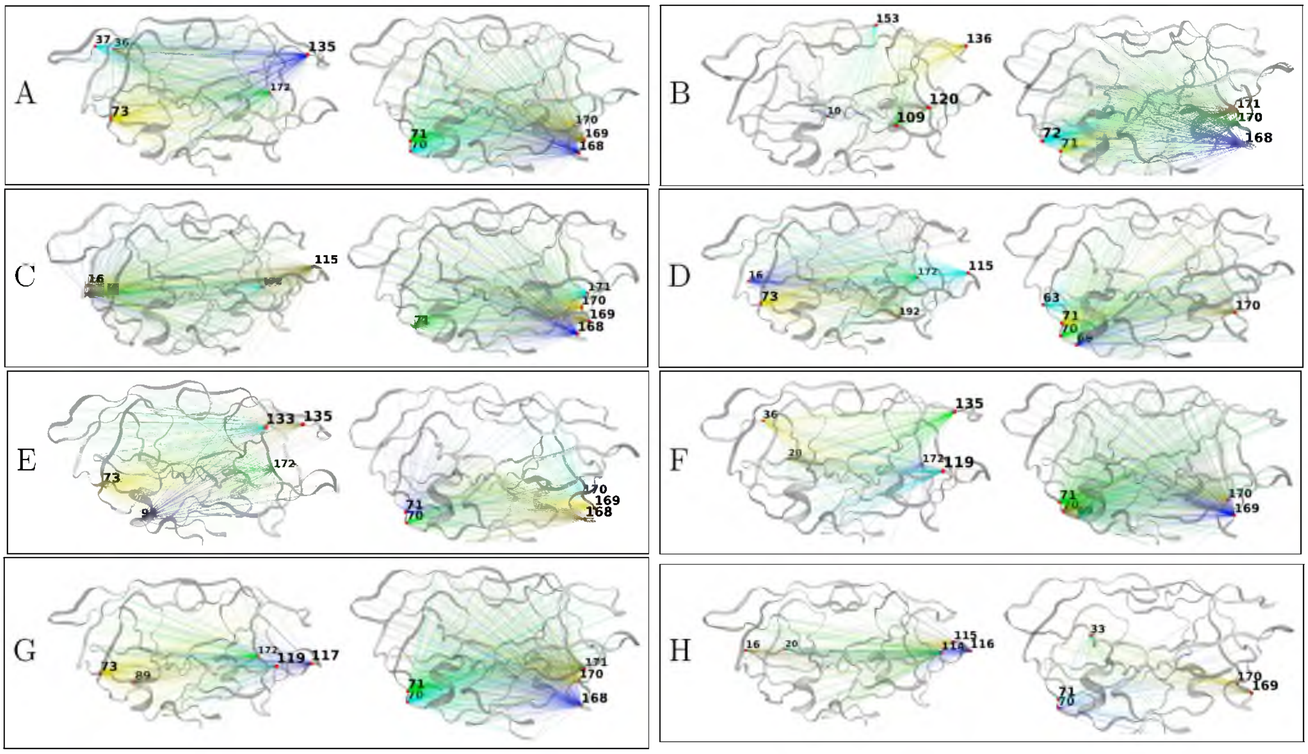
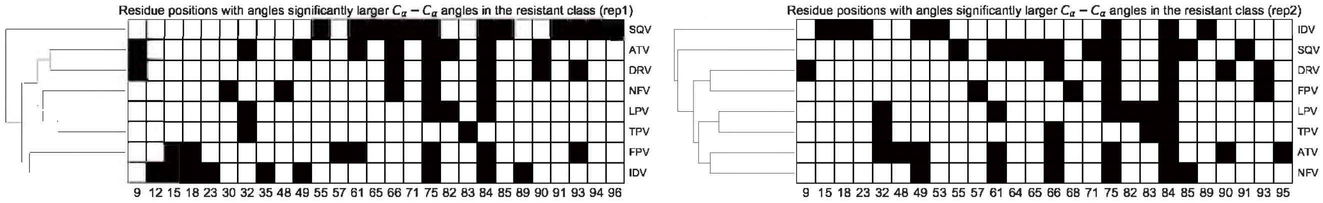
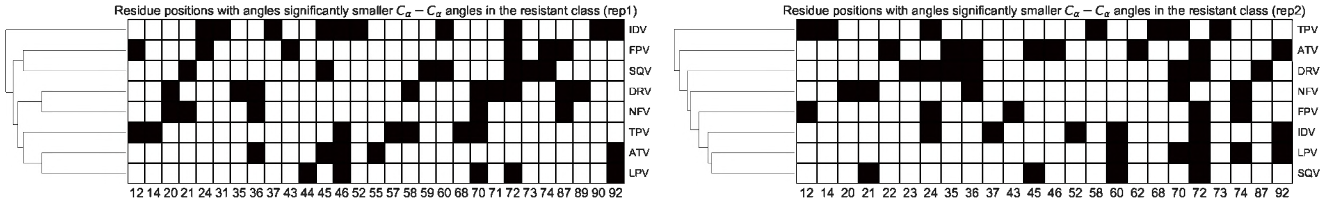


Figure 4.7: Top-ranked degree centralities mapped onto HIV protease structures for significantly larger (left) and smaller (right) distances observed in resistant ensembles for complexes containing ATV **(A)**, DRV **(B)**, FPV **(C)**, IDV **(D)**, LPV **(E)**, NFV **(F)**, SQV **(G)** and TPV **(H)**. Figure re-used from [303].



(a) Replicate 1 **(b)** Replicate 2

Figure 4.8: Heat map of residue positions with significantly larger C_α angles in the resistant ensemble for each PI. The hierarchical cluster tree is displayed on the left. The first replicate is at the left and the second replicate is at the right. Figure re-used from [303].



(a) Replicate 1 **(b)** Replicate 2

Figure 4.9: Heat map of residue positions with significantly smaller C_α angles in the resistant ensemble for each PI. The hierarchical cluster tree is displayed on the left. The first replicate is at the left and the second replicate is at the right. Figure re-used from [303].

4.4 Conclusions

In this work, we use degree centrality calculated from graphs built using statistically-moderated edges across ensembles of PI-bound protease complexes expressing resistance and susceptibility. By increasing the sensitivity of signal detection in the comparison of geometric differences present in ensembles of HIV protease sequences, we were able to highlight clear and statistically-significant differential residue motions occurring very early in cases of resistance and susceptibility for 8 FDA-approved ARV drugs. Despite being performed only on subtype B HIV, this approach can be of tremendous value in non-B subtypes, where predictions are currently less optimal. Given the level of stringency used and conservation observed in this experiment, it is not impossible that same or similar signals may extend to early resistance-related events in other subtypes. The brief simulation time means that similar designs may be trialled on consumer grade computers to supplement extant drug resistance algorithms in prescribing treatment. This level of conservation is novel and features a promising generic approach for the sensitive analysis of variation associated with differing phenotypes. More generally, such an approach can be repurposed for the extraction of dissimilarities between homologous systems or for feature augmentation in order to improve machine learning predictions of drug resistance. The experimental design inherits from requirements for performing the t-tests, that is generally 30 samples or more of each class for improved performance.

Chapter 5

Screening for potential protease inhibitors

5.1 Introduction

In this section we shift gears from the improvement of resistance prediction to use high-throughput virtual screening (HTVS) to find potential ARV scaffolds with activity against darunavir-resistant protease sequences. According to a UNAIDS report [305], an estimated 36.7 million people were infected with HIV worldwide in 2016 at a rate of 1.8 million new infections for the same year. There is no current cure for the disease and clinicians assist patients in maintaining virological suppression by prescribing Highly Active Antiretroviral Therapy (HAART) regimens [306], which target distinct key enzymes involved in HIV replication via drug combinations. There have been several cases where rational drug design has expedited the discovery of novel active compounds. Some examples include the development of a renin inhibitor aliskiren, which is used in the treatment of primary hypertension [307]; catechol diether derivatives potent against HIV RT highly resistant to rilpivirine and efavirenz [308, 309]; and in the repositioning of a non-steroidal anti-inflammatory drug celecoxib as a STAT3 protein inhibitor to be used as an anti-cancer drug [310]. The rational design of the first protease inhibitor saquinavir using computational methods has been decisive in substantially reducing the rate of HIV-related deaths since its incorporation as part of therapy in 1995 [19]. Since then with the exception of tipranavir, several PIs have been developed using the concept of peptide mimicry based on hydroxyethylamine isosteres [311]. However, because most of the current PIs share this substructure or rely on similar ligand-contacting protease residues [312], drug pressure exerted via treatment continuously selects for DRMs via mutation stacking, which presents a main challenge for current PI designs which have been used for decades [90]. HIV's tenacity resides in its ability to mutate [11] and replicate quickly [313], its capacity for accumulating mutations [314] and its existence as a quasispecies. Integration within the host genome marks permanent establishment of the retrovirus in HIV patients and for this reason timely use of effective ARVs is critical. DRMs harboured by HIV can either establish direct ARV contacts within the active site or act from a distance, in which case the DRMs are referred as major and accessory respectively [315]. The latter are individually less impactful but can act in concert to enhance drug resistance. For the reasons aforementioned, resistance against multiple members of a given drug class is not uncommon. After the previous gold standard lopinavir (LPV) [316], the latest FDA-approved protease inhibitor DRV is currently recommended by the World Health Organisation in combination with ritonavir for use in first line regimens as it is better-tolerated by patients and has a higher barrier for viral resistance. We use HTVS with a strong focus on finding drugs that will potentially be the next in line when patients fail DRV treatment. In this respect, the screen is performed to find new potential scaffolds targeted against a panel of DRV-resistant protease sequences using compounds from the South African natural compound database (SANCDB) [1]. Nine promising compounds are identified from the SANCDB database with good stabilities and hydrogen bonding capacities with their respective receptor complexes observed over periods of 20ns simulated by molecular dynamics. Modifications are also proposed to one of the compounds for improved efficacy.

5.2 Methods

5.2.1 Sequence retrieval and receptor preparation

The pre-filtered PhenoSense assay dataset was retrieved from Stanford HIVdb [83]. Each sequence entry was reconstituted by expansion using the Cartesian product of the tab-separated sequence residue positions. Using a DRV fold resistance cut-off ratio of 90, a total of 274 99-residue long were obtained and converted to FASTA format, with an identifier composed of the original sequence ID accompanied by an additional number corresponding to the variant rank after expansion. Following visual inspection of the multiple sequence alignment obtained from MUSCLE (version 3.8.31) [317], no further filtering was deemed necessary, as there were no indel mutations. In order to increase our chances of finding a drug of wide spectrum against the quasi-species, a series of divergent sequences was prepared. We proceed by building a hierarchical cluster tree from the aligned sequences and cutting it to obtain 10 sequence clusters. From each partition, one sequence was randomly picked. The selected sequences were aligned as depicted in Figure 5.2 to show the location of sampled non-synonymous mutations.

5.2.2 Receptor preparation for docking

In addition to sequence variance, different receptor conformations are also considered. The opened (PDB ID: 4YOA) and closed (PDB ID: 3UCB) conformation crystal structures were thus retrieved on the basis of (1) high-resolution (1.70 Å and 1.38 Å respectively), (2) the presence of co-crystallised DRV and (3) the absence of missing residues. The crystallized template protein structures were separated from any non-protein molecule and residue rotamers were removed using an in-house Python script so that protonation could be later performed. The rotamer removal algorithm was designed to retain higher-occupancy side chains. In the case of ties, only the last rotamer was kept. Complexed ligands were extracted and retained as separate files for docking preparations. MODELLER (version 9.16) [186] was used for homology modelling each sequence - 100 models were built under slow refinement for each of the 10 sequences in both conformations, resulting in 200 models, that were individually ranked in ascending order of their z-DOPE score to obtain 20 high-quality models (See Table 5.1).

5.2.3 Docking validation (with DRV) and virtual screening

Prior to ligand screening, a validation was performed to determine optimal experimental parameters for docking accuracy and speed. Apo-receptors were protonated to pH 7 with PDB2PQR (version 2.1.0) [272] using the PROPKA algorithm to produce 3D models with AMBER residue types. Based on results obtained from preliminary docking validations, co-crystallized water present at the interface of the protease flaps and ligand was retained for the closed conformation template (and homology models) as they improved the ligand binding pose. The open conformation template was aligned to the closed receptor conformation using PyMOL (version 1.7.2.1) to define a ligand docking centre. The C_{17} carbon atom of DRV from 3UCB PDB file was used as the docking centre with coordinates (13.455, -0.862, -10.106). AutoDockTools [273] was used to determine the grid box size (22 x 22 x 22 Å^3) to entirely surround the binding cavity around the docking centre. Extracted crystallized ligands were protonated using the *generic organic* algorithm and converted to PDB format using VEGA (version 3.1.1) [276] before merging non-polar hydrogen atoms, calculating partial charges and assigning rotatable bonds using the AutoDockTools *prepare_ligand4.py* script. These were then re-docked by flexible ligand docking against their respective prepared apo-receptor templates using AutoDock Vina (version 1.1.2) [200]. Once docking parameters were obtained, same were applied to screen all SANCDB ligands, finally summarizing the binding energies as an un-normalized and a quantile normalized (across compounds for each receptor) data set. In both cases, ligands were ranked in ascending order of average binding energy across receptor variants and conformations.

5.2.4 Assessment of hit compound stability

Each of the complexed ligands was subsequently monitored for their stability within the protease receptor variants via all-atom MD simulations using GROMACS (version 2016.4) [263] with the AMBER03 forcefield. Docked ligands were protonated to their original form and converted from PDBQT to PDB format using VEGA. Topologies were then obtained from the previously protonated receptors using the GROMACS *pdb2gmx* command while ACPYPE [264] was used for protonated ligands. The receptor-ligand complexes were solvated with SPC-modelled water in 0.15 M NaCl in a triclinic periodic box with a clearance of 1 nm from the protein. A cut-off distance of identical magnitude was used for short-range non-bonded interactions while PME was used for handling long distance charged interactions. After energy-minimization using the algorithm of steepest-descent with a gradient minimum of 1000 kcal/mol for a maximum of 50000 steps, the system underwent temperature and pressure equilibration for 50 ps in each case using time-steps of 2 fs. The LINCS algorithm was used to constrain bond lengths after unconstrained updates for both equilibration and the 20 ns production runs. All simulations were performed on the large queue of the Centre for High Performance Computing (CHPC) cluster using GNU Parallel (version 20160422) [274] with a 24-core node for each independent run. Hit compound stability was initially monitored by all-atom ligand RMSD with respect to the initial frame, but was replaced by the more sensitive calculation of the euclidean distance between the protein and ligand COM values for each simulation frame. Additionally, hydrogen bonding propensity between the receptors and ligands was monitored.

5.2.5 Hit-ligand modification

One of the hit SANCDB compounds (SANC00178) with very high receptor binding affinities across opened conformations, but poor performance in only one closed conformation variant was modified to improve flexibility. Modifications involved deletion of two carbon-carbon single bonds connecting cycloalkane derivatives around a central pyrazine group followed by filling the open valences with double bonds to produce a 2,5-dimethylpyrazine derivative (See Figure 5.9). The structure was minimized by conjugate gradient until convergence using the Avogadro software [318]. After docking, the ligand protonation state was restored using the Discovery Studio Visualizer (version 4.1), its topology was prepared and dynamics were monitored as described above. Stability of receptor-ligand interactions are inferred from the contact strengths estimated from the time-averaging of intermolecular contacts between ligand and protein heavy atoms around a radius of 4 Å from MD-sampled frames.

5.3 Results and Discussions

Here we present the ligand screening results, starting from receptor preparations, going through high-throughput ligand screening until the assessment of hit compound stability and the modification of a promising hit compound.

5.3.1 Characteristics of the protease receptors

With focus on DRV-resistance the number of sequences to be evaluated for drug screening was set at 10 to incorporate multiple cavity variations associated with resistance. The aim being to increase the probability of obtaining efficient binders by mitigating docking-related problems while still maintaining reasonable computational costs. Top-most divergent sequences were estimated using hierarchical clustering on the complete subset of DRV-resistant sequences retrieved from Stanford HIVdb. Cluster representatives were chosen by initially trimming the dendrogram to obtain 10 clusters, before randomly selecting a sequence from each cluster. We report that a previously trialled algorithm based on ranking of all sequences by average distance before selecting top ones does not work correctly as it accumulates sequences which are equally divergent to the remaining subset of sequences. In effect, such an approach led to a subset of similar sequences for such a small set of 10 sequence variants and was thus replaced by the partly stochastic approach described earlier in this subsection. Figure 5.1 depicts the relationships estimated between the sequences whereby distances were calculated using similarity while the alignment of cluster representatives is shown in Figure 5.2. Further, we show the location of sampled non-synonymous mutations with respect to the HIV reference protease by including the consensus B subtype in the multiple sequence alignment (Figure 5.2). Additionally, secondary structural elements have

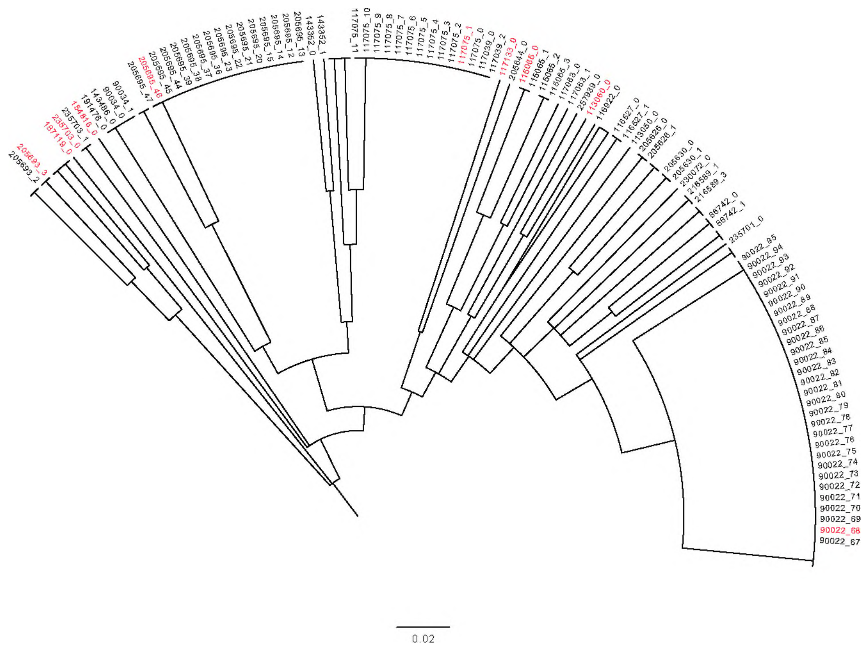


Figure 5.1: Hierarchical cluster tree based on the distance matrix obtained using the Fitch similarity metric from aligned protein sequences. The 10 selected sequences (sequence IDs: 90022_68, 205695_46, 117075_1, 113060_0, 115065_0, 117133_0, 154816_0, 187119_0, 205693_3, 235703_0) are labelled in red. Several clades were collapsed in order to improve label visibility.

been overlaid to show the positions of major DRMs. From the same figure, it can be seen that the mutations occur at multiple loci along the selected proteases, targeting both residues from the substrate-binding cavity and important secondary structural elements (β-hairpins) comprising the flap, fulcrum and cantilever regions. Sampled cavity residue mutations occurring at positions 46, 47, 82 and 84 are all involved in varying levels of multi-drug resistance, while the impacted β-hairpins act in concert to lead in flap opening [299]. By evaluating ligand performance against such a panel of slightly altered receptors, we attempted to mitigate chance events resulting from the partly stochastic sampling of the ligand search space by increasing the number of independent docking trials for each ligand. In effect, a hit compound would have to rank highly against most receptors for it to be considered further by more accurate but computationally-demanding methods. We improve the design further by also considering both opened and closed receptor conformations by using separate templates as described in the methods section. An array of 20 models was thus obtained for screening the drugs after performing independent quality evaluations using both QMEAN4 (from the SWISS-MODEL web server) and z-DOPE from MODELLER. As seen in Table 5.1, high quality models were obtained in all cases, with values very close to or below -1 in the case of z-DOPE, thus indicating native-like states [278]. Similarly, QMEAN4 evaluations indicated good model quality with values approaching 1. Having selected the sequences, they were brought forward for ligand screening.

5.3.2 High-throughput virtual screening

Before proceeding to HTVS, a control experiment was performed to validate docking parameters, which mainly comprised the grid size and centre in addition to the exhaustiveness. These are per-

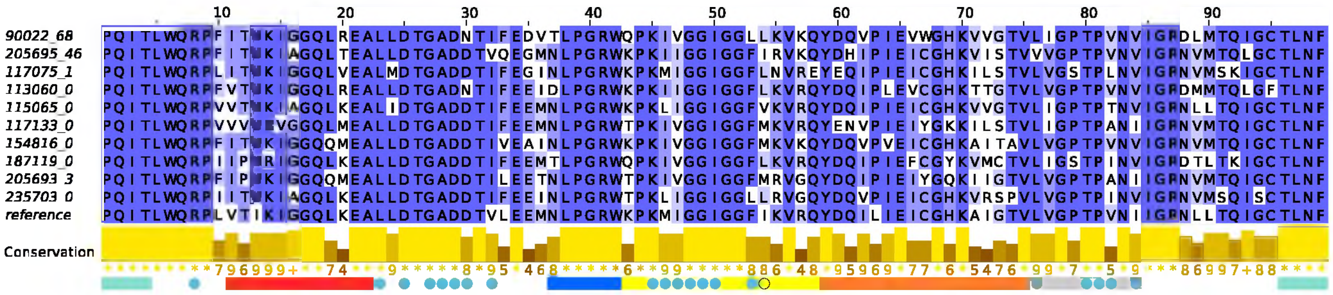


Figure 5.2: Multiple sequence alignment of the sequences used for drug screening showing the degree of residue conservation with respect to the subtype B consensus sequence, labelled as reference. Structural features have been added underneath the alignment as coloured strips, namely the protease dimer interface (cyan), the fulcrum (red), elbow (blue), flap (yellow), cantilever (orange) and the 80S loop (grey). Binding cavity residues are depicted as filled blue circles while black circular outlines denote major DRV-resistance positions.

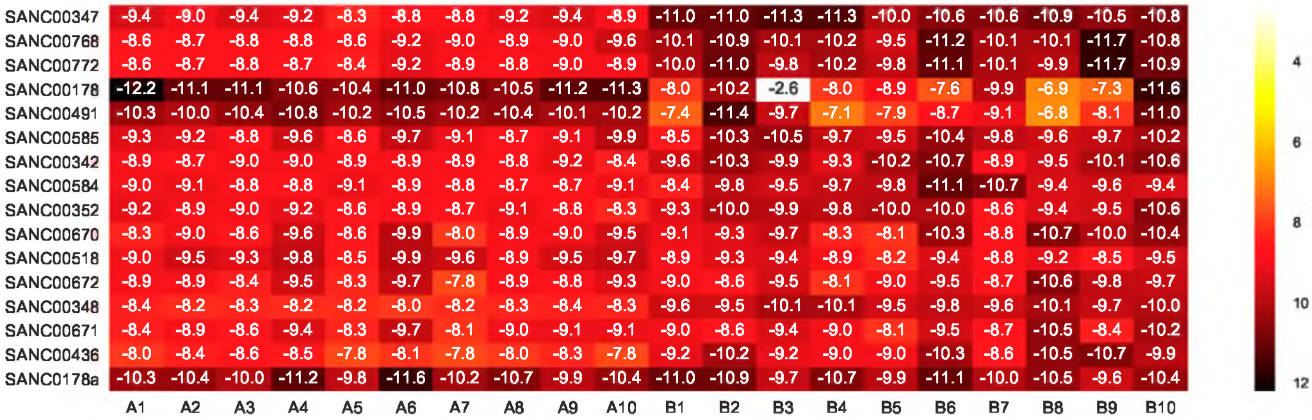
formed to assessing how well our approach reproduces experimentally-determined ligand poses. The correctness of docking parameters was estimated by calculating the heavy-atom RMSD between the docked DRV and its originally co-crystallized position within each template receptor. RMSD values of 0.9 Å and 8.6 Å were obtained for the closed and opened conformation templates respectively. A lack of ligand-stabilizing contacts in the opened receptor conformation was the main reason for not being able to reproduce closely the original pose. However, the positioning was close enough as estimated by a centre-of-mass distance of 3.7 Å with respect to the original pose. Ligand binding energies were estimated at -6.4kcal/mol and -8.7kcal/mol for the opened and closed template receptor conformations using AutoDock Vina. Of particular importance is an interfacial water molecule which was considered for screening in the case of the closed receptor conformations as it improved both the ligand binding energy and pose. Same could not be performed for the opened conformation as no crystallised water was anchoring DRV to the flaps. Subsequently, all compounds were retrieved from SANCDB for the screening experiment to make a total of 718 ligands. The docking jobs were performed on the CHPC large queue with 12 cores each. Screening results were then organized into an array with ligands along the row and receptors along the columns. Average binding energies were calculated for each ligand to rank the dataset in increasing order of average binding energy. The top-most ligands are shown in Figure 5.3 while details of their origin and recorded uses are given in Table 5.2. SANC0178a is a modification of compound SANC00178 based on observations made from dynamic simulations and is later discussed. From the heat map, it can be seen that all hit compounds displayed improved receptor affinities when compared to the re-docked DRV with the exception of a closed conformation receptor B3 in the case of SANC00178, despite a very good overall performance in opened conformations. Compound SANC00347 performed the best irrespective of receptor variation and conformation, with higher performances displayed towards the closed conformations. Co-incidentally most of the top hits comprised molecules from a select few compound classes each sharing similar structures, some being shown in Figure 5.8. The next step was to further narrow down the top hits by evaluating their stability under physiological conditions by molecular dynamics simulations.

Table 5.1: Sequence identifiers and model quality scores for the selected protease receptors

Labels used	Sequence IDs	Protease conformation	z-DOPE scores	QMEAN4 values
A1	113060_0	opened	-1.261	0.850
A2	115065_0	opened	-1.355	0.837
A3	117075_1	opened	-1.269	0.834
A4	117133_0	opened	-0.989	0.811
A5	154816_0	opened	-1.053	0.839
A6	187119_0	opened	-1.101	0.837
A7	205693_3	opened	-1.081	0.814
A8	205695_46	opened	-0.990	0.864
A9	235703_0	opened	-1.233	0.822
A10	90022_68	opened	-1.331	0.833
B1	113060_0	closed	-1.521	0.823
B2	115065_0	closed	-1.561	0.823
B3	117075_1	closed	-1.490	0.779
B4	117133_0	closed	-1.222	0.811
B5	154816_0	closed	-1.350	0.795
B6	187119_0	closed	-1.363	0.810
B7	205693_3	closed	-1.334	0.821
B8	205695_46	closed	-1.209	0.811
B9	235703_0	closed	-1.488	0.787
B10	90022_68	closed	-1.551	0.812

5.3.3 Stability of hit compounds within their receptor complexes

Hit compound stabilities were evaluated by estimating their tendency to be retained within receptors over time given their initially docked positions. For this matter, all-atom MD simulations were performed in explicit water with 0.15 M salt at a pH adjusted to 7 for the protease in a system pre-equilibrated at physiological temperature and pressure. Receptor protonation with PDB2PQR constantly resulted in only one catalytic aspartate being fully-protonated for the dimer. All ligand open valences were filled with hydrogen atoms using VEGA before preparing the topologies for use in MD. Of particular note is the failure of some methods (such as Open Babel or other algorithms from VEGA) to produce the right protonation states. Ligand RMSD values (Figure 5.4) were first evaluated but were observed not to give an accurate reflection of ligand stability when assessed visually, especially for longer and flexible hit compounds. As a possible explanation using compounds SANC00670-672 as example, the centre of mass (COM) is subtly changed with respect to that of the receptor, but high flexibility of the long chains inflates RMSD values to mask the overall lack of COM movement. The euclidean distance was instead calculated for the position vector of the ligand COM with respect to that of the protease receptor at each time point (Figure 5.5) and was observed to be more in line with ligand stability within the active site. A summary of the COM distances is given as a series of box plots for the time-averaged



	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
SANC00347	-9.4	-9.0	-9.4	-9.2	-8.3	-8.8	-8.8	-9.2	-9.4	-8.9	-11.0	-11.0	-11.3	-11.3	-10.0	-10.6	-10.6	-10.9	-10.5	-10.8
SANC00768	-8.6	-8.7	-8.8	-8.8	-8.6	-9.2	-9.0	-8.9	-9.0	-9.6	-10.1	-10.9	-10.1	-10.2	-9.5	-11.2	-10.1	-10.1	-11.7	-10.8
SANC00772	-8.6	-8.7	-8.8	-8.7	-8.4	-9.2	-8.9	-8.8	-9.0	-8.9	-10.0	-11.0	-9.8	-10.2	-9.8	-11.1	-10.1	-9.9	-11.7	-10.9
SANC00178	-12.2	-11.1	-11.1	-10.6	-10.4	-11.0	-10.8	-10.5	-11.2	-11.3	-8.0	-10.2	-2.6	-8.0	-8.9	-7.6	-9.9	-6.9	-7.3	-11.6
SANC00491	-10.3	-10.0	-10.4	-10.8	-10.2	-10.5	-10.2	-10.4	-10.1	-10.2	-7.4	-11.4	-9.7	-7.1	-7.9	-8.7	-9.1	-6.8	-8.1	-11.0
SANC00585	-9.3	-9.2	-8.8	-9.6	-8.6	-9.7	-9.1	-8.7	-9.1	-9.9	-8.5	-10.3	-10.5	-9.7	-9.5	-10.4	-9.8	-9.6	-9.7	-10.2
SANC00342	-8.9	-8.7	-9.0	-9.0	-8.9	-8.9	-8.9	-8.8	-9.2	-8.4	-9.6	-10.3	-9.9	-9.3	-10.2	-10.7	-8.9	-9.5	-10.1	-10.6
SANC00584	-9.0	-9.1	-8.8	-8.8	-9.1	-8.9	-8.8	-8.7	-8.7	-9.1	-8.4	-9.8	-9.5	-9.7	-9.8	-11.1	-10.7	-9.4	-9.6	-9.4
SANC00352	-9.2	-8.9	-9.0	-9.2	-8.6	-8.9	-8.7	-9.1	-8.8	-8.3	-9.3	-10.0	-9.9	-9.8	-10.0	-10.0	-8.6	-9.4	-9.5	-10.6
SANC00670	-8.3	-9.0	-8.6	-9.6	-8.6	-9.9	-8.0	-8.9	-9.0	-9.5	-9.1	-9.3	-9.7	-8.3	-8.1	-10.3	-8.8	-10.7	-10.0	-10.4
SANC00518	-9.0	-9.5	-9.3	-9.8	-8.5	-9.9	-9.6	-8.9	-9.5	-9.7	-8.9	-9.3	-9.4	-8.9	-8.2	-9.4	-8.8	-9.2	-8.5	-9.5
SANC00672	-8.9	-8.9	-8.4	-9.5	-8.3	-9.7	-7.8	-8.9	-8.8	-9.3	-9.0	-8.6	-9.5	-8.1	-9.0	-9.5	-8.7	-10.6	-9.8	-9.7
SANC00348	-8.4	-8.2	-8.3	-8.2	-8.2	-8.0	-8.2	-8.3	-8.4	-8.3	-9.6	-9.5	-10.1	-10.1	-9.5	-9.8	-9.6	-10.1	-9.7	-10.0
SANC00671	-8.4	-8.9	-8.6	-9.4	-8.3	-9.7	-8.1	-9.0	-9.1	-9.1	-9.0	-8.6	-9.4	-9.0	-8.1	-9.5	-8.7	-10.5	-8.4	-10.2
SANC00436	-8.0	-8.4	-8.6	-8.5	-7.8	-8.1	-7.8	-8.0	-8.3	-7.8	-9.2	-10.2	-9.2	-9.0	-9.0	-10.3	-8.6	-10.5	-10.7	-9.9
SANC0178a	-10.3	-10.4	-10.0	-11.2	-9.8	-11.6	-10.2	-10.7	-9.9	-10.4	-11.0	-10.9	-9.7	-10.7	-9.9	-11.1	-10.0	-10.5	-9.6	-10.4

Figure 5.3: Binding energies for 15 top binders across sequence variations and receptor conformations. Open and closed conformations are labelled A1-10 and B1-10 respectively.

COM distances for each complex (Figure 5.6), which facilitates interpretation of individual ligand performances across arrays of receptor conformations. From the distributions of COM distances in Figure 5.6, one can more easily spot the least and most stable compounds SANC00584 and SANC00672 respectively, by the height of their upper whiskers, which have been set to be the range in this case. SANC00672 also displayed a small median and a narrow range for the upper and lower quartiles. Compounds SANC00491 and SANC00772 fared less well in this respect, and together with SANC00584 hint at the possibility of ligand exit. Due to the fact that COM distance carries information from the receptor, we predict that the method will behave as expected when the receptor centre of mass does not shift largely - additional equilibrium conformations of the receptor sampled at more distal wells in the energy landscape will negatively affect this value. An appropriate correction then would be to restrict the receptor COM calculation to a smaller area centered around the ligand being investigated. After this preliminary analysis, the trajectories were assessed visually and we proceed by discussing the compounds in a class-wise manner.

Main classes for the hit molecules include the cephalostatins, kraussianones, scutiaquinones, saunderosides and marchantins, in addition to the compounds mamegakinone and 20(29)-lupene-3β-isoferulate (Table 5.2). Cephalostatins SANC00178 and SANC00491 were less flexible owing to their linear arrangement of fused heterocycles (Figure 5.8) but were generally stabilized by a centrally-located pyrazine ring in the initial receptor conformations. However, in one simulation with the opened conformation protease variant A7, SANC00471 was observed to force open one of the flaps whilst remaining strongly-bound to other flap over the course of simulation. Overall, SANC00178 is stabilized by a larger number of hydrogen bonds compared to SANC00491, especially in the opened receptor conformations, as shown in Figure 5.7 where higher maxima and bulks of the distributions are observed for SANC00178 in comparison SANC00491. The four kraussianones (Figure 5.8) share very similar substructures mainly composed of heterocycles. While SANC00342 and SANC00352 possessed a central rotatable bond, compounds SANC00347 and SANC00348 were less flexible due to their fused rings. The narrower COM distance ranges for SANC00342 and SANC00352 as seen from Figure 5.6 can be hypothesized being the result of combined central mobility and exposure of planar ligand moieties to the inner walls of the binding cavity. However, trajectory visualisation hinted at a weaker retention of SANC00342 in

Table 5.2: Top 15 natural compounds hits with reported uses from SANCDB[1]

Accession	Compound	Source organism	Reported use
SANC00178	Cephalostatin 1	*Cephalodiscus gilchristi*	Anticancer activity
SANC00491	Cephalostatin 17	*Cephalodiscus gilchristi*	Anticancer activity
SANC00342	Kraussianone 1	*Eriosema kraussianum*	Treatment of erectile dysfunction
SANC00347	Kraussianone 4	*Eriosema kraussianum*	NA
SANC00348	Kraussianone 5	*Eriosema kraussianum*	Contraction of corpus cavernosum tissue
SANC00352	Kraussianone 6	*Eriosema kraussianum*	NA
SANC00436	Mamegakinone	*Euclea natalensis*	Antibacterial
SANC00518	20(29)-Lupene-3β-isoferulate	*Euclea natalensis*	NA
SANC00584	Scutiaquinone A	*Scutia myrtina*	Anthelmintic Activity
SANC00585	Scutiaquinone B	*Scutia myrtina*	Anthelmintic Activity
SANC00670	Saundersioside F	*Ornithogalum saundersiae*	Cytostatic activity against HL-60 leukemia cells
SANC00671	Saundersioside G	*Ornithogalum saundersiae*	Cytostatic activity against HL-60 leukemia cells
SANC00672	Saundersioside H	*Ornithogalum saundersiae*	Cytostatic activity against HL-60 leukemia cells
SANC00768	Marchantin C	*Marchantia polymorpha*	NA
SANC00772	Marchantin H	*Marchantia polymorpha*	NA

comparison to SANC00352, with the latter showing reduced translational mobility owing to planar hydrophobic contacts in addition to polar interactions mediated by carbonyl and hydroxyl functional groups. Despite being the best hit molecule from docking, SANC00347 displayed some mobile interaction modes involving the protease flaps despite a reduced flexibility, especially when not initially docked at the binding cavity's floor. In two cases, namely the opened conformation variants A3 and A5, early signs of ligand exit though the 80S and flap loop were observed. Otherwise stabilization was mainly mediated by a central hydroxyl group interacting with the floor of the binding cavity. Compound SANC00348 was stabilized by the similarly-positioned hydroxyl functional group despite an increased asymmetry resulting from a lack of heterocycles on one side. However, a higher frequency of interactions with the flaps or 80S loop regions were observed in several complexes, namely in A2, A3, A9, B2 and B9 pointing to an increased probability of ligand exit. In particular we note that COM distance captured the decreased stability better than ligand RMSD. With respect to hydrogen bonding (Figure 5.7), the kraussianone members generally maintained at about 2 hydrogen bonds over the duration of MD simulations. Mamegakinone (compound SANC00436) was mainly unstable across protease conformations and variations, displaying increased interactions with receptor flaps and the 80S loop, at the expense of reduced contact with catalytic residues. In the initially-opened receptor conformation A6, SANC00436 leaves the active site. Compound SANC00518 was stable in several cases but left the active site in A2 and showed tendencies for ligand exit in A9 and B5. Additionally, hydrogen bonding propensity was low and less maintained in few cases, as seen in Figure 5.7. Both scutiaquinones (SANC00584 and SANC00585) tended not be retained within the active site due to their large planar size imparted by multiply-fused rings and limited flexibility, even though SANC00585 was slightly more flexible due to a different heterocycle arrangement. The latter compound was nevertheless found to be stable within a subset of the receptor conformations, namely A1, A2, A6, A8, A9, A10, B1, B3 and B6. As seen in Figure 5.7, both compounds displayed relatively poor hydrogen bonding properties, while SANC00584 additionally produced the largest maximum COM distance 5.6. The three saunderosides (SANC00670, SANC00671 and SANC00672) are also bulky and share a common scaffold, differing mainly terminally by few functional groups namely a hydroxyl, ketone and ether groups (Figure 5.8). Flexibility was achieved via rotation about ether linkages located on the one side of this class of moderately large compounds. This mainly resulted in self-folding within the active site and increased interactions with protease flaps, whilst allowing for linear ligand poses in few cases. As all three saunderosides displayed similar dynamics, it would seem that their different functional groups were not critical for their retention within the active site, at least for the period observed. Protease flaps, and in certain cases the 80S region, were key for stabilizing the saunderosides. Stabilization of the linear poses was mediated via the central hydroxyl groups involved in intermittent interactions with catalytic aspartates and via partial symmetry resulting from the presence of heterocycles at both ends of the molecules. The sheer bulk of the compounds meant that part of the saunderosides could be exposed outside of the binding site in two cases (compounds SANC00670 and SANC00672B1 in receptors B1 and B5 respectively), despite being stably folded within the active site. In all, these three compounds displayed variable hydrogen bonding propensities mostly sustained between 5 and 10 bonds (Figure 5.7) whilst showing no general trend with the two main receptor conformation. In addition to displaying the highest number of hydrogen bonds, the saunderosides were not showing much translational movement within their respective receptor variants, as observed by the quasi-stationary COM distances (Figure 5.5 and 5.6). In the case of these three compounds, it can be clearly seen that while RMSD is showing significant motion (Figure 5.4), COM distances are very stable, and corroborated with the absence of translational movement from visual inspection of the complexes. From this observation, we can see that protein-ligand COM distance may be a better descriptor of stability in cases where ligands are long and flexible. The two marchantins (SANC00768 and SANC00772) share a cyclic aromatic scaffold held together by multiple rotatable bonds. They only differ by the absence of a hydroxyl group on one of the phenolic moieties in SANC00768. Even though both compounds were initially among the top-scoring ligands from docking, both fit the active site only partly or displayed high mobility within the cavity in several receptor variants despite displaying some more stable poses in certain cases during dynamics simulations. Due to their high number of rotatable bonds and cyclic nature, both compounds also displayed self-compaction, which decreased the frequency of potential π contacts within the mainly hydrophobic walls of the binding cavity by placing the heterocycles along separate planes in a small space. Despite the variations in performance due to certain combinations of residue mutations, results from this screening experiment could be followed up by *in vitro* tests for protease inhibition. Additionally combinatorial optimization experiments can potentially be explored to enrich the pool of candidate ligands with improved performance and/or tolerability, if required. As an example of improving ligand effectiveness, we have investigated the modification of a top-performing compound, as discussed in the following subsection.

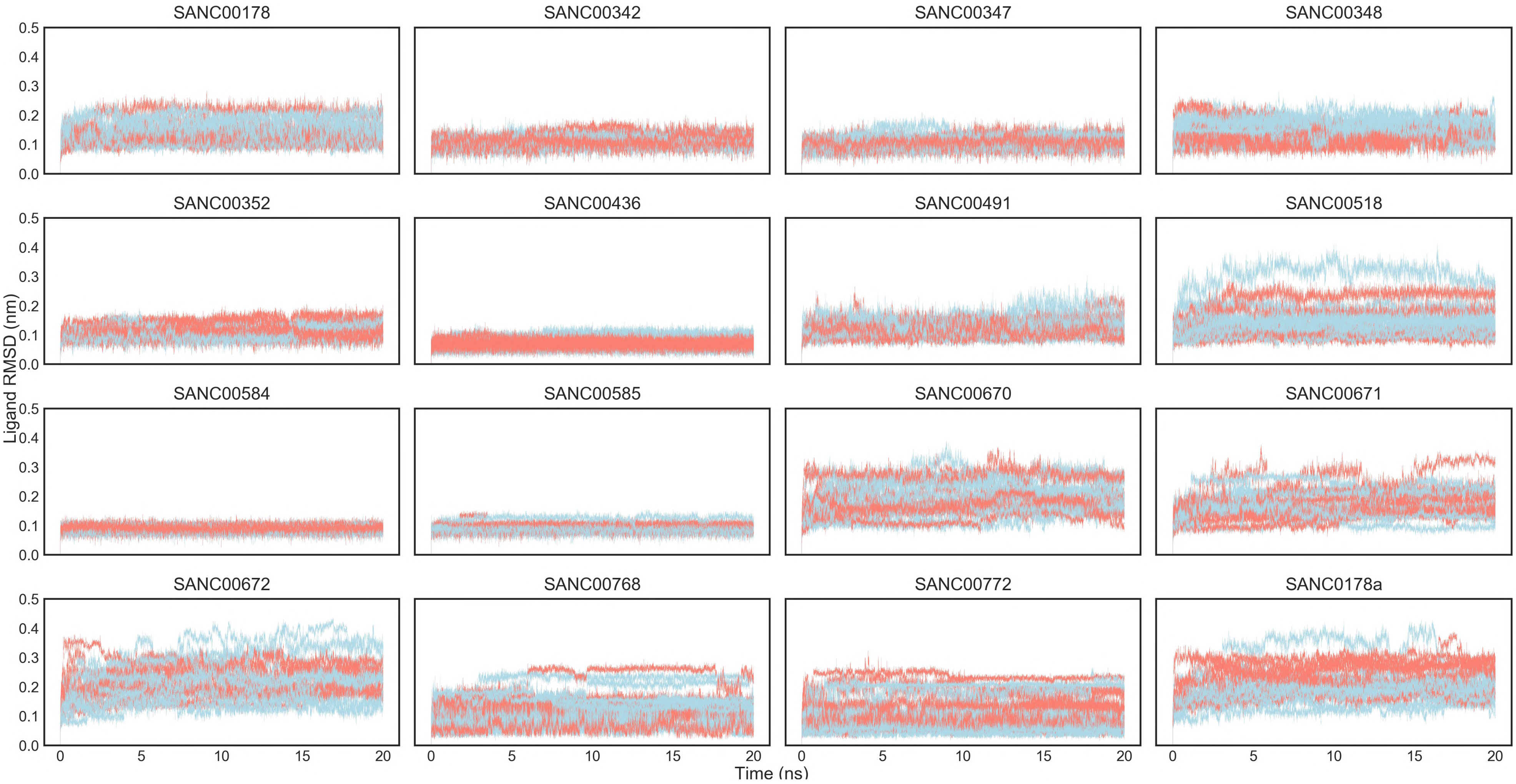


Figure 5.4: Ligand RMSD plots for top 15 SANCDB hit compounds. Opened conformations are in blue while closed ones are coloured red. The modified ligand SANC0178a is also included

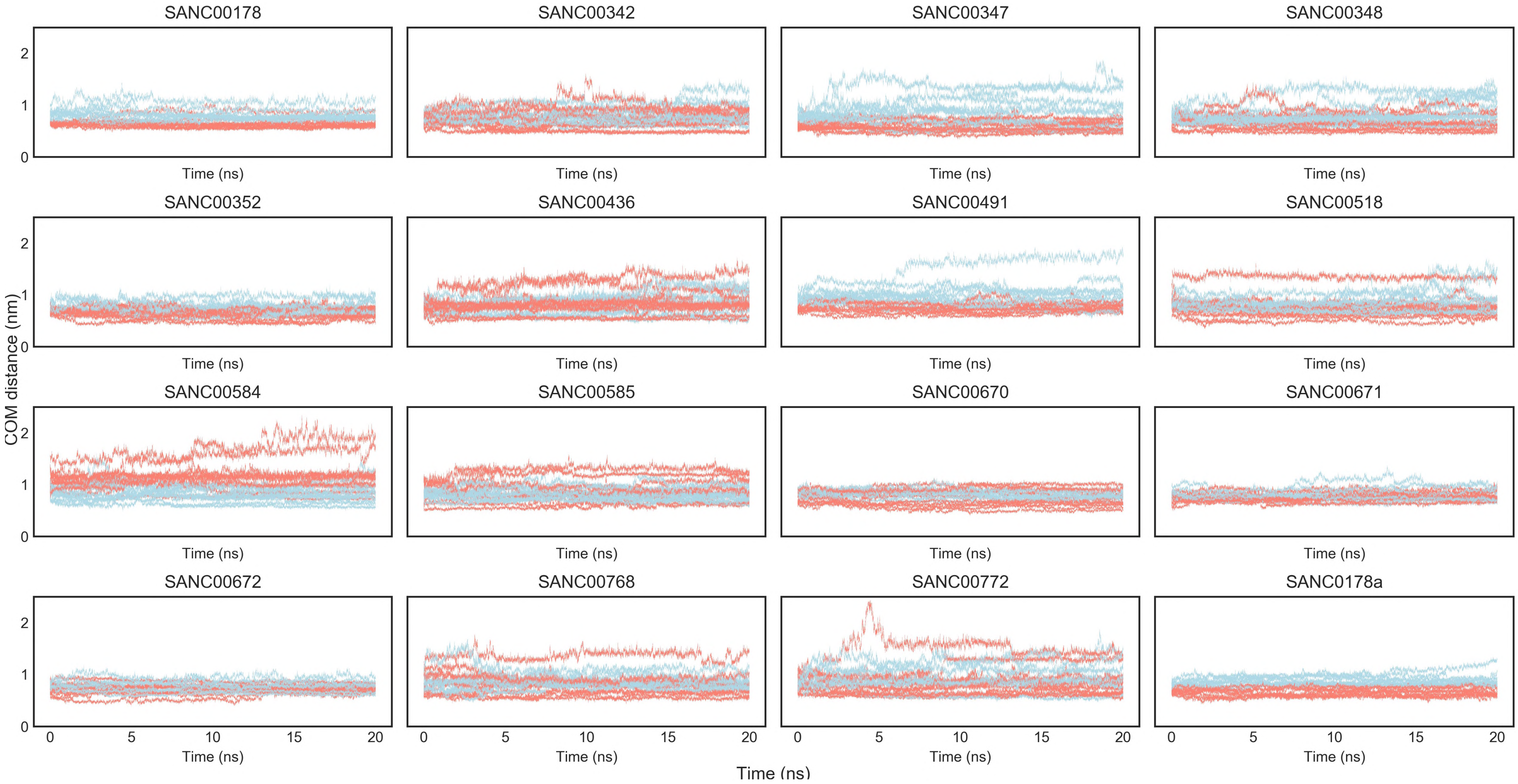


Figure 5.5: Protein-ligand COM distance plots for top 15 SANCDB hit compounds. Opened conformations are in blue while closed ones are coloured red. The modified ligand SANC0178a is also included

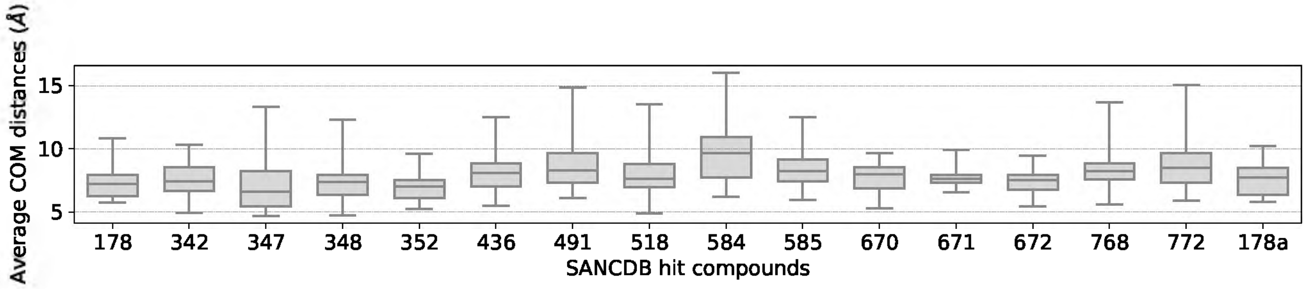


Figure 5.6: Box plots showing the distributions of observed average receptor-ligand COM distances for hit SANCDB compounds over the course of 20ns MD simulations. For each complex the COM distance was time-averaged and combined across all variants and conformations to give a distribution for each compound. Accessions are abbreviated to numbers for clarity.

5.3.4 Improving the performance of compound SANC00178

Compound SANC00178 showed very high affinities for the opened receptor conformations compared to all the other hit compounds, but displayed a very poor performance against the closed conformation receptor B3. After closer inspection, the diminished performance was found to be a result of two bad contacts occurring involving two rigid methyl groups peripheral to the central pyrazine group. One of the bumps occurred between one of the methyl groups and the retained interfacial water while the other bump involved the second methyl group and the protruding side chain of ILE47 from the protease flap. We surmised that the cause was a lack of ligand flexibility and proceeded by cleaving two single bonds found around the pyrazine moiety to result in a partially symmetric ligand topology. After investigating a few cleavage points involving asymmetric ones (not shown) and assessing their overall binding energies, the cleavage points were defined as shown in Figure 5.9. When compared against the hit compounds, the modified compound (SANC0178a) was found to perform exceedingly well overall, retaining the opened receptor conformation specificity while improving the affinity for closed conformations. A centrally-flexible polar pyrazine with partially symmetric long “arms” composed mainly of heterocycles was found to be beneficial, yielding in the majority of cases stable linear (Figure 5.10b) and folded (Figures 5.10c) ligand poses specific to the closed and opened receptor conformations respectively. To give an idea of ligand stability and the strength of these interactions, receptor-ligand contacts are shown as weighted undirected contact network graphs in Figures 5.10a and 5.10d, with each contact corresponding to time averages. The atomic contacts themselves are inferred using a cut-off distance of 4 Å and only those edges prevailing 40% of the time are shown, for clarity. The graphs themselves were produced using an in-house Python script. As a demonstration of applying an the network graph to show ligand performance, the linear and folded poses are shown in Figures 5.10d and 5.10e, corresponding to the complexes shown in Figures 5.10a and 5.10c respectively. The network graph corresponding to the linear pose is the worst performer within receptor B8, while that of the folded ligand pose is typically contact-rich, with no particular choice for selection. In this specific case, compound SANC0178a had migrated to one of the protease monomers to adopt a linear conformation thus decreasing the degree centrality while nevertheless maintaining strong connections with key residues mainly from chain A, namely the catalytic ASP25 (protonated)residue, and the residues GLY48 and ILE47 from the flap hairpin. In the case

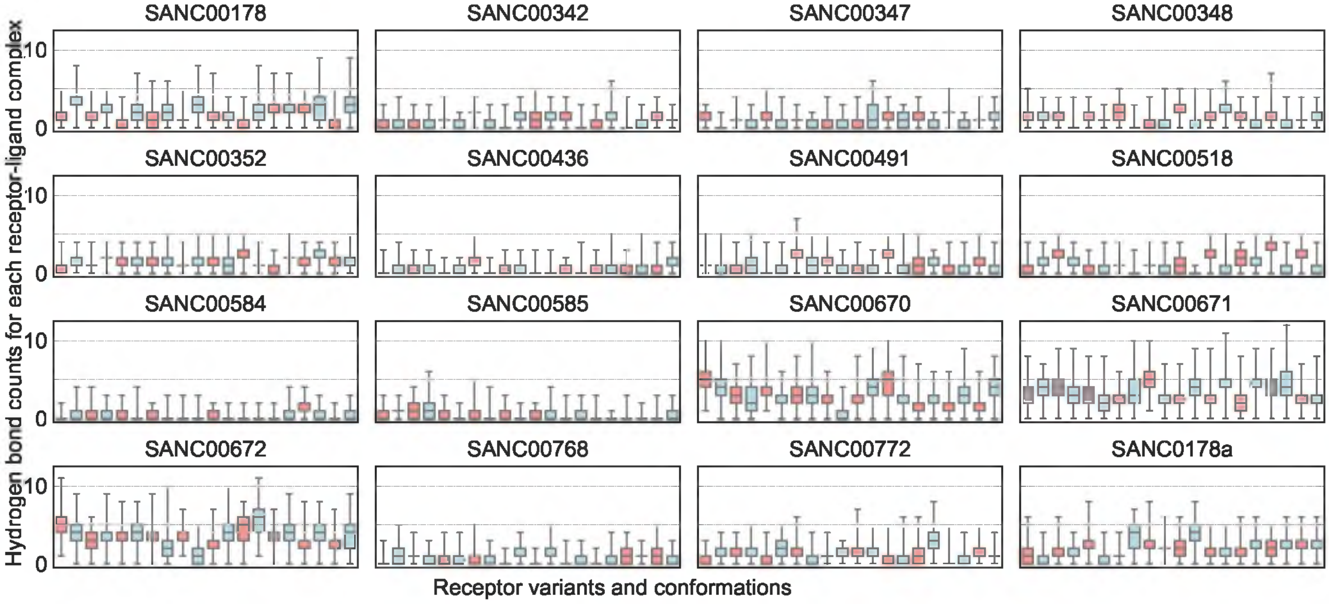


Figure 5.7: Box plots of hydrogen bonding distributions for hit SANCDB compounds over the course of 20ns molecular dynamics simulations. The modified ligand is also included. Opened conformation receptors are shaded blue and closed ones are coloured red. Box whiskers represent the range while a dotted grey lines have been introduced to ease visualization.

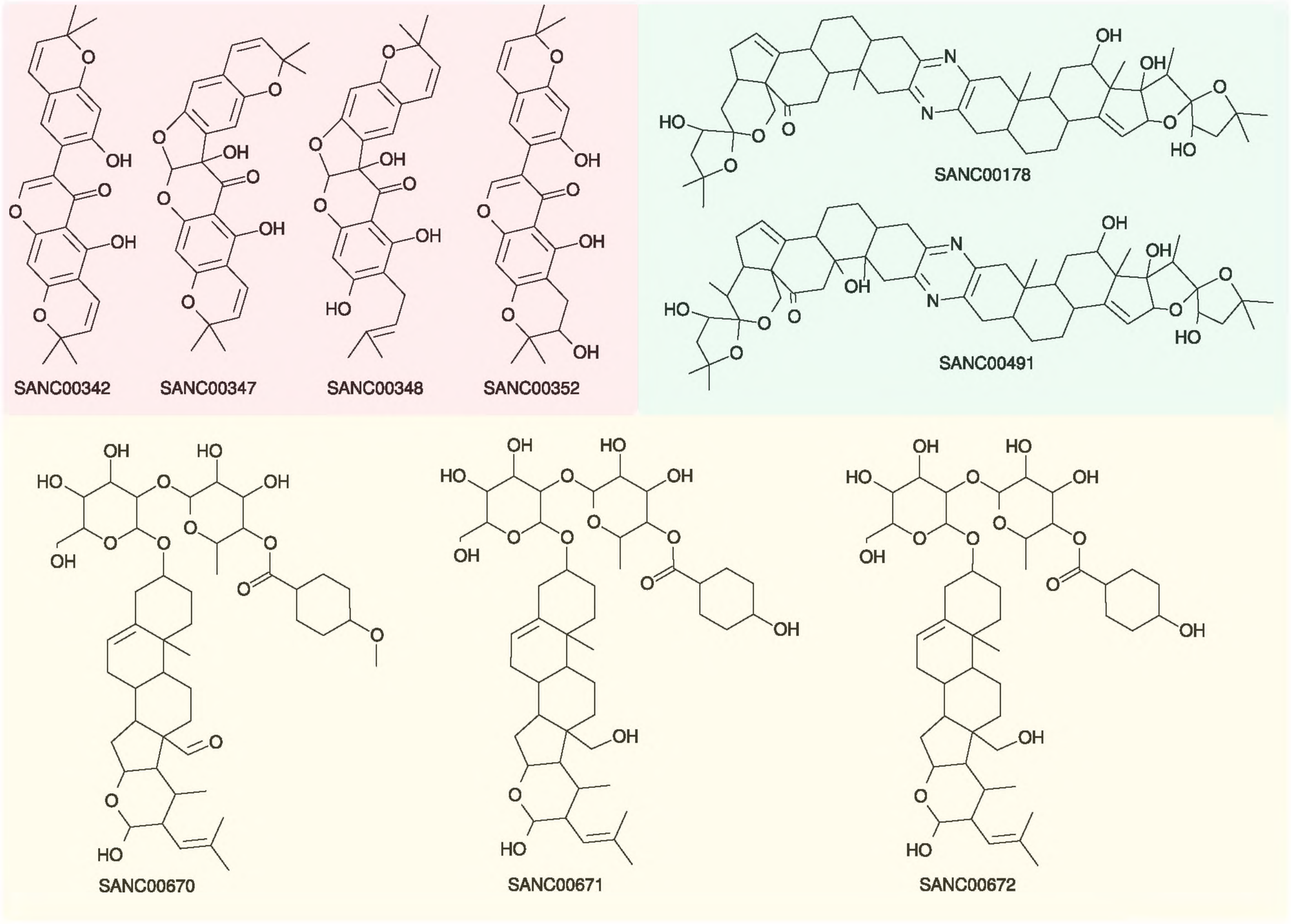


Figure 5.8: 2D structures of the short-listed scaffolds with potential inhibitory activity against DRV-resistant HIV protease sequences based on dynamic properties. Kraussianones, cephalostatins and saunderosides are shaded in red, blue and yellow respectively.

of initially-opened conformation receptors, an abundance of relatively long-lasting cavity contacts were formed with the folded ligand conformation involving both protease chains, as seen in Figure 5.10e. These contacts comprised the catalytic aspartate, residues from the flaps, 80S loops and additional internal surfaces within the binding pocket. Coincidentally, faulty restoration of ligand protonation states after docking initially yielded piperazine upon saturation but displayed similar dynamics with some added stabilization between one of the amine nitrogen and the catalytic aspartate.

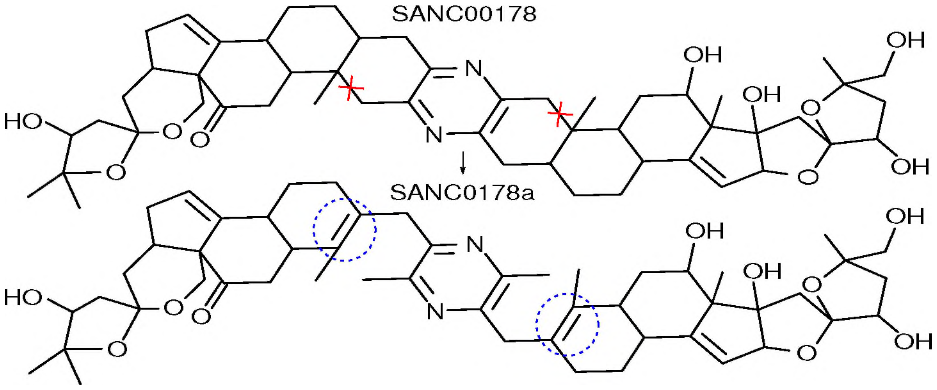
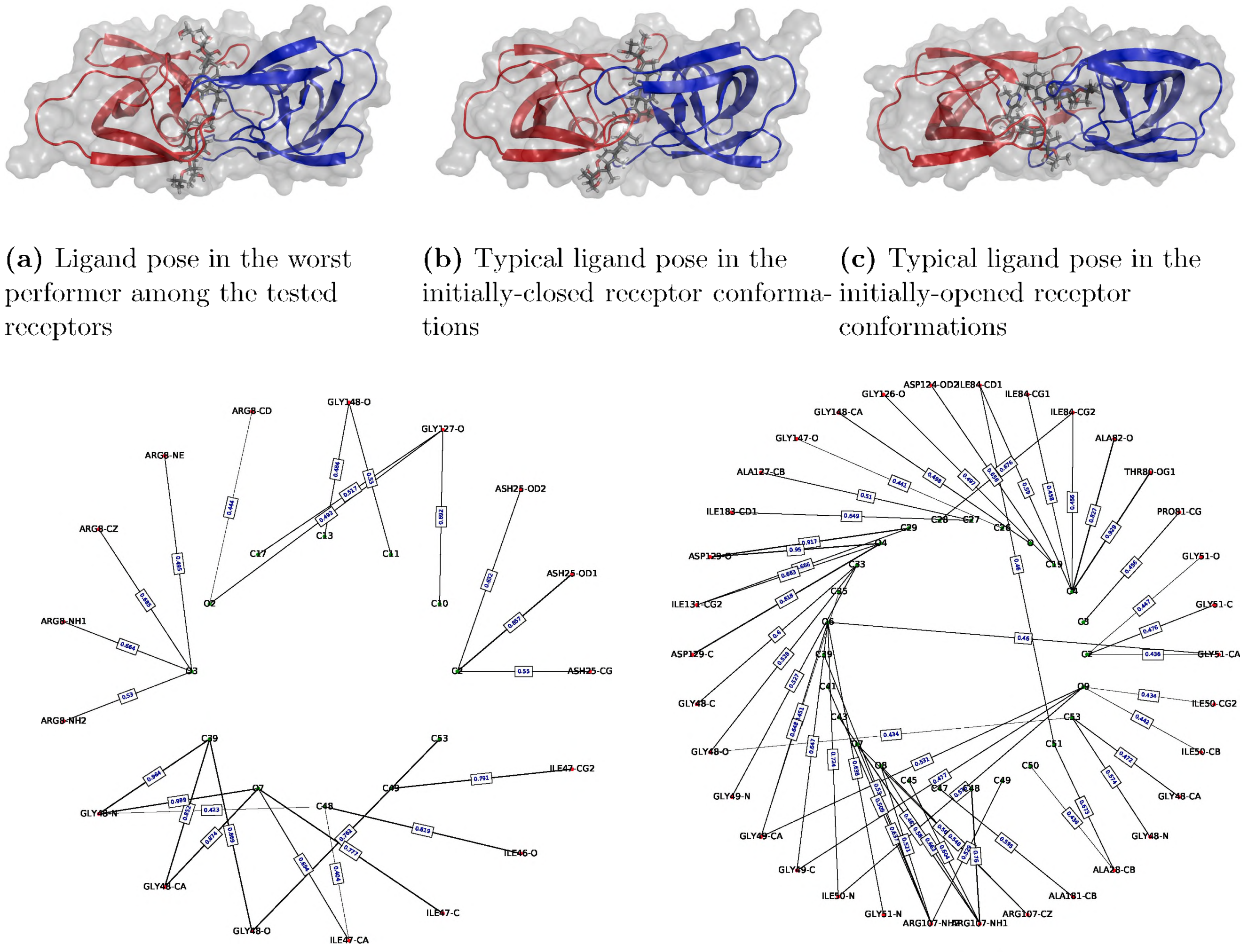


Figure 5.9: Modification of the SANCDB hit compound SANC00178 to SANC0178a in order to improve mobility while maintaining stability, determined by trial and error around the central pyrazine moiety. The red crosses show where the bonds were broken. A double bond (in blue) was used to fill opened valences on each of the affected proximal rings while hydrogen atoms were used for resolving any unfilled valences for the carbon atoms.



(a) Ligand pose in the worst performer among the tested receptors

(b) Typical ligand pose in the initially-closed receptor conformations

(c) Typical ligand pose in the initially-opened receptor conformations

(d) Contact map for the worst performer

(e) Contact map for an initially-opened receptor conformation

Figure 5.10: Final ligand poses (a-c) and weighted protein-ligand contact maps (d, e) estimated from ensemble-averaging of contacts over the course of molecular dynamics. Chains A and B are coloured red and blue respectively, showing the top view. In the bottom sub-figures, only intermolecular contacts above a frequency cut-off of 40% are shown for clarity, with the ligand and receptor atoms laid out as inner and outer concentric circles respectively. The arc labels and width denote the inferred contact weights.

5.4 Conclusions

In this chapter, high-throughput virtual screening was performed using 718 natural compounds from the SANCDB to find potential scaffolds to be used in protease inhibitor design in cases where patients fail DRV treatment, which is the latest FDA-approved PI drug. 20ns molecular dynamics simulations were performed against each of the top hits obtained from docking. Top hit compounds were defined here as molecules with the lowest average binding energies across opened and closed receptor conformations and mutations. The nine hit compounds consist of SANC00178, SANC00491, SANC00672, SANC00671, SANC00670, SANC00347, SANC00352, SANC00348 and SANC00342, enumerated in a generally decreasing order of performance based on several criteria observed from their dynamics. We went one step further to explore ligand modifications of a promising compound, which showed the highest performances in the opened conformation receptors. Increasing the central flexibility of SANC00178 improved its overall performance to produce receptor conformation-specific binding modes, which resulted in receptor affinities raised above all the other hit compounds used in our data set.

General conclusions and potential for further research for Part I

The main core of the study has been a multi-pronged approach, addressing to various extents the problem of drug resistance in HIV-1. Ideally, non-B HIV subtypes would have been a suitable data set to compare any conservation or divergence in drug resistance related behaviour with respect to subtype B, which forms the bulk of most research done in HIV. Non-B subtype information would be crucial in determining the effectiveness of current ARVs in the grand scheme, especially given that subtype B forms only about 12% of the world HIV subtype distribution. However, there is insufficient amount of publicly-available drug resistance-labelled data to perform such an investigation. The next best approach was thus to direct our research efforts towards improving drug resistance prediction and our understanding of it by seeking for a potentially more conserved resistance-related behaviour using the subtype B data. We further contribute by suggesting new scaffolds for use as inhibitory compounds, not from a single drug target but a diverse set of resistant receptors for improved drug activity.

In Chapter 2, the objective was to improve drug resistance prediction in HIV protease and reverse transcriptase using publicly-available data from the Stanford HIVdb. Due to the dataset composition, including a mixture of true biological but also technical variability, a series of filtering criteria were applied for each labelled protease and reverse transcriptase sequence entry in the initially unfiltered data record. These criteria included the variation of ANN architectures (nodes and layers), control of the number of sequence ID-specific variant combinations and the removal of sequences with very low frequency within each of the drug datasets. This resulted in a total of 16 predictive models for each of available FDA-approved drugs from the PI, NRTI and NNRTI classes, which faired well in terms of overall regression and classification performance when compared against Stanford HIVdb and SHIVA web servers, and relatively recent regression models developed by Shen and co-workers [127]. Another main output from this work lies in the filtering methodology, which may be applied to other subtypes should such similarly-labelled data be obtained. From our performance and explanations of phylogenetic concerns we also back the idea of subtype dependence in drug resistance prediction in the wait of a larger and more diverse dataset in terms of subtype composition.

In Chapter 3, we search for a highly-conserved differentiating signal characteristic of drug resistance using a series of existing structure-based tools (homology modelling, energy-minimization, docking, MD, dynamic cross-correlations, modal analysis, PRS and contact network analysis) on large numbers of variants, with the objective of finding a characteristic so conserved that it might extrapolate to non-B subtypes, either mechanistically or in terms of methodology. Unfortunately, none of these approaches showed the existence of such a signal. However, they lead to understand that both drug susceptible and resistant protease conformations share subsets of conformational space, and that the answer may instead lie in the probability of sampling such conformations. With this, the MD data were re-analysed by devising a more sensitive approach, described in the following chapter. Additionally, we have strong evidence backing the known quantitative inexactness of non-equilibrium potential energy estimation methods (using Vina and X-Score) due to their low correlations to lab-based resistance values for multiple drugs. Interestingly, the hydrophobic term was the highest contributing term to drug resistance, though not providing very strong correlations.

In Chapter 4, the redesign of the network construction approach by using stringent statistical tests over averages of pairwise residue distances across resistance states and using ranked degree centrality lead us to find a very strong degree of conserved behaviour, across PI drugs and very closely reproduced after replication in each case. Though re-analysed from Chapter 3, the extent of conformational sampling from MD was influenced by the number of samples needed to retain sufficient variant diversity while keeping the time required for sampling realistic, to obtain results of statistical significance. A lateral expansion and an associated contraction (base of the cantilever moving towards the central catalytic core) were observed, showing some degree of chain symmetry in many cases. Such signal may directly extend to non-B subtypes in similar simulated conditions, or may establish a sensitive method to detect subtype-specific motions in them. Additionally, the approach is generalisable to other systems where a minimum number of 30 variants per ensemble of some labelled phenotype may be made available, the number being generally acceptable for performing t-tests.

In Chapter 5, we switch gears to delve into the domain of protease inhibition. HTVS is used to identify potential inhibitor scaffolds derived from natural compounds obtained by screening 718 SANCDB compounds. Focus was placed on DRV-failing patients and hit effectiveness against multiple variants in both opened and closed receptor conformations, using AutoDock Vina for *in silico* docking before performing 20ns MD simulations on the top ligands, sorted by binding energy. We thus prioritised 9 compounds and proposed modifications to cephalostatin 1 (a known compound with anticancer activity) to generate a dual-conformation lead compound able to stably bind the active site in both the opened and closed conformations. The latter compound showed outstanding overall affinities over all complexes and conformations.

Part II

Side projects

Chapter 6

Overview of the side projects

This section deals with various sub-projects that branched from findings coming from the main research theme in addition to collaborative work.

Starting from the initial idea to make static networks from docked ligand complexes we extended the concept to factor in time for the study of proteins in different states, namely a diseased (case) and a wild-type (control) for the renin-angiotensinogen complex. A contact network is constructed over multiple frames sampled from a molecular dynamics trajectory whereby each residue-residue contact is then time averaged to result in a weighted network in the hopes that these will be a better representation of the overall protein dynamics at a residue position of interest. This forms part of my main contribution in terms of network analysis for this work using MD trajectories. PhD student at the time of the experiment, Dr David Brown investigated betweenness and reachability as global network metrics. The work resulted in a journal publication.

We then introduce the tool suites MD-TASK and MODE-TASK, which are sets of tools designed to study protein dynamics using non-conventional techniques and modes of motion respectively. Both resulted in journal publications, in which I mainly added a Python script for inferring weighted contacts and contributed in software testing respectively.

6.1 Side project 1: Analysis of non-synonymous mutations in the renin-angiotensinogen complex

This project draws from and reproduces certain figures used in the publication listed below. Credit for the reproduced material is given as citations in the respective figure captions.

- Brown DK, **Sheik Amamuddy O** and Tastan Bishop Ö. "Structure-Based Analysis of Single Nucleotide Variants in the Renin-Angiotensinogen Complex." *Global Heart*, 2017 Mar 13. pii: S2211-8160(17)30006-6. doi: 10.1016/j.gheart.2017.01.006. PMID: 28302554.

My contributions for this work are enumerated in the **Publications and contributions** section, item number 5.

6.1.1 Summary

The renin-angiotensin-aldosterone system (RAAS or RAS) plays a vital role in the regulation of arterial blood pressure, plasma sodium concentration and extracellular volume [319]. Over-activity of this system may lead to the onset of multiple pathologies, which can be chronic, acute or even result in death [320]. At the heart of RAS are pressure-sensing juxtaglomerular cells which upregulate secretion of the endopeptidase renin into the blood stream in a cascade of reactions [321] as a response to lowered blood pressure [94]. After activation by proteolytic cleavage of prorenin by the enzymes cathepsin B or neuroendocrine convertase 1 (or even by non-proteolytic mechanisms) [319], the 340-residue long renin [322] catalyses conversion of the liver-produced precursor protein angiotensinogen into the decapeptide angiotensin I [323]. The Angiotensin Converting Enzyme (ACE) subsequently converts the decapeptide into an octapeptide, called angiotensin II [323], which then binds adrenal gland receptors to stimulate secretion of the hormone aldosterone, which increases both the retention of sodium by the kidneys and blood pressure. As shown in Figure 6.1, this system influences the behaviour of several organs in both the cardiovascular and central nervous systems and is an important target used for drug-treatment of hypertension-related morbidities. For these reasons, RAS was thus chosen as a target system to investigate the effect of potentially-damaging renin single nucleotide variations (SNVs) as predicted by *in silico* approaches. For this case study the *in silico* phenotypic predictions of non-synonymous mutations are investigated at the structural level using molecular dynamics and dynamic residue network analysis [204]. Severity of the mutations were inferred by Dr David Brown using the Variant Analysis Portal (VAPOR) server, which aggregates SNV functional effect predictions from five servers, namely PolyPhen-2 [325], PROVEAN [326], Phd-SNP [327], Panther-PSEP [328] and FATHMM [329]. These web servers implement different algorithms for predicting functional effects including Support Vector Machines, Hidden Markov Models, sequence similarity searches and Naïve Bayes classifiers and have various levels of accuracy. With this approach, we hoped that the use of independent predictors would increase the likelihood of finding cases where non-synonymous mutations would lead to conformational or dynamic changes, which would hopefully contrast the protein behaviour in diseased versus a healthy individual. By coupling network analysis to protein structural information in RAS, we cross-evaluated the effects of various SNVs predicted to be pathogenic *in silico* and found significant associated changes with respect to the WT complex, namely in (1) the angiotensinogen variant P40L, which reduced the complex stability, (2) the renin A188V variation, which resulted in increased residue fluctuations and (3) renin D104N, which increased overall rigidity of the complex.

6.1.2 Methods

Data retrieval, filtering and homology modelling: Dr David Brown, who is also first author in our publication [204] performed all the steps described in this paragraph, which lead to the acquisition of high-quality wild-type and mutant models as input for molecular dynamics. He retrieved RAS variants using the HUMA tool and filtered them to retain only non-synonymous variants, before fetching the sequences from UniProt and modelling templates from PDB. He

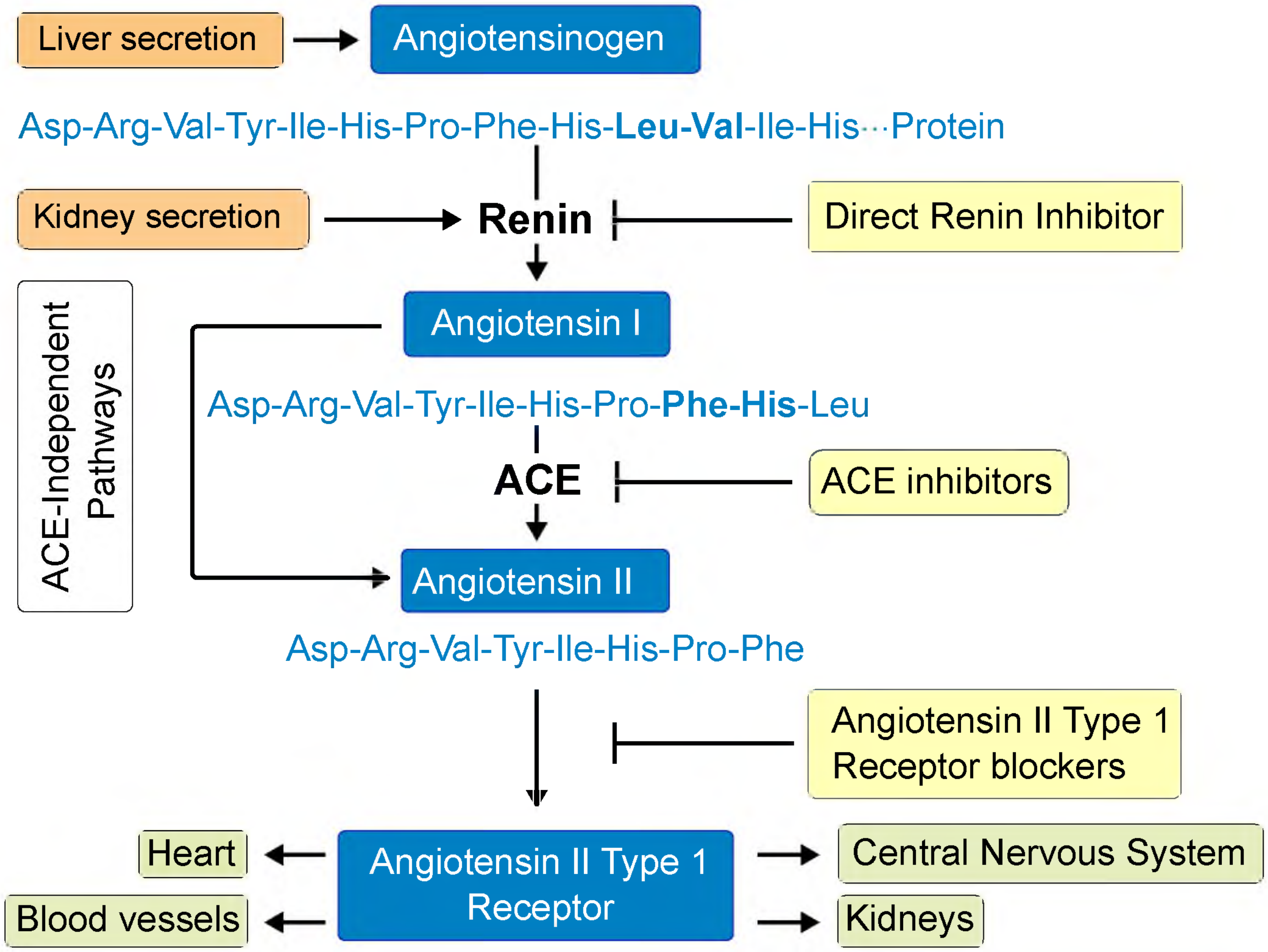

Liver secretion
Angiotensinogen
Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu-Val-Ile-His···Protein
Kidney secretion
Renin
Direct Renin Inhibitor
ACE-Independent Pathways
Angiotensin I
Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu
ACE
ACE inhibitors
Angiotensin II
Asp-Arg-Val-Tyr-Ile-His-Pro-Phe
Angiotensin II Type 1 Receptor blockers
Heart
Blood vessels
Angiotensin II Type 1 Receptor
Central Nervous System
Kidneys

Figure 6.1: The Renin Angiotensin System and strategies for inhibition. In blue is the main pathway forming part of the RAS system. Classes of inhibiting compounds are in yellow. Secreting organs are in orange while affected parts of the body are in green. Adapted from [324] and [323]

homology-modelled the variants and chose high-quality models. These models contained mutations individually introduced in the enzyme, its substrate and in both. The positions of these mutations with reference to the template PDB structures (PDB IDs: 2X0B, 2WXW, 2WXY and 2WXZ) used for modelling are given in Table 6.1. The templates obtained had missing residues at position 1-73 for renin, and for angiotensinogen at positions 1-32 and for the last 3 residues, after target template sequence alignments.

Molecular dynamics simulations: GROMACS 5.1 [263] was used for the simulations, with 480 cores per protein complex at the CHPC. The all-atom AMBER03 force-field was used to define the atom types and parameters for the potential functions. After explicitly solvating the proteins with SPC-modelled water and neutralising charges with NaCl at a final concentration of 0.15 M in a triclinic periodic box (with a minimum image distance of 1.5 nm), the system was energy-minimized by the steepest descent algorithm. An energy gradient tolerance of 1000 $kJ^{-1}mol^{-1}nm^{-1}$ was used for a maximum of 50,000 iterations with the default step size. A cut-off distance of 1 nm was used for short-range non-bonded interactions (van der Waals and electrostatics) while the particle-mesh Ewald algorithm was used for long-range electrostatic interactions for the steps that followed. The LINCS algorithm was used here onwards to correct for rotational bond lengthening after unconstrained updates [209]. After a 100 ps temperature equilibration at

310 K (using the modified Berendsen thermostat [207]), pressure was equilibrated at 1 bar (using the Parrinello-Rahman barostat [208]) for the same amount of time. Finally, 100 ns production dynamics runs were performed with time steps of 2 fs.

Traditional MD analysis: The simulations were analysed by calculating the protein root mean squared deviation (C_α RMSD) and fluctuation (RMSF, per-residue) for the backbone-aligned proteins after correcting for periodic boundary conditions (PBC) and removing rotational/translational motions. As the periodic boundary was not completely removed initially using the molecular centre of mass, each protein complex had to be made whole before removing the periodic jumps and finally box-centering. Performing these steps sequentially was crucial for the success of molecular reconstruction.

Local network analysis: Additionally, we start developing scripts to construct residue interaction networks (RINs) along frames of given MD trajectories and combine them to produce time-averaged contact maps, referred to as Dynamic Residue Networks (DRNs) for each trajectory. The algorithm mainly consists in the definition of C_β atoms (or C_α atoms in the case of glycine) as nodes and edges as the set of node pairs that are within a distance of 6.7 Å apart [261]. These contacts are evaluated over every residue pair for the protein being investigated for the given frame. The network graph then is built by aggregating these contacts and dividing by the number of frames being considered to give a weighted contact. The original graphing functions were used from the igraph library [330] as implemented in the R scripting language, required expansion of the binary XTC trajectory file into PDB format using the trjconv tool from GROMACS. This multi-PDB file was parsed with an in-house Python script designed to calculate the residue contacts (1 representing a contact and 0 for non-contact within a given frame). As the complete graph was too information-rich with the whole complement of residues and their pairwise relationships, the network was instead displayed for the non-synonymous mutation only in each case and compared to the analogous position in the reference RAS complex. We improve visibility and information content by representing the edge weights as log_2-scaled edge widths with the actual frequencies shown as edge labels. These weighted maps were then compared to assess the local impacts of individual mutations over the simulated period with respect to gain and loss of contact. Although generated and developed independently, our approach of combining networks as a weighted was found to share similarities to previous work done by Doshi and co-workers, where the dynamic contacts were instead defined as those being neither rare nor absolutely conserved [262].

Global network analysis: Dr David Brown calculated the betweenness centrality (BC) and average geodesics (L) for each residue using Brandes [331] and Dijkstra's [248] (for pairwise geodesics calculation) algorithms respectively, from the NetworkX library from Python. As explained in Chapter 3, BC includes the number of geodesics that go through a node, while L determines the averaged geodesics to a node, for every residue pair in the protein contact network. To enable evaluation over dynamics, a network graph and the two network centrality metrics were evaluated at 10 ns intervals along the 100 ns trajectory, before computing the average and standard deviation for each individual residue position for the respective metrics. $\Delta L/L$ was evaluated by dividing the difference in L between wild type and the variant by the wild type L value. $\Delta\Delta BC$ values were calculated similarly. Mean and standard deviations were evaluated for both metrics, with focus on two variants of interest, namely angiotensinogen P40L and renin A188L.

Table 6.1: Mutations evaluated

Label	Renin	Angiotensin
R1	D104N	-
R2	R148C	-
R3	R148H	-
R4	A188V	-
R5	L318R	-
R6	F319V	-
A1	-	H39R
A2	-	P40L
A3	-	L43F
A4	-	E48K
A5	-	S49G
A6	-	S49N
A7	-	A104T
A8	-	M105V
A9	-	D168Y
B1	D104N	L43F
B2	R148C	E48K
B3	R148C	S49G
B4	R148C	S49N
B5	R148H	E48K
B6	R148H	S49G
B7	R148H	S49N
B8	L318R	A104T
B9	F319V	A104T
B10	F319V	M105V

6.1.3 Results and discussions

Variant selection and homology modelling: All steps in this paragraph are a highlight of work done by Dr David Brown. Variation data from HUMA contained various SNVs from dbSNP, comprising both synonymous and non-synonymous variants, were retrieved for both renin and angiotensinogen. After discarding non-sense (codons leading to premature termination of translation) and synonymous variation data, only interfacial variations predicted to be pathogenic by VAPOR (a consensus prediction method for functional effects of variations) were considered, totalling 9 and 6 variants for angiotensinogen and renin, respectively. Homology models, once made had z-DOPE values ranging between -1.17 and -1.2, which hinted to near-native conformations. Further, SNVs from renin in close structural vicinity to SNVs from the substrate protein were modelled, yielding a total of 25 variant complexes, in addition to the wild type RAS.

Conventional analysis of MD simulations: Out of all the mutants investigated, R5 MD simulation data was not obtained - bad contacts with water occurring during minimization were not resolved, as the variation had been deemed non-damaging by the VAPOR tool - the gathered variation data was also judged sufficient to proceed with the experiment [204]. Due to the high number of complexes evaluated, the C_α RMSD values are summarized as box plots. Periodic boundary condition were corrected based on initial results from RMSD plots, by sequentially making the proteins whole, before removing jumps and centering these molecules within their respective simulation box. From Figure 6.2, it can easily be seen from the RMSD distributions that complex A2 displays the highest and most divergent RMSD distribution compared to all the others, inclusive of WT RAS. Complex A2 has a higher 25^{th}, 50^{th} and 75^{th} percentile in addition to the highest

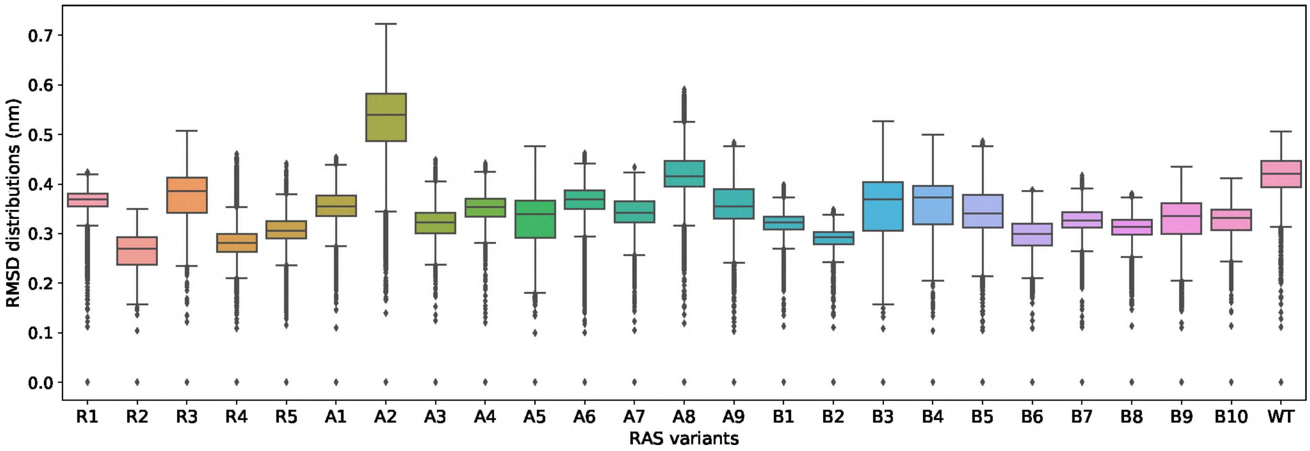
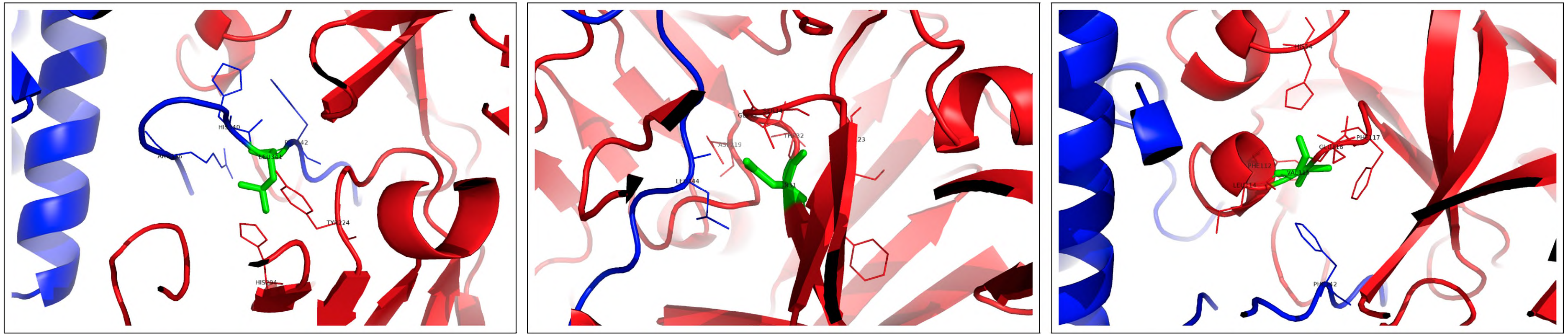


Figure 6.2: Box plots of C_α RMSD values for variants of the RAS complex.

observed maximum RMSD, suggesting a possible differential equilibrium state for the complex, which is significantly shifted from those of all the other RAS variants examined, inclusive of the WT complex, as they were all modelled using a same reference template. As RMSD only shows displacement from the initial reference frame, the calculation of RMSF values for all the complexes were evaluated to highlight local residue flexibility recorded over the entire 100 ns of simulation. We compactly represent RMSF results as the difference with respect to the WT RMSF values (Δ RMSF) in Figure 6.4a. From the heat map, it can be seen that the mutation position, whether in renin or angiotensinogen, did not lead to similar outcomes in terms of RMSF changes as there were no clear renin or angiotensinogen "clades" on the dendrogram. Several patterns of residue fluctuation appear to be conserved (deeper blue hues on the heat map) across variants with respect to the WT protein along the C-terminus from renin (at several regions within renin spanning residue positions 260 to 380), and the N-terminus of its substrate (around positions 80 and 195), suggesting a higher relatively conserved rigidity in these areas. The topmost cluster, mutants A2 and R4 appear to share mainly higher Δ RMSF values towards the C-terminus of angiotensinogen within positions 330-390, when compared to all remaining variants. For more detail, we display the residue fluctuations for complexes A2 and R4, in Figure 6.4b, along with that of R1. It can be seen that A2 and R4 share patterns of residue fluctuation, both being generally more flexible than the WT at the C-terminus of angiotensinogen. Analysis of individual RMSF plots also showed similarly increased flexibility in the double variant B2 (not shown). Variant R1 displayed highest overall rigidity, with respect to the WT complex, as shown in the same figure. In all, each of the tested RAS variants shared a relatively higher rigidity within the angiotensinogen substrate at residue positions 80-100 (surface-exposed loop) and 175-200 (spans an internal alpha helix within the serpin domain), as shown in Figure 6.4a. As the changes experienced by the surface-exposed loop are far-reaching within the complexes, they suggest allostery. On the other hand, the internal alpha helix is sandwiched at the dimer interface, which likely reduces mobility. As both changes seem to occur across all disease-causing variants, they may be correlated movements that lead to associated pathologies. Other areas, such as residue positions 435-450 (surface-exposed loop) in the substrate show less conservation in flexibility, probably associated with decreased stabilizing interactions with the β-strands lying underneath.

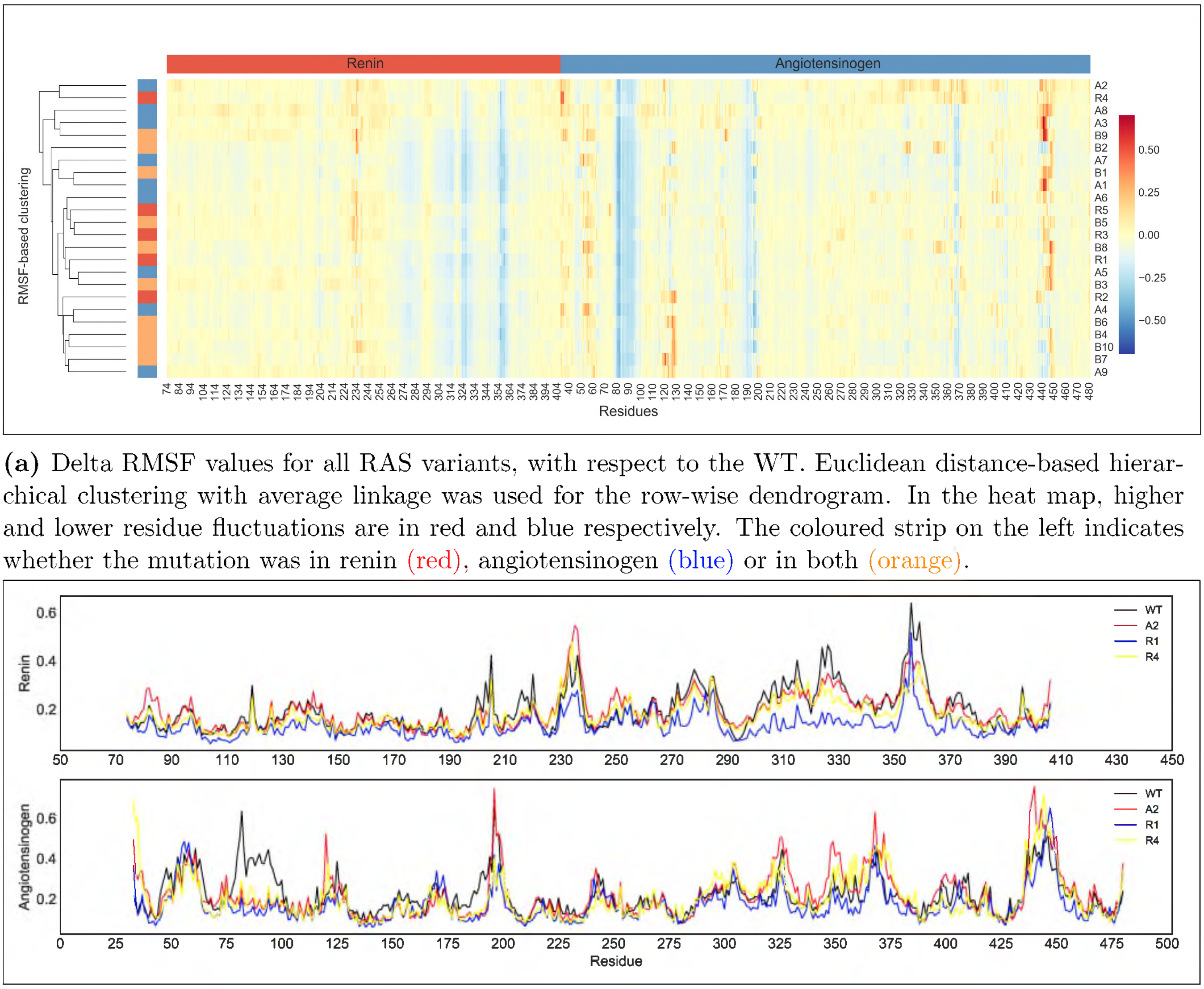


(a) Position of the renin P40L.

(b) Position of renin D104N.

(c) Position of angiotensinogen A188V.

Figure 6.3: Variations of interest, found at the enzyme-substrate interface are shown in green with numberings from the models, while renin and angiotensinogen are coloured red and blue respectively in each of the sub-figures.

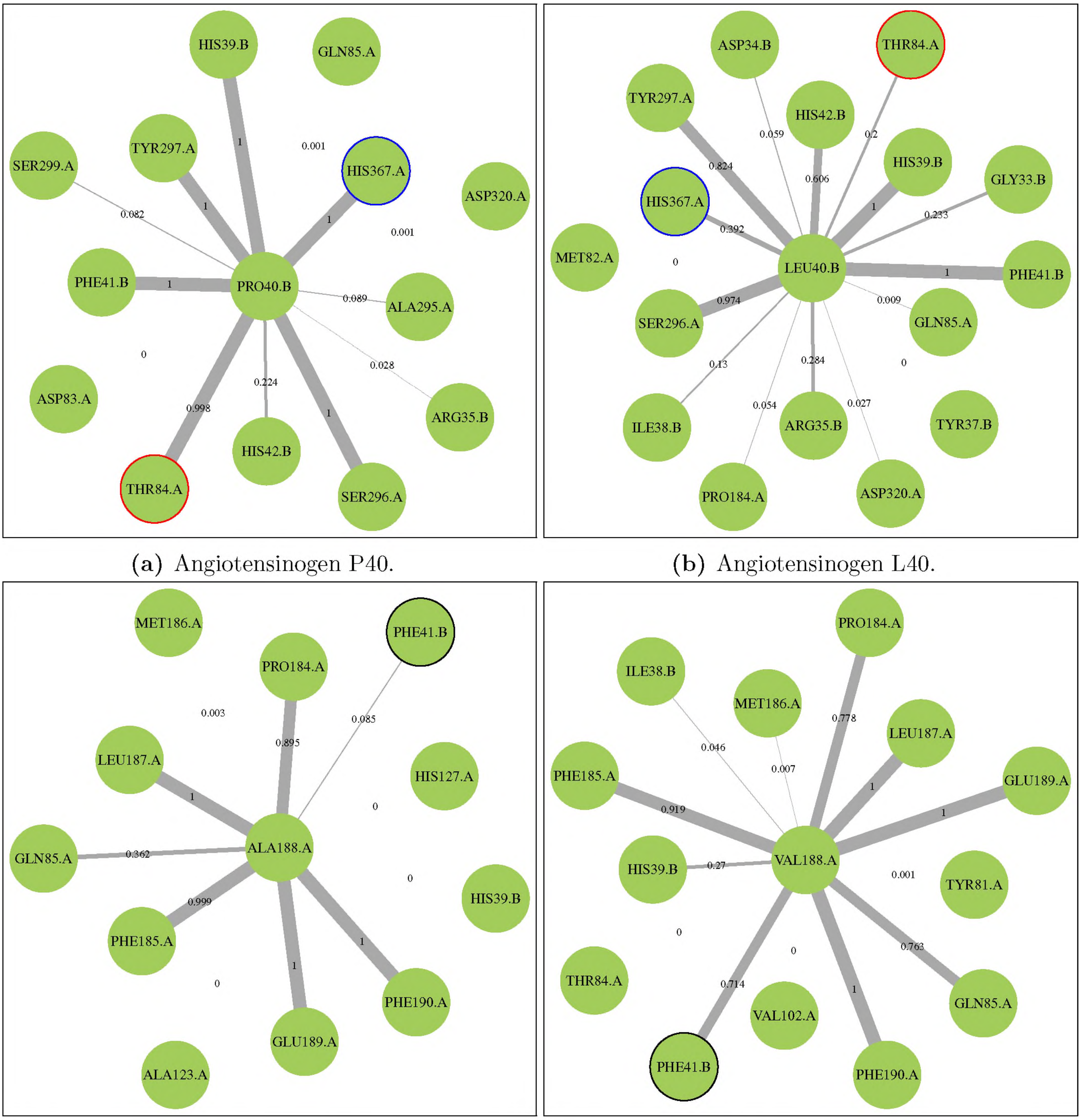


(a) Delta RMSF values for all RAS variants, with respect to the WT. Euclidean distance-based hierarchical clustering with average linkage was used for the row-wise dendrogram. In the heat map, higher and lower residue fluctuations are in red and blue respectively. The coloured strip on the left indicates whether the mutation was in renin (red), angiotensinogen (blue) or in both (orange).

(b) RMSF values for variants A2, B4 and R1, including WT RAS.

Figure 6.4: RMSF values (nm) from 100ns MD simulation for the RAS variants. In sub-figures 6.4a and 6.4b, renin and angiotensinogen begin at residue positions 74 and 33 respectively.

Analysis of Residue Contact Networks: We showcase the application of network analysis for investigating the impact of residue variations over the MD trajectory by evaluating and comparing residue contacts prevailing within a subset of pathogenic variants against those occurring for the wild-type. The simplest method uses a cut-off distance and aggregates the contacts for each residue pair over the course of an MD simulation. Instead of representing the whole network, only one residue is chosen for comparing against the homologous position in the WT protein, to give an idea of contact strength (via weighted edges) and gain/loss of contact at that locus - the SNVs of interest in this case. We show such graphs for the angiotensinogen P40L (Figure 6.5 a-b), and the renin A188V (Figure 6.5 c-d) variants. Edge labels are the observed contact frequencies in each case. The angiotensinogen P40L variation (chain B) significantly weakens the contacts THR84 and HIS367 within renin (chain A) centered around residue position 40. On the other hand, the renin A188V variation strengthened the PHE41 contact around residue position 188 in angiotensinogen. While ignoring any energy metric, the analysis highlights very pertinent information with respect to intra-molecular relationships. At the time of publication, it was not possible to order the



(a) Angiotensinogen P40.

(b) Angiotensinogen L40.

(c) Renin A188.

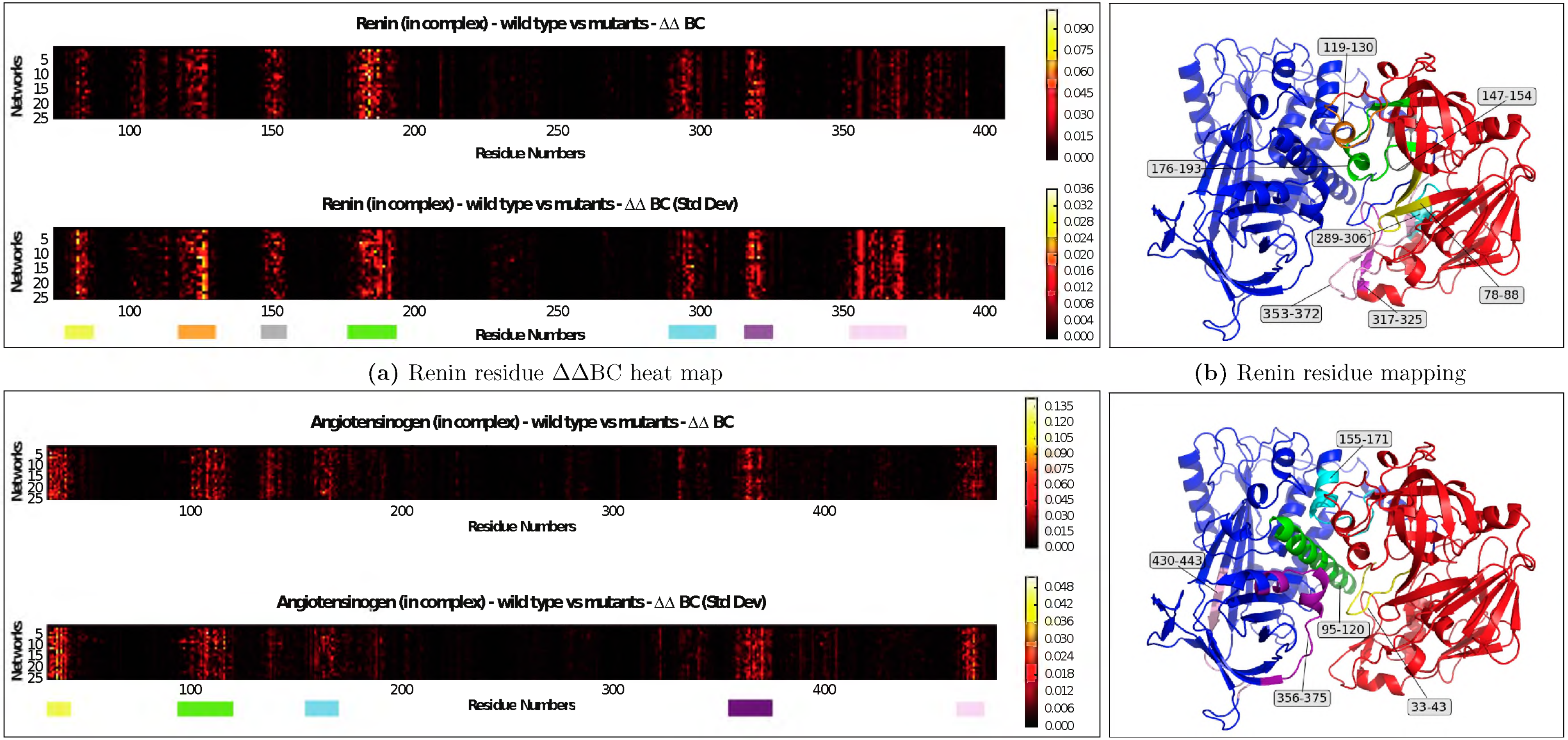
(d) Renin V188.

Figure 6.5: Dynamic residue contact networks for angiotensinogen P40L and renin A188V. Each node is a protein residue, while edge labels denote the observed contact frequencies inferred from MD simulations. The central node represents the WT residue in the left sub-figures while the corresponding variants are shown on the right. Chain information is added after a dot, for convenience. Residues with significant changes in contact behaviour are circled in matching colours (red, blue and black) for the variant and corresponding WT protein. Figures re-used from [204].

nodes to ease comparison across homologous systems due to library-specific (igraph) limitations for manipulating the network graph layout within R. Since then, the R code has been rewritten for Python (2x and 3x) using the NetworkX library, thus facilitating comparisons and reducing the number of dependencies and making the tool more easily maintainable over time.

Analysis of network betweenness centrality: BC values evaluated for the potentially-pathogenic variants with respect to the wild type RAS showed high degree of conservation for this network property (though not absolute) at multiple locations for all of the pathogenic variations examined, in both renin and angiotensinogen. Regions of high averaged $\Delta\Delta$BC values were associated with higher relative variabilities, indicating to a greater extent of conformational differences, involving 7 loci in renin and 5 loci in angiotensinogen, as shown in Figures 6.6b and 6.6a for the enzyme, and 6.6d and 6.6c for the substrate, respectively. Mapping of the impacted residue locations shows that the regions of high $\Delta\Delta$BC values within rennin were found mostly at substrate-interfacing residues, which is not unexpected given that the region connects and mediates tactile information between the enzyme and substrate, being referred to a "high traffic zone" in [204]. A similar trend was observed within angiotensinogen, whereby 4 regions were interfacial, except fo the case of an externally-located β-strand at residue position 430-443. Such a scenario may be a result of differential packing of this secondary structure with respect the substrate's core structure in the pathogenic variants. Overall, these mutation-driven ΔBC differences suggests a compensatory energetic pathway driving the protein dynamics within the complex, which could be important for retaining activity despite the associated pathogenicity of the variations. [204].

Analysis of network geodesics: It can be seen that comparing the averaged geodesics across cases and a control (wild type) highlights a different network property compared to BC. There is a relative extent of centrality conservation across the potentially-pathogenic variants, however the variants angiotensinogen P40L and renin A188L stood out with more divergent characteristics from the rest in terms of average and variance, respectively. The renin enzyme from the angiotensinogen P40L variant displayed a consistently higher average difference in averaged geodesics, from residue 250 onwards to the C-terminus in renin, as shown in Figure 6.7 A and B. In the case of renin A188L, the averaged ΔLL showed less average divergence from the wild type, but displayed higher variability, as shown in Figure 6.8 A and B for both chains of the RAS complex. Contact maps at the renin A188L variation show weakened contacts with the angiotensinogen PHE41 residue with respect to the wild type, as seen in Figure 6.5 (c) and (d). The presence of PHE41 within a loop may explain the increased variability in reachability, resulting from the continuous formation and breakage of contacts due to a higher mobility.



(a) Renin residue ΔΔBC heat map

(b) Renin residue mapping

(c) Angiotensinogen residue ΔΔBC heat map

(d) Angiotensinogen residue mapping

Figure 6.6: ΔΔBC heat maps (left) and the corresponding 3D mapping of residue segments with differential ΔBC values. The enzyme renin and its bound substrate angiotensinogen are depicted as red and blue cartoon representations respectively. Sub-figures (a) and (c) are mainly re-used from [204].

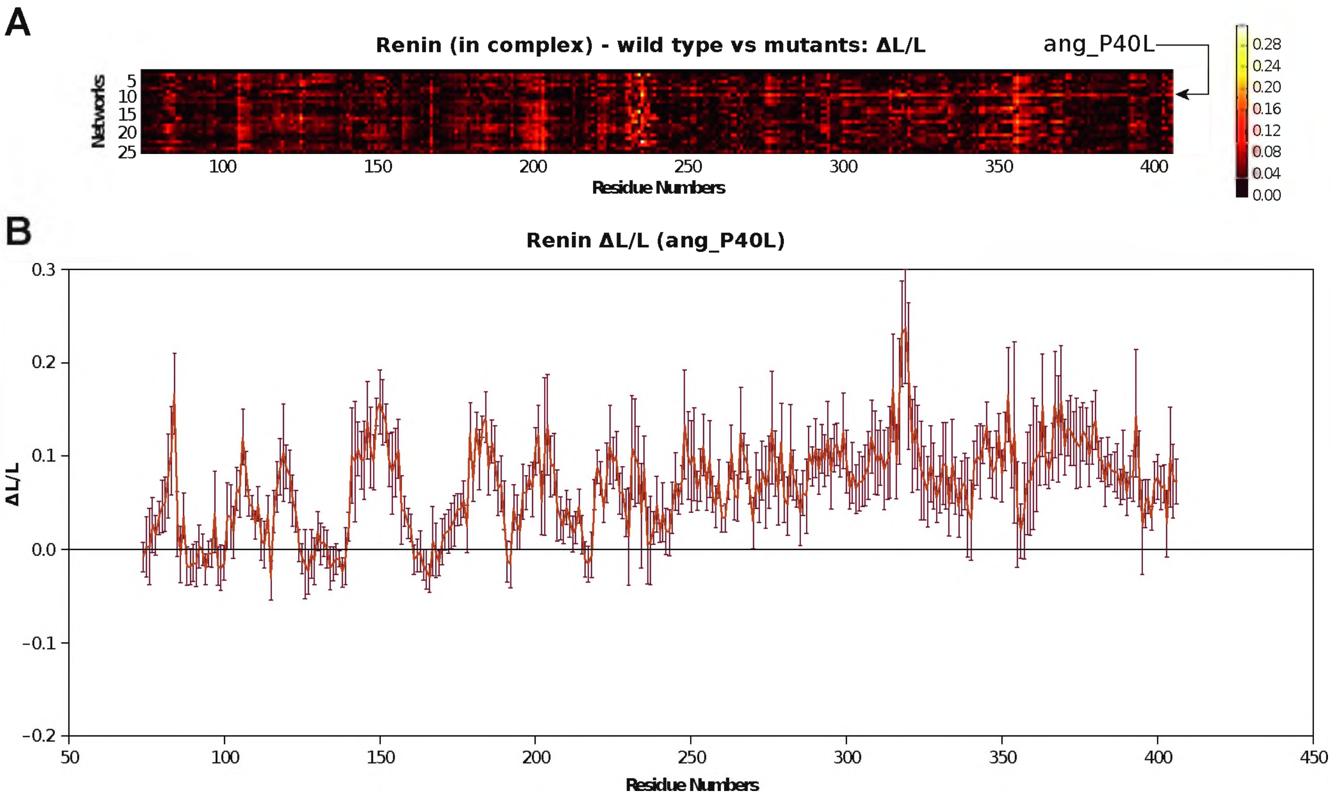


Figure 6.7: Differences in geodesics for the variants with respect to the wild type RAS (only renin is shown). All the variants are shown in sub-figure A, while a more detailed profile is shown for the angiotensinogen P40L variant in sub-figure B with error bars depicting standard deviations. Re-used from [204]

6.1.4 Conclusions

This case study application of dynamic network analysis for the context of analysing residue behaviour from MD data has proved very useful in highlighting both local and distal events, as weighted residue contact maps and via the use of network centrality metrics, respectively. Weighted contact maps visually represent the strength of residue contacts via the frequency of connections, thus proving to be a very useful and easily-interpretable non-energetic characterisation of the effect of non-synonymous variations between protein systems. The two explored network centrality metrics L and BC, prove very useful in extracting global effects not arising directly at the site being investigated, thus potentially highlighting allosteric effects, if such exist across homologous systems. Presented methods thus show an additional layer of information, compared to metrics such as radius of gyration, RMSF and RMSD, suggesting that developed methods can be used to complement the latter.

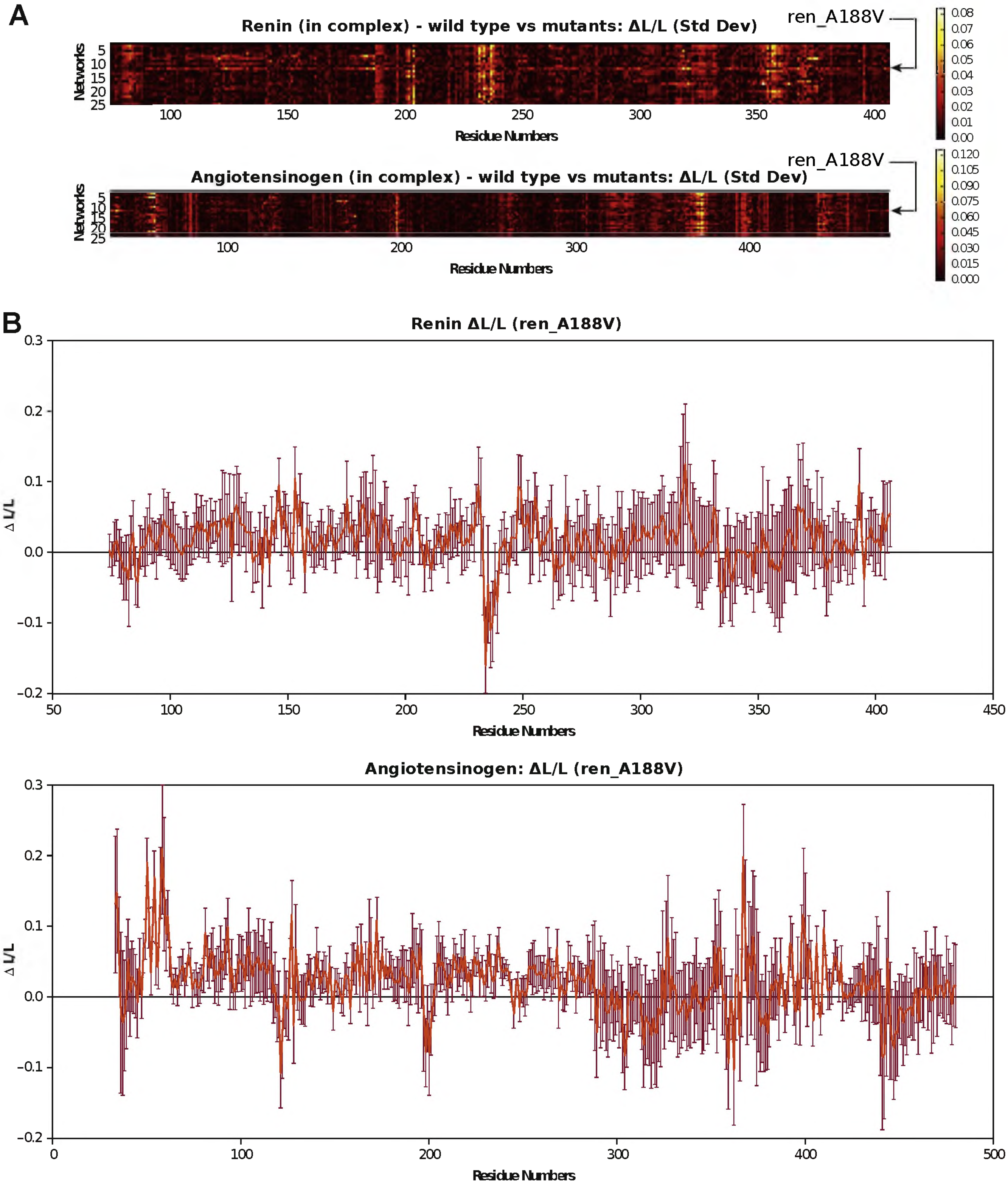


Figure 6.8: Differences in geodesics for the variants with respect to the wild type RAS (for both renin and angiotensinogen). Standard deviations along the whole complex are shown in sub-figure A. Sub-figure B gives the detailed profile for the renin A188V variant, along both chains of the complex. Error bars depict the standard deviations. Re-used from [204]

6.2 Side project 2: An MD analysis tool: MD-TASK

This project draws from and reproduces certain figures used in the publication listed below. Credit for the reproduced material is given as citations in the respective figure and table captions.

- Brown DK, Penkler DL, **Sheik Amamuddy O**, Ross C, Atilgan AR, Atilgan C and Tastan Bishop Ö. "MD-TASK: a software suite for analyzing molecular dynamics trajectories." *Bioinformatics*, 2017 May 31. doi: 10.1093/bioinformatics/btx349. PMID: 28575169.

My contributions for this work are enumerated in the **Publications and contributions** section, item number 4.

6.2.1 Summary

After developing the weighted contact network representation from the previous work with the RAS system, the script was refined and generalized before being incorporated in the MD-TASK suite of tools meant to supplement traditional methods of analysing results from MD simulations. Available tools include dynamic residue contact network analysis, perturbation response scanning (PRS), and dynamic cross-correlation (DCC) scripts, which for the main part use topologies and trajectories as input. My main contribution in the tool suite is in the writing and testing of a weighted residue contact map, which gives the time-averaged contacts around a chosen target residue, accumulated from an MD simulation. A typical application of residue contact maps would be in the comparison of conserved contacts between cases and controls, which could be for example (1) a mutant versus its wild-type protein, (2) to monitor the effect of ligand-binding to receptors versus the apo-protein or (3) to investigate a residue of interest such as one found at protein a interface region, following exposure to a given environmental perturbation. All of the available tools can be used in conjunction with conventional methods such as RMSF, to highlight protein functional/mechanistic properties for biological inference, to the extent permissible by the forcefield. This work [237] was designed as an Application Note paper and thus emphasis is laid on how the MD-TASK tool can be used, rather than a research paper describing a molecular phenomenon. Wild type and drug-resistant HIV proteases were used as a test system. Scripts for the MD-TASK suite are available on GitHub at the following address, https://github.com/RUBi-ZA/MD-TASK, where documentation is also available. In the following sub-section, we give a brief overview of the tool suite and then discuss a case using HIV protease as example.

6.2.2 Methods

Example results from the MD-TASK tool suite were generated using the 198-residue long HIV protease dimer. A wild type sequence was obtained from the crystal structure with the PDB ID 4ZIP [332], while the mutant sequence was obtained from the crystal structure with PDB ID 3S54. The major DRMs comprised variations V32I and I47V with the accessory DRM V82I. Closed and opened receptor conformations were modelled using the crystal structures with PDB IDs 3S54 [333] and 1TW7 [109] respectively for the wild type and resistant protease variants. Further details are provided in the following subsections.

Dynamic residue interaction networks: Dr David Brown (former PhD student at RUBi) wrote the scripts for performing betweenness centrality and average shortest paths calculations. These involved the computation of such metrics at each selected time frame along an MD trajectory before estimating their average and standard deviation, at each residue position. For this evaluation, he used 40ns of simulated MD data as input, to show the L and BC profiles in the opened conformations of both the wild type and mutant.

Dynamic residue contact maps: My main contribution was in the generalisation of a script for carrying out residue contact map calculations, which displays the time-averaged local contacts around a designated residue using a distance cut-off value from an MD simulation. Consequently the edge label values for the contacts will be in the range [0,1], denoting very poor and conserved, contacts respectively. The values are rounded to a maximum of 3 decimal places for improved readability and aesthetics. More emphasis will be laid on the contact maps for this case study.

DCC: Caroline Ross (PhD student at RUBi) wrote the script for performing dynamic cross correlations, which estimates trends in pairwise residue motions over time, from an MD simulation. She evaluated DCC using the opened conformation mutant protein. The equation used (Eq. 3.14) is the same described previously, in Chapter 3.

PRS: David Penkler (PhD student at RUBi) wrote the tool for performing perturbation response scanning, which sequentially applies random uniform forces around each residue via the dot product of independent force vectors to inverted Hessian matrix obtained from an MD trajectory, before calculating correlations to a target conformation. At the heart of PRS is the equation given earlier in Chapter 3 (Eq. 3.19), which describes the dot product between an inverted Hessian matrix and a force vector. For demonstration purposes, he selected a 20ns equilibrated portion from MD data obtained from the mutant to evaluate perturbations leading towards the closed conformation triple-mutant structure 3S54 as a target state. 50 random uniform forces were applied at each residue. Due to the reliance on an energy minimum for the construction of a Hessian, it is conceivable that it will be negatively affected by a system that traverses multiple such minima.

6.2.3 Results and discussions

With its suite of tools, MD-TASK supplements traditional ways of analysing MD trajectories, packaging network-based methods together with PRS and DCC. Our method of constructing contact networks benefits from averaged values and avoids possible over-minimization, which is in contrast to work by Ozbaykal suggested several thousands of energy minimization steps [261]. We strongly believe that the accuracy of current forcefields is insufficient to cater for the more intricate quantum effects that become more apparent and significant as atoms draw closer, such that allowing a complex biomolecule to proceed too far downhill in the realm of standard molecular mechanical forcefields may lead to non-realistic conformations. When assessed over 2000 frames for a 599 residue-long protein using a provided example trajectory (example_small.dcd), analysis run times by MD-TASK scripts varied from minutes to several hours, as shown in Table 6.2, where performances were evaluated by Dr David Brown on a desktop PC (Intel Core i5-6300U quad-core

Table 6.2: Time of execution for MD-TASK scripts. Re-used from [237]

Script	Average time (sec)
calc_network.py (–calc-L)	37 298
calc_network.py (–calc-BC)	62 109
calc_delta_BC.py	16 852
calc_delta_L.py	1 864
avg_network.py	1 713
compare_networks.py	4 230
delta_networks.py	2 289
contact_map.py	19 806
calc_correlation.py	39 703
prs.py	95 480

CPU clocked at 2.4 GHz, with 8 Gb RAM) running the Ubuntu 16.04 operating system. Results from all MD-TASK scripts are shown in Figure 6.9. Averaged residue reachability (Figure 6.9A) showed that some regions are similarly highly central for global communication in both chains A and B for both the wild type and the mutant proteases around the mid portion of the sequence (residues 46-56, the flap), and also highlight some partial symmetry across chains, indicating some level of similarity of the protein dynamics. A high reachability value suggests high peripheral network connectivity (a denser network), which may possibly occur via compaction and may also be reinforced by stable immediate contacts at the flap tips. Betweenness centrality (Figure 6.9B) was less symmetrical and showed some more differentiation between the wild type and mutant. We note that BC and farness profiles are not subsets of each other - they do not display any obvious signs absolute correlation, as seen in Figure 6.9A and B. As pairwise geodesics, comparative residue betweenness hints at distal effects. High BC residues are important information-mediating nodes, able to cause significant disruption, despite their possibly low local connectivity. High BC was observed at residue position 25, which is not only the catalytic aspartate, but also forms part in a very strong hydrogen-bonding network within a loop, termed the fireman's grip that maintains dimer stability. Despite their flexibility due to lack of regular secondary structure, the connectivity is lowered, however BC appropriately captures its importance as an information mediation node. By evaluating pairwise residue correlations over the protein length in Figure 6.9D, DCC reveals that isolated stretches of residues within the same chain move in a correlated fashion (chain A: residues 1-99 or chain B: residues 100-198), while anti-parallel motion is observed between chains. Sequential residue perturbation by the PRS method shows multiple residues coming from both chains A and B from the closed conformation wild-type protease that when perturbed lead to the targeted opened receptor conformation for the drug resistant variant, with part of the cantilever area (residues 60-72) ranking highest in terms of correlation whilst showing conservation across chains for the analogous positions. The weighted residue contact map in Figure 6.9 has been re-generated and magnified in Figure 6.10 using the latest version of the "contact_map.py" script. As an example, residue contacts around the major DRM position I47V are shown for both the wild-type and the drug-resistant variant. The script "contact_map.py" takes as input the MD trajectory and topology files and uses the Python library MDTraj for reading the inputs. The pairwise residue distances are then calculated within a radius of 6.7Å by default around each C_β

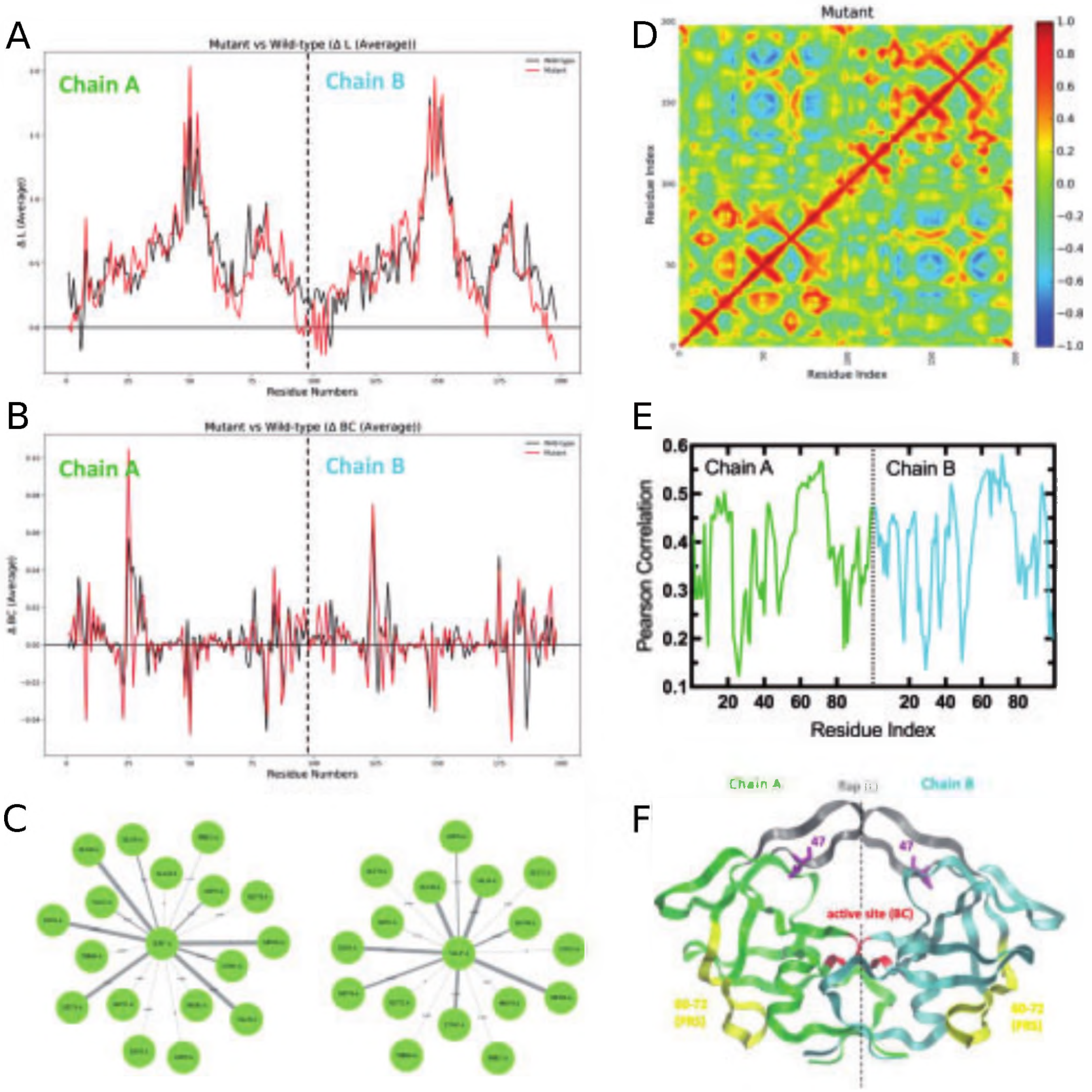
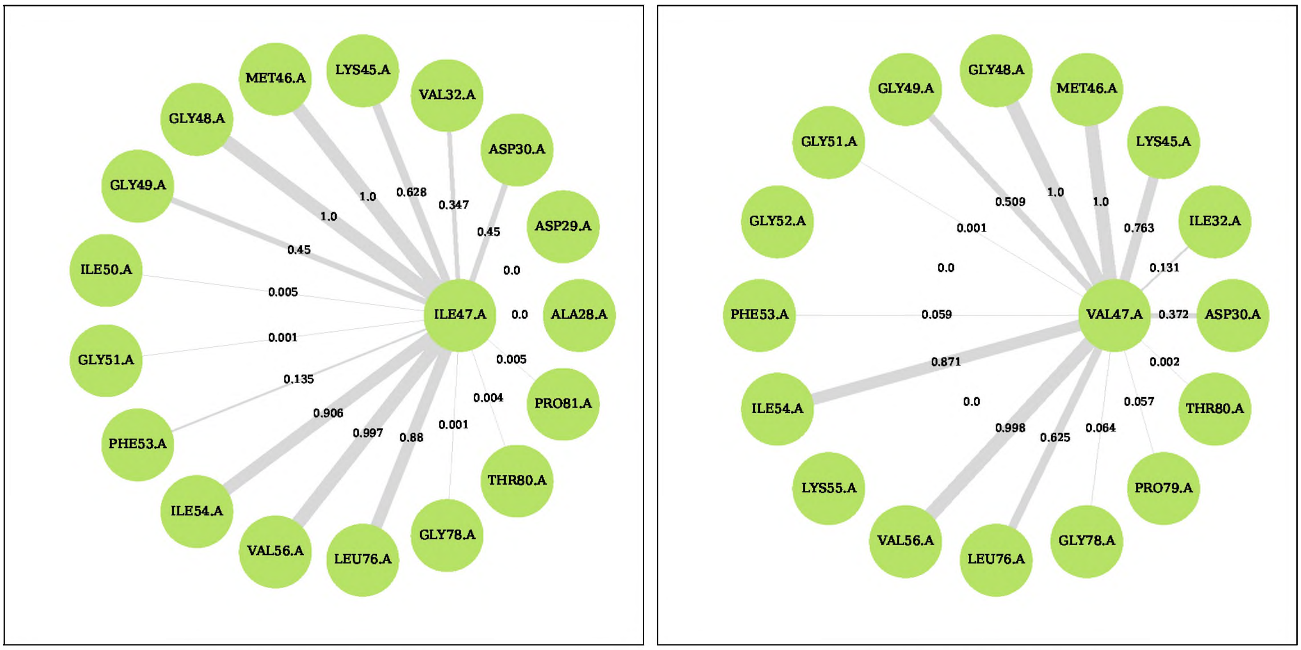


Figure 6.9: Example outputs from MD-TASK, using HIV protease, showing network properties in (A) ΔL, (B) ΔBC and (C) residue contact maps, and then (D) DCC and (E) PRS. 3D mapping of PRS correlations (F) was done separately by David Penkler. Re-used from [237]



(a) Wild-type ILE47

(b) Variant VAL47

Figure 6.10: Magnified sub-figures for the weighted contact maps for the major DRM variation I47V in the initially-opened conformation HIV protease, re-generated using a revised version of the "contact_map.py" script. Edge thickness denotes the contact frequency (shown as edge labels) and indirectly reflects the strength of atomic interactions holding the residues together.

(or GLY C_α) atom, representative of each residue. The C_β atoms are chosen to factor in side-chain information, which is otherwise absent for C_α atoms. Given the distance is satisfied, all contacts are weighed equally, with a value 1 or 0 for their presence or absence respectively. This is stored as an edge list instead of a matrix to limit physical memory consumption. In the initial implementation, the graph was processed via the "igraph" library from the R scripting language. The R dependency was removed and re-written in a later version to facilitate installation and increase portability by replacing it with Python's "NetworkX" library. Doing so enabled ordering of the plotted nodes, such that comparisons of multiple graphs was made easier. This was not possible in the R implementation, which used a random node layout instead. The code was written for strong portability across Python 2.7x and 3.x. Additional options were also included in the most recent version, such as file prefixes, CSV output of the contacts in a compact form and cleaner PDF outputs of the graphs more geared to generate publication-ready figures. Additionally, new graphical parameters were also added such that the user could interact more with the graphical output, for instance the node size, node font size, a scaling factor for the edge width and an edge label font size. Minor optimizations were also done in the new code for increased running efficiency, such as the early reduction of the trajectory size to limit the repetitive reading of large chunks from the trajectory. Adjustable step sizes gives a trade-off between the edge weight accuracy and the speed of network construction. A strong point of this network inference method is its non-requirement for energy minima - the network is basically a summary of all relevant contacts irrespective of conformation, but however may suffer from the contact discontinuity problem. The definition of cut-offs, though taken from literature can be subjective, and hence ignore genuine contacts that are close to but not below the set radius. The weighted contact map is nevertheless a very easy and straightforward approach to hone in on, and interpret local residue behaviour.

6.2.4 Conclusion

The presented tool suite MD-TASK for analysing MD trajectory using non-conventional methods is a very versatile and freely-available set of scripts that facilitates the calculation of network metrics, enables the determination of possible trigger residues responsible for driving conformational changes and implements a DCC method for the investigation of correlative motion from proteins. A demonstration of applicability was shown using MD trajectories from modelled HIV protease structures.

6.3 Side project 3: Analysis of modes of motion: MODE-TASK

This project draws from and reproduces certain figures used in the publication listed below. Credit for the reproduced material is given as citations in the respective figure captions.

- Ross CJ, Nizami B, Glenister M, **Sheik Amamuddy O**, Atilgan AR, Atilgan C and Tastan Bishop Ö "MODE-TASK: Large-scale protein motion tools" *Bioinformatics*, 2018 May 29. doi: 10.1093/bioinformatics/bty427.

My contributions for this work are enumerated in the **Publications and contributions** section, item number 2.

6.3.1 Summary

MODE-TASK was put together by the two main authors (Caroline Ross, a PhD student and Dr Bilal Nizami, a post-doctoral fellow) as a PyMOL plugin "pyMODE-TASK" based on the TkInter graphical user interface package for a series of independent Python scripts implementing code handling modal analysis from proteins [334]. Basically the package implements multiple dimension reduction techniques (Principal Components Analysis, multidimensional scaling and t-Distributed Stochastic Neighbour Embedding) and normal mode analysis (based on the Elastic Network Model). The main functionality of MODE-TASK is thus centered around the application of matrix decomposition techniques onto a Hessian or onto a covariance matrix prior to determining eigenvector and eigenvalue solutions, which will then give normal modes or essential modes, respectively, which both can be used to analyse protein motion. Normal modes are basically a set of orthonormal vectors that breakdown complex movement into a set of independent simpler motions that bear different weights on the observed movement. Large eigenvalues represent the most predominant motions in normal mode analysis and a similar principle is used in essential dynamics to compress a multidimensional array of molecular conformations down to few (usually 2 or 3) dimensions, which can then be represented as scatter plots [217]. While there are many ways of modelling the potential energy to get a Hessian, the Anisotropic Network model, which forms part of the family of elastic network models (Hooke's law for modelling springs), was used to obtain the modes from a single coarse-grained protein structure. Similarly essential dynamics is mainly a visualisation method that tries to represent sampled protein conformations by the first two principal components obtained from a trajectory covariance matrix. As a tester I evaluated functionality, compatibility and recommended changes. The MODE-TASK suite is freely available on github, at the following address: https://github.com/RUBi-ZA/MODE-TASK, and the PyMOL plugin is available here: https://github.com/RUBi-ZA/pyMODE-TASK. Software documentation is also included. As for MD-TASK, the MODE-TASK publication was prepared as an application note to showcase the available tools and demonstrate use cases.

6.3.2 Methods

To run the performance tests, a PC running the Ubuntu 16.04.2 LTS operating system was used with an Intel Core i7-4790 CPU with a processor speed of 3.60GHz with 32GB of Random Access Memory. Software documentation was written by Doctor Bilal Nizami and Caroline Ross (a PhD student at RUBi) and technical assistance was provided by Michael Glenister (software developer at RUBi). Command line codes were run several times in Python 2 and Python 3, checking that the outputs were generated and identical. Errors were reported to the developers, while minor typographical mistakes were corrected. Remote executions were run via SSH connections to the dedicated desktop computer. The whole steps were written as simple bash scripts. The GUI interface (pyMODE-TASK) was tested locally and was found to be limited to Python 2, due to a requirement for available PyMOL version for the operating system via the apt package manager.

Essential dynamics: Tools for dimension reduction were written by Doctor Bilal Nizami. These tools mainly used dimension reduction functionality available from the scikit-learn Python library together with trajectory-handling and manipulation methods from the MDTraj library. Dimension reduction techniques mainly included Principal Components Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE) and Multidimensional scaling (MDS). Each tool was parametrised by various options, including the decomposition method (eg. eigenvalue versus singular value decomposition methods), coordinate type (pairwise distances, bond angles and dihedrals) amongst many more, wrapping several of the actual scikit-learn functionalities. For the purpose of the publication, I tested the performance of the tools using previously generated MD data (100ns;10,000 frames) for the renin-angiotensinogen variant P40L (in angiotensinogen). The tools “pca.py”, “internal_pca.py”, “mds.py” and “tsne.py” were evaluated for performance and bugs.

Normal Mode Analysis: Tools for vibrational analysis from static protein 3D structures based on the anisotropic network model were implemented by Caroline Ross. These may be used sequentially, starting from the coarse-graining of a protein structure, through the calculation of normal modes, to finally analyse and visualise features associated with modes of interest. The scripts were evaluated for performance and bugs. Coarse-graining was performed on the full capsid of the coxsackievirus A16 (CAV-16; PDB ID: 5C4W [335]) to enable faster and cheaper downstream computation of modes, by evenly sampling from surrounding C_β (or GLY C_α) atoms with a coarse-graining level of 4 and starting at residue 3. The “ANM.py” tool was used for extracting vibrational modes from the coarse-grained capsid with a cut-off radius of 50 Å. In order to test the "conformationMode.py" tool, a CAV-16 capsid conformation for the uncoating intermediate (PDB ID: 5C4W) was used to identify any mode with a high degree of association to the conformational change via the calculation of mode overlap and correlation. This mode was finally visualised in VMD using input generated by the tool "visualiseVector.py". Additionally a covariance matrix was calculated for the first non-trivial eigenvector (mode 7) of the coarse-grained capsid using the "assemblyCovariance.py" tool, after which the mean square fluctuation was performed for the same mode to assess the coarse-graining level suitability using the "meanSquareFluctuation.py" tool.

6.3.3 Results and discussions

Essential dynamics: For the purpose of demonstration, standard PCA was performed on 100ns-long MD simulations for each of the 781-residue WT RAS complex and P40L variant, the effects of which were previously described structurally in [204]. As the program used separate reference conformations (first frame from each simulation) for alignment before evaluating the principal components analysis, the plots are not exactly comparable although some initial conformational similarity exists due to the choice of common templates for homology modelling. However, one can see conformational drifts occurring over the course of each simulation. It is possible to make the plots more comparable by aligning all trajectories to a common reference before performing PCA, however this is not directly implemented in MODE-TASK as the first frame is automatically selected for alignment. Times of execution are listed in Table 6.3 and it can be seen that standard PCA works fastest with the default linear kernel and SVD solver using Cartesian coordinates for 10000 MD frames as input.

Normal Mode Analysis: Coarse-graining the full CAV-16 capsid at a level of 4, starting from residue 3 reduced the number of atoms from 399720 to 2460, comprising only C_β and GLY C_α atoms. For ANM construction, the default cut-off of 15 Å was increased to 50 Å to account for more distal effects coming from the large capsid (≈ 300 Å in diameter), while also verifying for the leading 6 non-trivial modes, which correspond to the 3 degrees of freedom for each of the rotational and translational movements. As an example of a use case, the "conformationMode.py" was used to reveal that mode 33 more closely matched the capsid conformation in the process of uncoating (PDB ID: 4JGY) using the previously coarse-grained structure together with the full capsid of the target conformation and the matrix of transposed eigenvectors (V^T) as input. The 3D vector plot was obtained by processing outputs from the "visualiseVector.py" script into VMD for the mode shown in Figure 6.12. Subsequently, the "combinationMode.py" script gives a break down of the mode contributions towards the observed conformational change, which in the case of mode 33 gives overlap values of -0.90, -0.79 and -0.92 for the matching chains A, B and C for the asymmetric unit 1. A covariance matrix was obtained using the "assemblyCovariance.py" tool for the coarse-grained capsid using mode 7, requiring about 20Gb of physical memory to compute. The mean square fluctuation was performed for the same mode to assess the coarse-graining level suitability using the "meanSquareFluctuation.py" tool with coarse-graining levels of 4 and 9. The pyMODE-TASK plugin for use within PyMOL is very straight-forward for evaluating dimension

Table 6.3: Performance for MODE-TASK scripts. Table used from [334].

	Main parameters	Time
Essential dynamics scripts		
pca.py	SVD solver; linear kernel; 10000 frames	26 secs
pca.py	RBF kernel; 10000 frames	40 mins
internal_pca.py	Phi angles; 1 000 frames	10 secs
mds.py	RMSD; 10000 frames	88 mins
tsne.py	10000 frames	68 mins
NMA scripts		
coarseGrain.py	5C4W full capsid; Coarse grain level 4; C_β atoms; starting atom 3	<1s
ANM	5C4W coarse-grained; Cut off 50 Å; 2460 nodes	97 mins
conformationMode.py	4JGY full capsid; 5C4W coarse-grained	26 secs
combinationMode.py	4JGY full capsid; 5C4W coarse-grained	26 secs
visualiseVector.py	5C4W coarse-grained; mode 7	1 secs
assemblyCovariance.py	5C4W coarse-grained; mode 7	317 secs
meanSquareFluctuations.py	4JGY coarse-grain level 9; 5C4W coarse-grained; mode 7	275 secs

reduction methods on an MD trajectory and displays the normal mode tools in a sequential manner, as shown in Figure 6.13.

6.3.4 Conclusion

MODE-TASK provides an open source collection of diverse tools for examining modes of motion from static proteins in the case of normal mode analysis as per anisotropic network model, and also allows a high degree of parametrisation for performing essential dynamics using various methods. The set of command line scripts are convenient for batch and remote execution, while the graphical interface provides a more user-friendly option for the same functionality. Compatibility with both Python 2.7x and 3.x allow for a wider audience of users.

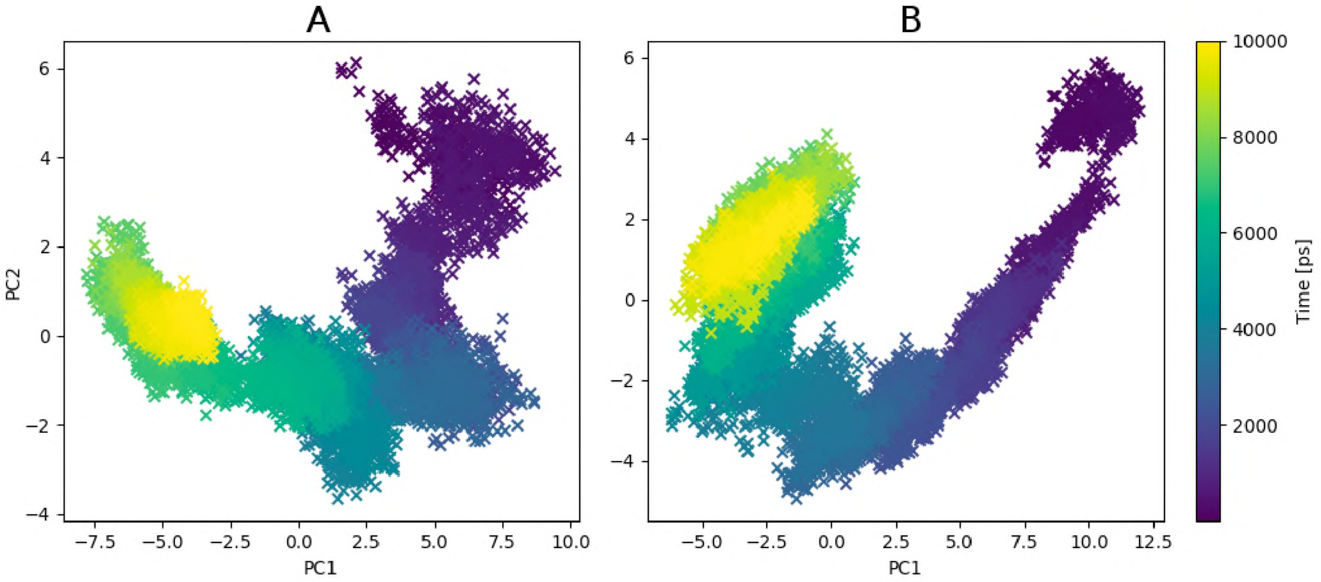


Figure 6.11: Scatter plot of the first two principal components obtained from standard PCA for the WT (A) and mutant (B) RAS complexes, coloured according to time. Colours denote time, in picoseconds. Figure taken from [334].

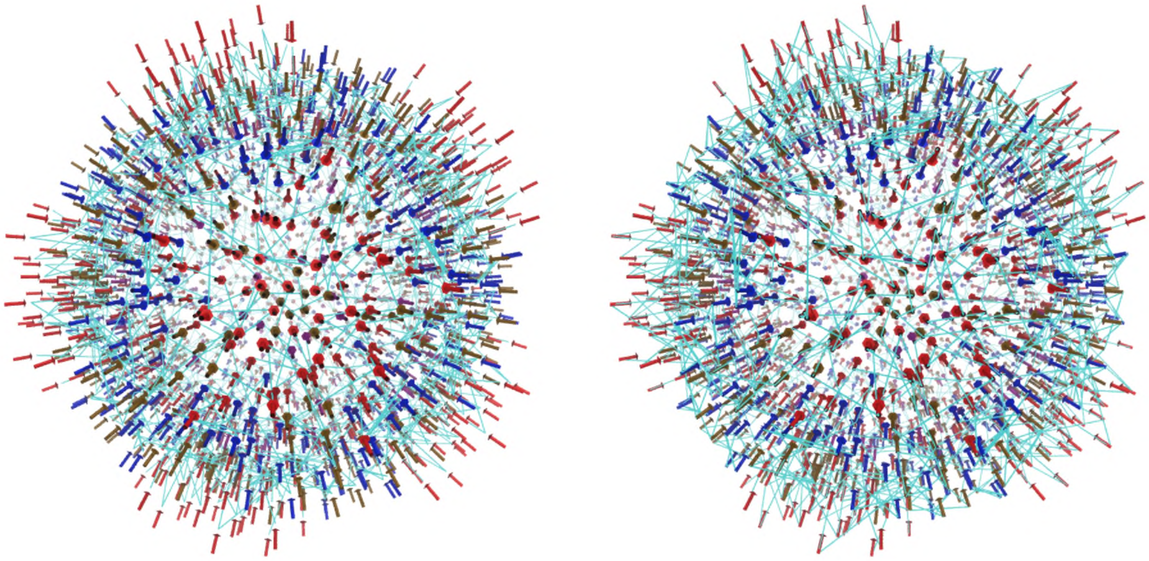
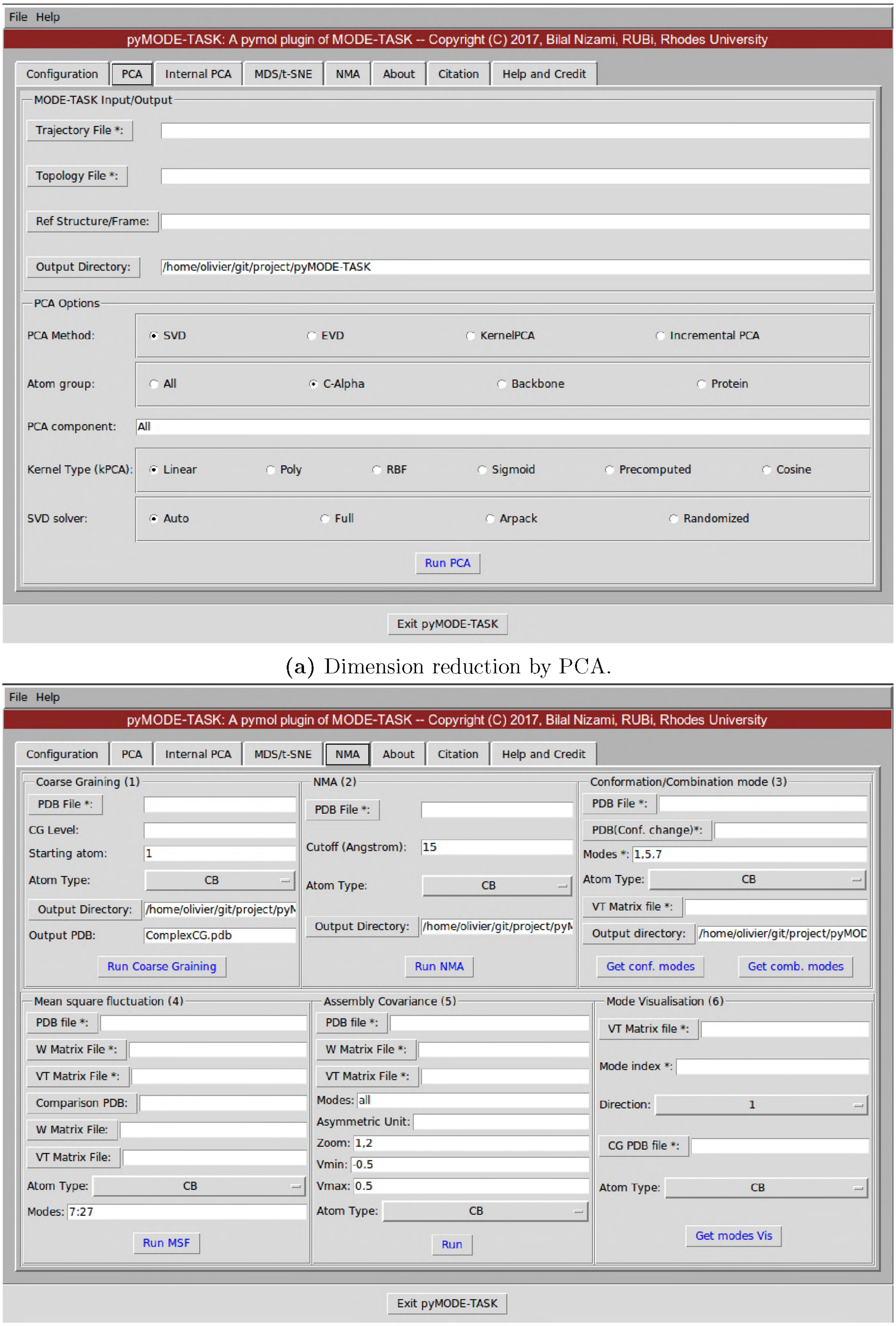


Figure 6.12: Expansion motion associated with the target conformation observed in the capsid uncoating intermediate, shown as the radially-expanded (right) and constricted (left) conformations obtained after stacking 50 frames of mode 33 onto the initial coarse-grained structure. The arrow heads and lengths reflect the direction and unit displacements along the respective components for the mode. Arrows for the viral capsid proteins VP1 (chain A), VP2 (chain B), VP3 (chain C) and VP4 (chain D) are coloured red, blue, ochre and purple respectively. Figure adapted from [334].



File Help
pyMODE-TASK: A pymol plugin of MODE-TASK -- Copyright (C) 2017, Bilal Nizami, RUBi, Rhodes University
Configuration
PCA
Internal PCA
MDS/t-SNE
NMA
About
Citation
Help and Credit
Run PCA
Exit pyMODE-TASK

MODE-TASK Input/Output

Trajectory File *:

Topology File *:

Ref Structure/Frame:

Output Directory: /home/olivier/git/project/pyMODE-TASK

PCA Options

PCA Method: ● SVD ○ EVD ○ KernelPCA ○ Incremental PCA

Atom group: ○ All ● C-Alpha ○ Backbone ○ Protein

PCA component: All

Kernel Type (kPCA): ● Linear ○ Poly ○ RBF ○ Sigmoid ○ Precomputed ○ Cosine

SVD solver: ● Auto ○ Full ○ Arpack ○ Randomized

(a) Dimension reduction by PCA.

Coarse Graining (1)

PDB File *:

CG Level:

Starting atom: 1

Atom Type: CB

Output Directory: /home/olivier/git/project/pyM

Output PDB: ComplexCG.pdb

NMA (2)

PDB File *:

Cutoff (Angstrom): 15

Atom Type: CB

Output Directory: /home/olivier/git/project/pyM

Conformation/Combination mode (3)

PDB File *:

PDB(Conf. change)*:

Modes *: 1,5,7

Atom Type: CB

VT Matrix file *:

Output directory: /home/olivier/git/project/pyMOD

Mean square fluctuation (4)

PDB file *:

W Matrix File *:

VT Matrix File *:

Comparison PDB:

W Matrix File:

VT Matrix File:

Atom Type: CB

Modes: 7:27

Assembly Covariance (5)

PDB file *:

W Matrix File *:

VT Matrix File *:

Modes: all

Asymmetric Unit:

Zoom: 1,2

Vmin: -0.5

Vmax: 0.5

Atom Type: CB

Mode Visualisation (6)

VT Matrix file *:

Mode index *:

Direction: 1

CG PDB file *:

Atom Type: CB

(b) Normal Mode Analysis interface.

Figure 6.13: The pyMODE-TASK graphical user interface. The dimension reduction tools can be used with various parameters, while numbered widgets suggest a flow of execution for normal mode analysis.

General conclusions for Part II

These side projects showcase the development, testing and application of various tools developed in collaboration with members of the RUBi research group for which I have participated to various extents, as enumerated in the **Publications and contributions** section.

In side project 1, the RAS system was investigated via a weighted residue contact network that showed the local differences prevailing at a chosen homologous residue position between a mutant and a wild type complex. This tool provides a new functionality which enables a compact representation of existing residue contacts and provides a non-energetic estimate of their strength via a simple frequency calculation and can be used in other contexts to locally analyse or compare protein systems, and is a useful supplement to traditional MD analysis approaches.

In side project 2, the contact map is generalised and packaged as part of the MD-TASK suite, in which various novel MD analysis tools are also showcased using HIV protease as an example system.

Finally, side project 3 mainly describes my involvement in evaluating the command-line and graphical user interfaces for the MODE-TASK tool kit, developed for a wide audience due to its simplicity, portability and convenience.

Bibliography

[1] Rowan Hatherley, David K Brown, Thommas M Musyoka, David L Penkler, Ngonidzashe Faya, Kevin A Lobb, and Özlem Tastan Bishop. SANCDB: A South African natural compound database. *Journal of Cheminformatics*, 7(1):29, 2015.

[2] Robert C. Gallo and Luc Montagnier. The Discovery of HIV as the Cause of AIDS. *New England Journal of Medicine*, 349(24):2283–2285, dec 2003.

[3] CDC. HIV surveillance–United States, 1981-2008. *Morbidity and mortality weekly report*, 60(21):689–693, 2011.

[4] WHO. Global Health Observatory (GHO) data: HIV/AIDS, 2018.

[5] S Vella. HIV therapy advances. Update on a proteinase inhibitor. *AIDS (London, England)*, 8 Suppl 3:S25–9, sep 1994.

[6] Arun K. Ghosh, Heather L. Osswald, and Gary Prato. Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS. *Journal of Medicinal Chemistry*, 59(11):5172–5208, 2016.

[7] Anna Hake and Nico Pfeifer. Prediction of HIV-1 sensitivity to broadly neutralizing antibodies shows a trend towards resistance over time. *PLoS Computational Biology*, 13(10):1–23, 2017.

[8] Food and Drug Administration. HIV/AIDS Treatment - Antiretroviral drugs used in the treatment of HIV infection, 2018.

[9] Jennifer S Orman and Caroline M Perry. Tipranavir: a review of its use in the management of HIV infection. *Drugs*, 68(10):1435–63, 2008.

[10] R. Bruno, P. Sacchi, L. Maiocchi, S. Patruno, and G. Filice. Hepatotoxicity and antiretroviral therapy with protease inhibitors: A review. *Digestive and Liver Disease*, 38(6):363–373, jun 2006.

[11] José M Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely High Mutation Rate of HIV-1 In Vivo. *PLOS Biology*, 13(9):e1002251, sep 2015.

[12] Jiuping Ji and Lawrence A. Loeb. Fidelity of HIV-1 Reverse Transcriptase Copying RNA in Vitro. *Biochemistry*, 31(4):954–958, 1992.

[13] Andrew E. Armitage, Koen Deforche, Chih-hao Chang, Edmund Wee, Beatrice Kramer, John J. Welch, Jan Gerstoft, Lars Fugger, Andrew McMichael, Andrew Rambaut, and Astrid K. N. Iversen. APOBEC3G-Induced Hypermutation of Human Immunodeficiency Virus Type-1 Is Typically a Discrete All or Nothing Phenomenon. *PLoS Genetics*, 8(3):e1002550, mar 2012.

[14] Mahdis Monajemi, Claire F. Woodworth, Katrin Zipperlen, Maureen Gallant, Michael D. Grant, and Mani Larijani. Positioning of APOBEC3G/F Mutational Hotspots in the Human Immunodeficiency Virus Genome Favors Reduced Recognition by CD8+ T Cells. *PLoS ONE*, 9(4):e93428, apr 2014.

[15] Harold C. Smith. APOBEC3G: A double agent in defense. *Trends in Biochemical Sciences*, 36(5):239–244, 2011.

[16] Suzannah J. Rihn, Joseph Hughes, Sam J. Wilson, and Paul D. Bieniasz. Uneven Genetic Robustness of HIV-1 Integrase. *Journal of Virology*, 89(1):552–567, 2015.

[17] Ashraf Brik and Chi-Huey Wong. HIV-1 protease: mechanism and drug discovery. *Organic & Biomolecular Chemistry*, 1(1):5–14, 2003.

[18] Marie-Pierre de Béthune. Non-nucleoside reverse transcriptase inhibitors (NNRTIs), their discovery, development, and use in the treatment of HIV-1 infection: a review of the last 20 years (1989-2009). *Antiviral research*, 85:75–90, 2010.

[19] Jana Pokorná, Ladislav Machala, Pavlína ezáčová, and Jan Konvalinka. Current and novel inhibitors of HIV protease, 2009.

[20] Yong Wang, Zhigang Liu, Joseph S. Brunzelle, Iulia a. Kovari, Tamaria G. Dewdney, Samuel J. Reiter, and Ladislau C. Kovari. The higher barrier of darunavir and tipranavir resistance for HIV-1 protease. *Biochemical and Biophysical Research Communications*, 412(4):737–742, 2011.

[21] Gerhard (Editor) Klebe. *Drug Design.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[22] Madhavi N L Nalam and Celia A Schiffer. New approaches to HIV protease inhibitor drug design II: Testing the substrate envelope hypothesis to avoid drug resistance and discover robust inhibitors, nov 2008.

[23] Irene T. Weber and Johnson Agniswamy. HIV-1 Protease: Structural Perspectives on Drug Resistance. *Viruses*, 1(3):1110–1136, dec 2009.

[24] D. Paraskevis, E. Kostaki, G. Magiorkinis, P. Gargalianos, G. Xylomenos, E. Magiorkinis, M. Lazanas, M. Chini, G. Nikolopoulos, A. Skoutelis, V. Papastamopoulos, A. Antoniadou, A. Papadopoulos, M. Psichogiou, G.L. Daikos, M. Oikonomopoulou, A. Zavitsanou, G. Chrysos, V. Paparizos, S. Kourkounti, H. Sambatakou, N.V. Sipsas, M. Lada, P. Panagopoulos, E. Maltezos, S. Drimis, and A. Hatzakis. Prevalence of drug resistance

among HIV-1 treatment-naive patients in Greece during 20032015: Transmitted drug resistance is due to onward transmissions. *Infection, Genetics and Evolution*, 54:183–191, oct 2017.

[25] SG Deeks, SR Lewin, and DV Havlir. The End of AIDS: HIV Infection as a Chronic Disease. *Lancet*, 382(9903):1525–1533, 2013.

[26] T.-W. Chun, D Engel, M M Berrey, T Shea, L Corey, and A S Fauci. Early establishment of a pool of latently infected, resting CD4+ T cells during primary HIV-1 infection. *Proceedings of the National Academy of Sciences*, 95(15):8869–8873, jul 1998.

[27] A Boasso, G M Shearer, and C Chougnet. Immune dysregulation in human immunodeficiency virus infection: know it, fix it, prevent it? *Journal of Internal Medicine*, 265(1):78–96, jan 2009.

[28] Erick Wekesa Bunyasi and David John Coetzee. Relationship between socioeconomic status and HIV infection: findings from a survey in the Free State and Western Cape Provinces of South Africa. *BMJ open*, 2017.

[29] B G Turner and M F Summers. Structural biology of HIV. *Journal of molecular biology*, 285:1–32, 1999.

[30] Larry O. Arthur, Julian W. Bess, Raymond C. Sowder, Raoul E. Benveniste, Dean L. Mann, Jean Claude Chermann, and Louis E. Henderson. Cellular proteins bound to immunodeficiency viruses: Implications for pathogenesis and vaccines. *Science*, 1992.

[31] Su Li, Christopher P. Hill, Wesley I. Sundquist, and John T. Finch. Image reconstructions of helical assemblies of the HIV-1 CA protein. *Nature*, 407(6802):409–413, 2000.

[32] Edward M. Campbell and Thomas J. Hope. HIV-1 capsid: the multifaceted key player in HIV-1 infection. *Nature Reviews Microbiology*, 13(8):471–483, 2015.

[33] Francesca Caccuri, Stefania Marsico, Simona Fiorentini, Arnaldo Caruso, and Cinzia Giagulli. HIV-1 Matrix Protein p17 and its Receptors. *Current Drug Targets*, 2015.

[34] Lezelle Botes, Wilhelmina M. J. van den Heever, and Gert H. J. Pretorius. The Structural Biology of HIV. *Medical Technology SA*, 21(June):13–18, 2007.

[35] M. Lapadat-tapolsky, H. De Rocquigny, D. Van Gent, B. Roques, R. Plasterk, and J. L. Darlix. Interactions between HIV-1 nucleocapsid protein and viral DNA may have important functions in the viral life cycle. *Nucleic Acids Research*, 21(8):2024, 1993.

[36] G. Krishnamoorthy, Bernard Roques, Jean Luc Darlix, and Yves Mély. DNA condensation by the nucleocapsid protein of HIV-1: A mechnism ensuring DNA protection. *Nucleic Acids Research*, 31(18):5425–5432, 2003.

[37] Alan D. Frankel and John A. T. Young. HIV-1: Fifteen Proteins and an RNA. *Annual Review of Biochemistry*, 67(1):1–25, 1998.

[38] Melissa Hill, Gilda Tachedjian, and Johnson Mak. The packaging and maturation of the HIV-1 Pol proteins. *Current HIV research*, 3:73–85, 2005.

[39] Klaus Strebel. HIV accessory proteins versus host restriction factors. *Current Opinion in Virology*, 3(6):692–699, 2013.

[40] Frank Kirchhoff. Immune evasion and counteraction of restriction factors by HIV-1 and other primate lentiviruses, 2010.

[41] Steven G. Deeks, Julie Overbaugh, Andrew Phillips, and Susan Buchbinder. HIV infection. *Nature Reviews Disease Primers*, 1(October), 2015.

[42] Alan Engelman and Peter Cherepanov. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nature Reviews Microbiology*, 10(4):279–290, 2012.

[43] Stefan G. Sarafianos, Bruno Marchand, Kalyan Das, Daniel M. Himmel, Michael A. Parniak, Stephen H. Hughes, and Eddy Arnold. Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition, jan 2009.

[44] Olivier Delelis, Kevin Carayon, Ali Saïb, Eric Deprez, and Jean-François Mouscadet. Integrase and integration: biochemical activities of HIV-1 integrase. *Retrovirology*, 5:114, dec 2008.

[45] Marco Salemi, Tulio De Oliveira, Marcelo A. Soares, Oliver Pybus, Ana T. Dumans, Anne Mieke Vandamme, Amilcar Tanuri, Sharon Cassol, and Walter M. Fitch. Different epidemic potentials of the HIV-1B and C subtypes. *Journal of Molecular Evolution*, 60:598–605, 2005.

[46] Eduardo Castro-Nallar, Marcos Pérez-Losada, Gregory F. Burton, and Keith a. Crandall. The evolution of HIV: Inferences using phylogenetics. *Molecular Phylogenetics and Evolution*, 62(2):777–792, 2012.

[47] Paulo C.C. Dos Santos, Helder F.S. Lopes, Rosana Alcalde, Cláudio R. Gonsalez, Jair M. Abe, and Luis F. Lopez. Paraconsistents artificial neural networks applied to the study of mutational patterns of the F subtype of the viral strains of HIV-1 to antiretroviral therapy. *Anais da Academia Brasileira de Ciencias*, 88(1):323–334, 2016.

[48] Ana B Abecasis, Annemarie MJ J Wensing, Dimitris Paraskevis, Jurgen Vercauteren, Kristof Theys, David AMC M C Van de Vijver, Jan Albert, Birgitta Asjö, Claudia Balotta, Danail Beshkov, Ricardo J Camacho, Bonaventura Clotet, Cillian De Gascun, Algis Griskevicius, Zehava Grossman, Osamah Hamouda, Andrzej Horban, Tatjana Kolupajeva, Klaus Korn, Leon G Kostrikis, Claudia Kücherer, Kirsi Liitsola, Marek Linka, Claus Nielsen, Dan Otelea, Roger Paredes, Mario Poljak, Elisabeth Puchhammer-Stöckl, Jean-Claude Schmit, Anders Sönnerborg, Danika Stanekova, Maja Stanojevic, Daniel Struck, Charles AB B Boucher, and Anne-Mieke Vandamme. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology*, 10(1):7, jan 2013.

[49] Joris Hemelaar. The origin and diversity of the HIV-1 pandemic. *Trends in Molecular Medicine*, 18(3):182–192, 2012.

[50] P. M. Sharp and B.H. Hahn. Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspectives in Medicine*, 1(1), 2011.

[51] D. R. Matute. The role of founder effects on the evolution of reproductive isolation. *Journal of Evolutionary Biology*, 26(11):2299–2311, 2013.

[52] Bette Korber, Brian Gaschen, Karina Yusim, Rama Thakallapally, Can Kesmir, and Vincent Detours. Evolutionary and immunological implications of contemporary HIV-1 variation. *British Medical Bulletin*, 58(August):19–42, 2001.

[53] D L Robertson, J P Anderson, J A Bradac, J K Carr, B Foley, R K Funkhouser, F Gao, B H Hahn, M L Kalish, C Kuiken, G H Learn, T Leitner, F McCutchan, S Osmanov, M Peeters, D Pieniazek, M Salminen, P M Sharp, S Wolinsky, and B Korber. HIV-1 nomenclature proposal. *Science (New York, N.Y.)*, 288(5463):55–6, apr 2000.

[54] Katherine A Lau and Justin J.L. Wong. Current trends of HIV recombination worldwide. *Infectious Disease Reports*, 5(1S):4, jun 2013.

[55] Hongshuo Song, Elena E. Giorgi, Vitaly V. Ganusov, Fangping Cai, Gayathri Athreya, Hyejin Yoon, Oana Carja, Bhavna Hora, Peter Hraber, Ethan Romero-Severson, Chunlai Jiang, Xiaojun Li, Shuyi Wang, Hui Li, Jesus F. Salazar-Gonzalez, Maria G. Salazar, Nilu Goonetilleke, Brandon F. Keele, David C. Montefiori, Myron S. Cohen, George M. Shaw, Beatrice H. Hahn, Andrew J. McMichael, Barton F. Haynes, Bette Korber, Tanmoy Bhattacharya, and Feng Gao. Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nature Communications*, 9(1):1928, dec 2018.

[56] Joost Louwagie, Francine E. McCutchan, Martine Peeters, Terrence P. Brennan, Eric Sanders-Buell, Gerald A. Eddy, Guido Van Der Groen, Katrien Fransen, Guy Michel Gershy-Damet, Robert Deleys, and Donald S. Burke. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS*, 1993.

[57] L G Kostrikis, E Bagdades, Y Cao, L Zhang, D Dimitriou, and D D Ho. Genetic analysis of human immunodeficiency virus type 1 strains from patients in Cyprus: identification of a new subtype designated subtype I. *Journal of virology*, 1995.

[58] T Leitner, A Alaeus, S Marquina, E Lilja, K Lidman, and J Albert. Yet another subtype of HIV type 1? *AIDS research and human retroviruses*, 11(8):995–7, aug 1995.

[59] Karine Triques, Anke Bourgeois, Sentob Saragosti, Nicole Vidal, Eitel Mpoudi-Ngole, Nzila Nzilambi, Christian Apetrei, Michel Ekwalanga, Eric Delaporte, and Martine Peeters. High diversity of HIV-1 subtype F strains in Central Africa. *Virology*, 1999.

[60] K Triques, A Bourgeois, N Vidal, E Mpoudi-Ngole, C Mulanga-Kabeya, N Nzilambi, N Torimiro, E Saman, E Delaporte, and M Peeters. Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. *AIDS Res Hum Retroviruses*, 2000.

[61] C. Pasquier, N. Millot, R. Njouom, K. Sandres, M. Cazabat, J. Puel, and J. Izopet. HIV-1 subtyping using phylogenetic analysis of pol gene sequences. *Journal of Virological Methods*, 94(1-2):45–54, 2001.

[62] Lavinia Fabeni, Giulia Berno, Joseph Fokam, Ada Bertoli, Claudia Alteri, Caterina Gori, Federica Forbici, Desiré Takou, Alessandra Vergori, Mauro Zaccarelli, Gaetano Maffongelli, Vanni Borghi, Alessandra Latini, Alfredo Pennica, Claudio Maria Mastroianni, Francesco Montella, Cristina Mussini, Massimo Andreoni, Andrea Antinori, Carlo Federico Perno, Maria Mercedes Santoro, Daniele Armenia, Maria Concetta Bellocchi, Andrea Biddittu, Massimiliano Bruni, Anna Rita Buonomini, Luca Carioti, Francesca Ceccherini-Silberstein, Carlotta Cerva, Novella Cesta, Domenico Di Carlo, Luca Dori, Luca Foroghi, Elisa Gentilotti, Sara Giannella, Tania Guenci, Vincenzo Malagnino, Alessandra Ricciardi, Marzia Romani, Omina Salpini, Loredana Sarmati, Rossana Scutari, Valentina Serafini, Pasquale Sordillo, Francesca Stazi, Cristof Stingone, Valentina Svicher, Elisabetta Teti, Magdalena Viscione, Isabella Abbate, Rosa Acinapura, Lucia Alba, Adriana Ammassari, Franco Baldini, Rita Bellagamba, Evangelo Boumis, Maria Rosaria Capobianchi, Stefania Carta, Stefania Cicalini, Fabio Continenza, Gabriella De Carli, Roberta D'Arrigo, Gianpiero D'Offizi, Valentina Fedele, Vincenzo Galati, Alberto Giannetti, Enrico Girardi, Susanna Grisetti, Raffaella Libertone, Giuseppina Liuzzi, Patrizia Lorenzini, Rita Maddaluno, Andrea Mariano, Assunta Navarra, Emanuele Nicastri, Giuseppina Nurra, Nicoletta Orchi, Antonio Palummieri, Carmela Pinnetti, Silvia Pittalis, Daniele Pizzi, Vincenzo Puro, Alessandro Sampaolesi, Maria Rosaria Sciarrone, Paola Scognamiglio, Catia Sias, Ubaldo Visco-Comandini, Manuela Colafigli, Antonio Cristaudo, Massimo Giuliani, Anna Pacifici, Fiorella Di Sora, Filippo Iebba, Miriam Lichtner, Raffaella Marocco, Stefania Bernardi, Enza Anzalone, Maria Elena Bonaventura, Mauro Marchili, Antonella Pitorri, Luigi Falconi Di Francesco, Dante Di Giammartino, Antonio Caterini, Orlando Armignacco, Rinalda Mariani, Maurizio Paoloni, Giustino Parruti, Alessandro Pieri, Federica Sozio, Antonio Cellini, Alessandro Grimaldi, Maurizio Mariani, Giovanna Picchi, William Gennari, Aubin J. Nanfack, Alexis Ndjolo, and Judith N. Torimiro. Comparative evaluation of subtyping tools for surveillance of newly emerging HIV-1 strains. *Journal of Clinical Microbiology*, 55(9):2827–2837, 2017.

[63] David L. Robertson, Paul M. Sharp, Francine E. Mc Cutchan, and Beatrice H. Hahn. Recombination in HIV-1. *Nature*, 1995.

[64] Luis Menéndez-Arias. Molecular basis of human immunodeficiency virus drug resistance: An update. *Antiviral Research*, 85:210–231, 2010.

[65] Yvonne Geiß and Ursula Dietrich. Catch Me If You Can The Race Between HIV and Neutralizing Antibodies. *AIDS Reviews*, pages 107–113, 2015.

[66] Gang Wang, Na Zhao, Ben Berkhout, and Atze T. Das. CRISPR-Cas based antiviral strategies against HIV-1. *Virus Research*, 244(May 2017):321–332, 2018.

[67] Weijun Zhu, Rongyue Lei, Yann Le Duff, Jian Li, Fei Guo, Mark A Wainberg, and Chen Liang. The CRISPR/Cas9 system inactivates latent HIV-1 proviral DNA. *Retrovirology*, 12(1):22, 2015.

[68] Youdiil Ophinni, Mari Inoue, Tomohiro Kotaki, and Masanori Kameoka. CRISPR / Cas9 system targeting regulatory genes of HIV-1 inhibits viral replication in infected T-cell cultures. *Scientific Reports*, 8(November 2017):1–12, dec 2018.

[69] Janet L. Paulsen, Florian Leidner, Debra A. Ragland, Nese Kurt Yilmaz, and Celia A. Schiffer. Interdependence of Inhibitor Recognition in HIV-1 Protease. *Journal of Chemical Theory and Computation*, 13(5):2300–2309, 2017.

[70] Alice K Pau and Jomy M George. Antiretroviral therapy: Current drugs, sep 2014.

[71] Michele W. Tang and Robert W. Shafer. HIV-1 antiretroviral resistance: Scientific principles and clinical applications. *Drugs*, 72(9):1–25, 2012.

[72] Lianhong Xu, Hongtao Liu, Bernard P. Murray, Christian Callebaut, Melody S. Lee, Allen Hong, Robert G. Strickley, Luong K. Tsai, Kirsten M. Stray, Yujin Wang, Gerry R. Rhodes, and Manoj C. Desai. Cobicistat (GS-9350): A potent and selective inhibitor of human CYP3A as a novel pharmacoenhancer. *ACS Medicinal Chemistry Letters*, 1(5):209–213, 2010.

[73] Christopher F. Rowley. Developments in CD4 and viral load monitoring in resource-limited settings. *Clinical Infectious Diseases*, 58(3):407–412, 2014.

[74] Soo-Yon Rhee, Michael R. Jordan, Elliot Raizes, Arlene Chua, Neil Parkin, Rami Kantor, Gert U. Van Zyl, Irene Mukui, Mina C. Hosseinipour, Lisa M. Frenkel, Nicaise Ndembi, Raph L. Hamers, Tobias F. Rinke de Wit, Carole L. Wallis, Ravindra K. Gupta, Joseph Fokam, Clement Zeh, Jonathan M. Schapiro, Sergio Carmona, David Katzenstein, Michele Tang, Avelin F. Aghokeng, Tulio De Oliveira, Annemarie M. J. Wensing, Joel E. Gallant, Mark A. Wainberg, Douglas D. Richman, Joseph E. Fitzgibbon, Marco Schito, Silvia Bertagnolio, Chunfu Yang, and Robert W. Shafer. HIV-1 Drug Resistance Mutations : Potential Applications for Point-of-Care Genotypic Resistance Testing. *PLoS ONE*, 10(12):1–17, 2015.

[75] Omar Sued, María Inés Figueroa, and Pedro Cahn. Clinical challenges in HIV/AIDS: Hints for advancing prevention and patient management strategies. *Advanced Drug Delivery Reviews*, 103:5–19, 2016.

[76] Kazuhisa Yoshimura. Current status of HIV/AIDS in the ART era. *Journal of Infection and Chemotherapy*, 23(1):12–16, 2017.

[77] Victoria A Johnson, Vincent Calvez, Huldrych F Gunthard, Roger Paredes, Deenan Pillay, Robert W Shafer, Annemarie M Wensing, and Douglas D Richman. Update of the drug resistance mutations in HIV-1: March 2013. *Top Antivir Med*, 21(1):6–14, 2013.

[78] Annemarie M Wensing, Vincent Calvez, Huldrych F Günthard, Victoria A Johnson, Roger Paredes, Deenan Pillay, Robert W Shafer, and Douglas D Richman. 2014 update of the drug resistance mutations in HIV-1. *Topics in Antiviral Medicine*, 22(3):642–650, 2014.

[79] Annemarie M. Wensing, Vincent Calvez;, Huldrych F. Günthard, Victoria A. Johnson, Roger Paredes, Deenan Pillay, Robert W. Shafer, and Douglas D. Richman. 2015 Update of the Drug Resistance Mutations in HIV-1. *Topics in antiviral medicine*, pages 132–141, 2015.

[80] AM Wensing, V Calvez, HF Gunthard, VA Johnson, R Paredes, D Pillay, RW Shafer, and DD Richman. 2017 Update of the Drug Resistance Mutations in HIV-1. *Top Antivir Med*, 24(4):132–133, 2017.

[81] Benson Edagwa, Jo Ellyn McMillan, Brady Sillman, and Howard E. Gendelman. Long-acting slow effective release antiretroviral therapy. *Expert Opinion on Drug Delivery*, 14(11):1281–1291, 2017.

[82] James M. McMahon, Julie E. Myers, Ann E. Kurth, Stephanie E. Cohen, Sharon B. Mannheimer, Janie Simmons, Enrique R. Pouget, Nicole Trabold, and Jessica E. Haberer. Oral Pre-Exposure Prophylaxis (PrEP) for Prevention of HIV in Serodiscordant Heterosexual Couples in the United States: Opportunities and Challenges. *AIDS Patient Care and STDs*, 28(9):462–474, 2014.

[83] Soo-Yon Rhee, Matthew J Gonzales, Rami Kantor, Bradley J Betts, Jaideep Ravela, and Robert W Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic acids research*, 31(1):298–303, 2003.

[84] Sarah Wagner, Mario Kurz, and Thomas Klimkait. Algorithm evolution for drug resistance prediction: Comparison of systems for HIV-1 genotyping. *Antiviral Therapy*, 20(6):661–665, 2015.

[85] Fabio Pietrucci, Attilio Vittorio Vargiu, and Agata Kranjc. HIV-1 Protease Dimerization Dynamics Reveals a Transient Druggable Binding Pocket at the Interface. *Scientific Reports*, 5(1):18555, nov 2016.

[86] Samuel Broder. The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic. *Antiviral Research*, 85(1):1–18, 2010.

[87] AIDSinfo. AIDSinfo — Drugs, 2018.

[88] Food and Drug Administration. Fuzeon (Product Label), 2013.

[89] Food and Drug Administration. Ibalizumab (Product Label), 2018.

[90] A. M. J. Wensing, Noortje M. van Maarseveen, and Monique Nijhuis. Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Research*, 85:59–74, 2010.

[91] Ming-Hua Chen, Shan-Shan Chang, Biao Dong, Li-Yan Yu, Ye-Xiang Wu, Ren-Zhong Wang, Wei Jiang, Zeng-Ping Gao, and Shu-Yi Si. Ahmpatinin i Bu, a new HIV-1 protease inhibitor, from Streptomyces sp. CPCC 202950. *RSC Advances*, 8(10):5138–5144, 2018.

[92] M A Navia, P M Fitzgerald, B M McKeever, C T Leu, J C Heimbach, W K Herber, I S Sigal, P L Darke, and J P Springer. Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature*, 337(6208):615–620, 1989.

[93] Joanna Trylska, Valentina Tozzini, Chia-en A Chang, and J Andrew McCammon. HIV-1 protease substrate binding and product release pathways explored with coarse-grained molecular dynamics. *Biophysical journal*, 92(12):4179–87, jun 2007.

[94] Robert D. Finn, Teresa K. Attwood, Patricia C. Babbitt, Alex Bateman, Peer Bork, Alan J. Bridge, Hsin Yu Chang, Zsuzsanna Dosztanyi, Sara El-Gebali, Matthew Fraser, Julian Gough, David Haft, Gemma L. Holliday, Hongzhan Huang, Xiaosong Huang, Ivica Letunic, Rodrigo Lopez, Shennan Lu, Aron Marchler-Bauer, Huaiyu Mi, Jaina Mistry, Darren A. Natale, Marco Necci, Gift Nuka, Christine A. Orengo, Youngmi Park, Sebastien Pesseat, Damiano Piovesan, Simon C. Potter, Neil D. Rawlings, Nicole Redaschi, Lorna Richardson, Catherine Rivoire, Amaia Sangrador-Vegas, Christian Sigrist, Ian Sillitoe, Ben Smithers, Silvano Squizzato, Granger Sutton, Narmada Thanki, Paul D. Thomas, Silvio C.E. Tosatto, Cathy H. Wu, Ioannis Xenarios, Lai Su Yeh, Siew Yit Young, and Alex L. Mitchell. InterPro in 2017 - beyond protein family and domain annotations. *Nucleic Acids Research*, 2017.

[95] Irene T Weber, Daniel W Kneller, and Andres Wong-Sam. Highly resistant HIV-1 proteases and strategies for their inhibition. *Future medicinal chemistry*, 7(8):1023–38, 2015.

[96] Moses Prabu-Jeyabalan, Ellen Nalivaika, and Celia A. Schiffer. Substrate shape determines specificity of recognition for HIV-1 protease: Analysis of crystal structures of six substrate complexes. *Structure*, 10:369–381, 2002.

[97] A Wlodawer, M Miller, M Jaskólski, B K Sathyanarayana, E Baldwin, I T Weber, L M Selk, L Clawson, J Schneider, and S B Kent. Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science (New York, N.Y.)*, 245(4918):616–621, aug 1989.

[98] Natalie L. Dawson, Tony E. Lewis, Sayoni Das, Jonathan G. Lees, David Lee, Paul Ashford, Christine A. Orengo, and Ian Sillitoe. CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, 2017.

[99] Olivier Sheik Amamuddy, Nigel T. Bishop, and Özlem Tastan Bishop. Improving fold resistance prediction of HIV-1 against protease and reverse transcriptase inhibitors using artificial neural networks. *BMC bioinformatics*, 18(1):369, aug 2017.

[100] Aleksejs Kontijevskis, Jarl E. S. Wikberg, and Jan Komorowski. Computational proteomics analysis of HIV-1 protease interactome. *Proteins: Structure, Function, and Bioinformatics*, 68(1):305–312, apr 2007.

[101] Steve C. Pettit, Scott F. Michael, and Ronald Swanstrom. The specificity of the HIV-1 protease. *Perspectives in Drug Discovery and Design*, 1:69–83, 1993.

[102] Uwe Lendeckel and Nigel M Hooper. *Viral Proteases and Antiviral Protease Inhibitor Therapy: Proteases in Biology and Disease*, volume 8. Springer Science & Business Media, 2009.

[103] József Tözsér and József. Comparative studies on retroviral proteases: Substrate specificity. *Viruses*, 2(1):147–165, jan 2010.

[104] Warner C. Greene, Zeger Debyser, Yasuhiro Ikeda, Eric O. Freed, Edward Stephens, Wes Yonemoto, Robert W. Buckheit, José a. Esté, and Tomas Cihlar. Novel targets for HIV therapy. *Antiviral Research*, 80:251–265, 2008.

[105] John Randolph and David DeGoey. Peptidomimetic Inhibitors of HIV Protease. *Current Topics in Medicinal Chemistry*, 4(10):1079–1095, 2004.

[106] Mukesh M Mudgal, Nagaraju Birudukota, and Mayur A Doke. Applications of Click Chemistry in the Development of HIV Protease Inhibitors. *International Journal of Medicinal Chemistry*, 2018:1–9, jul 2018.

[107] Nanjie Deng, Stefano Forli, Peng He, Alex Perryman, Lauren Wickstrom, R. S.K. Vijayan, Theresa Tiefenbrunn, David Stout, Emilio Gallicchio, Arthur J. Olson, and Ronald M. Levy. Distinguishing binders from false positives by free energy calculations: Fragment screening against the flap site of HIV protease. *Journal of Physical Chemistry B*, 119(3):976–988, 2015.

[108] S. Spinelli, Q. Z. Liu, P. M. Alzari, P. H. Hirel, and R. J. Poljak. The three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU. *Biochimie*, 1991.

[109] Philip Martin, John F. Vickrey, Gheorghe Proteasa, Yurytzy L. Jimenez, Zdzislaw Wawrzak, Mark A. Winters, Thomas C. Merigan, and Ladislau C. Kovari. Wide-Open" 1.3 A Structure of a Multidrug-Resistant HIV-1 Protease as a Drug Target. *Structure*, 13:1887–1895, 2005.

[110] Yuqi Yu, Jinan Wang, Zhaoqiang Chen, Guimin Wang, Qiang Shao, Jiye Shi, and Weiliang Zhu. Structural insights into HIV-1 protease flap opening processes and key intermediates. *RSC Advances*, 7(71):45121–45128, 2017.

[111] Youla S. Tsantrizos. Peptidomimetic therapeutic agents targeting the protease enzyme of the human immunodeficiency virus and hepatitis C virus. *Accounts of Chemical Research*, 41(10):1252–1263, 2008.

[112] M. N. L. Nalam, A. Peeters, T. H. M. Jonckers, I. Dierynck, and C. A. Schiffer. Crystal Structure of Lysine Sulfonamide Inhibitor Reveals the Displacement of the Conserved Flap

Water Molecule in Human Immunodeficiency Virus Type 1 Protease. *Journal of Virology*, 81(17):9512–9518, 2007.

[113] Mariusz Jaskólski, Alexander Wlodawer, Alfredo G. Tomasselli, Tomi K. Sawyer, Douglas G. Staples, Robert L. Heinrikson, Jens Schneider, Stephen B.H. Kent, and Stephen B.H. Kent. Structure at 2.5-Å Resolution of Chemically Synthesized Human Immunodeficiency Virus Type 1 Protease Complexed with a Hydroxyethylene-Based Inhibitor. *Biochemistry*, 30(6):1600–1609, 1991.

[114] Lin Li, Lili Chen, Shaomin Yang, Tianyi Li, Jianjian Li, Yongjian Liu, Lei Jia, Bihui Yang, Zuoyi Bao, Hanping Li, Xiaolin Wang, Daomin Zhuang, Siyang Liu, and Jingyun Li. Recombination Form and Epidemiology of HIV-1 Unique Recombinant Strains Identified in Yunnan, China. *PLoS ONE*, 7(10), 2012.

[115] Shuntai Zhou, Corbin Jones, Piotr Mieczkowski, and Ronald Swanstrom. Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next-Generation Sequencing of HIV-1 Genomic RNA Populations. *Journal of Virology*, 89(16):8540–8555, 2015.

[116] Soyeon Ahn, Ziqi Ke, and Haris Vikalo. Viral quasispecies reconstruction via tensor factorization with successive read removal. *Bioinformatics*, 34(13):i23–i31, 2018.

[117] S.-L. Liu, A. G. Rodrigo, R. Shankarappa, G. H. Learn, L. Hsu, O. Davidov, L. P. Zhao, and J. I. Mullins. HIV Quasispecies and Resampling. *Science*, 273(5274):415–416, jul 1996.

[118] AIDSinfo. AIDSinfo Glossary of HIV/AIDS-Related Terms, 2015.

[119] Niko Beerenwinkel, Barbara Schmidt, Hauke Walter, Rolf Kaiser, Thomas Lengauer, Daniel Hoffmann, Klaus Korn, Joachim Selbig, Barbara Schmidt, Hauke Walter, Klaus Korn, Rolf Kaiser, and Daniel Hoffmann. Geno2pheno: Interpreting genotypic HIV drug resistance tests. *IEEE Intelligent Systems and Their Applications*, 16(December):35–41, 2001.

[120] Andrea Clemencia Pineda-Peña, Nuno Rodrigues Faria, Stijn Imbrechts, Pieter Libin, Ana Barroso Abecasis, Koen Deforche, Arley Gómez-López, Ricardo J. Camacho, Tulio De Oliveira, and Anne Mieke Vandamme. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. *Infection, Genetics and Evolution*, 2013.

[121] Mona Riemenschneider, Thomas Hummel, and Dominik Heider. SHIVA - a web application for drug resistance and tropism testing in HIV. *BMC Bioinformatics*, 17(1):314, dec 2016.

[122] Niko Beerenwinkel, Martin Däumer, Mark Oette, Klaus Korn, Daniel Hoffmann, Rolf Kaiser, Thomas Lengauer, Joachim Selbig, and Hauke Walter. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Research*, 31(13):3850–3855, 2003.

[123] Jurgen Vercauteren and Anne Mieke Vandamme. Algorithms for the interpretation of HIV-1 genotypic drug resistance information. *Antiviral Research*, 71(2-3 SPEC. ISS.):335–342, 2006.

[124] B. Masquelier, F. Brun-Vezinet, D. Descamps, V. Calvez, A. Ruffault, G. Peytavin, M. Vray, L. Morand-Joubert, F. Telles, D. Costagliola, J-L Meyard, and Narval (ANRS 088) Study Group. Clinically relevant interpretation of genotype for resistance to abacavir. *Aids*, 17(12):1795–1802, 2003.

[125] Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, may 1982.

[126] Los Alamos National Laboratory. HIV Sequence Database: HIV and SIV Nomenclature, 2012.

[127] ChenHsiang Shen, Xiaxia Yu, Robert W. Harrison, and Irene T. Weber. Automated prediction of HIV drug resistance from genotype data. *BMC Bioinformatics*, 17(S8):278, aug 2016.

[128] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2012.

[129] Fran\c{c}ois and others Chollet. Keras, 2015.

[130] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, and Google Brain. TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, pages 265–284, 2016.

[131] Adam Paszke, Adam and Gross, Sam and Chintala, Soumith and Chanan, Gregory and Yang, Edward and DeVito, Zachary and Lin, Zeming and Desmaison, Alban and Antiga, Luca and Lerer. Automatic differentiation in PyTorch, 2017.

[132] Max Kuhn. caret Package. *Journal Of Statistical Software*, 2008.

[133] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[134] Rishabh Choudhary and Hemant Kumar Gianey. Comprehensive Review On Supervised Machine Learning Algorithms. In *2017 International Conference on Machine Learning and Data Science (MLDS)*, volume 2018-Janua, pages 37–43. IEEE, dec 2017.

[135] Rachid Darnag, Brahim Minaoui, and Mohamed Fakir. QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression. *Arabian Journal of Chemistry*, 10:S600–S608, 2017.

[136] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks CORINNA. *Chemical biology & drug design*, 20(3):273–297, aug 1995.

[137] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer New York, New York, NY, 2013.

[138] Trevor; Hastie. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

[139] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579—-2605, 2008.

[140] Martin Ester, Hans-Peter Kriegel, Jörg Sander, , and Xiaowei others Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226—-231, 1996.

[141] Nadeem Tariq. Breast Cancer Detection using Artificial Neural Networks. *Journal of Molecular Biomarkers & Diagnosis*, 09(01):1–6, 2018.

[142] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalena Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1124.e9, 2018.

[143] Paras Lakhani and Baskaran Sundaram. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, 284(2):574–582, 2017.

[144] Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4):290–298, 2018.

[145] Isis Bonet, María M García, Yvan Saeys, Yves Van de Peer, and Ricardo Grau. Predicting Human Immunodeficiency Virus (HIV) Drug Resistance Using Recurrent Neural Networks. Number November, pages 234–243. 2007.

[146] Gary B. Fogel, Susanna L. Lamers, Enoch S. Liu, Marco Salemi, and Michael S. McGrath. Identification of dual-tropic HIV-1 using evolved neural networks. *Biosystems*, 137:12–19, 2015.

[147] Ben Otange, Zephania Birech, Ronald Rop, and Julius Oyugi. Estimation of HIV-1 viral load in plasma of HIV-1-infected people based on the associated Raman spectroscopic peaks. *Journal of Raman Spectroscopy*, pages 1–9, jan 2019.

[148] Ashok Kumar Dwivedi and Usha Chouhan. Multilayer perceptron and evolutionary radial basis function neural network models for discrimination of HIV-1 genomes. *Current Science*, 115(11), 2018.

[149] Xinyu Lu, Lifang Wang, and Zejun Jiang. The Application of Deep Learning in the Prediction of HIV-1 Protease Cleavage Site. In *2018 5th International Conference on Systems and Informatics (ICSAI)*, number Icsai, pages 1299–1304. IEEE, nov 2018.

[150] Hailin Hu, An Xiao, Sai Zhang, Yangyang Li, Xuanling Shi, Tao Jiang, Linqi Zhang, Lei Zhang, and Jianyang Zeng. DeepHINT: understanding HIV-1 integration via deep learning with attention. *Bioinformatics*, pages 1–16, oct 2018.

[151] Abolfazl Barzegar, Elham Zamani-Gharehchamani, and Ali Kadkhodaie-Ilkhchi. ANN QSAR workflow for predicting the inhibition of HIV-1 reverse transcriptase by pyridinone non-nucleoside derivatives. *Future Medicinal Chemistry*, 9(11):1175–1191, jul 2017.

[152] Ctlin Buiu, Mihai V. Putz, and Speranta Avram. Learning the relationship between the primary structure of HIV envelope glycoproteins and neutralization activity of particular antibodies by using artificial neural networks. *International Journal of Molecular Sciences*, 17(10):1–14, 2016.

[153] Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

[154] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, pages 1–14, 2016.

[155] Ning Qian. On the momentum term in gradient descent learning algorithms, 1999.

[156] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 2011.

[157] Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *arXiv*, 2012.

[158] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations 2015*, 2015.

[159] Patrick Winston. 6.034 Artificial Intelligence. Fall 2010. Massachusetts Institute of Technology, 2010.

[160] Seare Tesfamichael Araya and Scott Hazelhurst. Support vector machine prediction of HIV-1 drug resistance using the viral nucleotide patterns. *Transactions of the Royal Society of South Africa*, 64(1):62–72, 2009.

[161] Jaideep Ravela, Bradley J Betts, Francoise Brun-Vézinet, Anne-Mieke Vandamme, Diane Descamps, Kristel van Laethem, Kate Smith, Jonathan M Schapiro, Dean L Winslow, Caroline Reid, and Robert W Shafer. HIV-1 protease and reverse transcriptase mutation patterns responsible for discordances between genotypic drug resistance interpretation algorithms. *Journal of acquired immune deficiency syndromes (1999)*, 33(1):8–14, 2003.

[162] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, jul 1944.

[163] Donald W. Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, jun 1963.

[164] Peter Rice, Ian Longden, and Alan Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6):276–277, 2000.

[165] Haley Hedlin. Genotype-Phenotype Datasets: DRMcv, 2014.

[166] RJ Lessells, DK Katzenstein, and T de Oliveira. Are subtype differences important in HIV drug resistance? *Current Opinion in Virology*, 2(5):636–643, oct 2012.

[167] Kamalendra Singh, Jacqueline A. Flores, Karen A. Kirby, Ujjwal Neogi, Anders Sonnerborg, Atsuko Hachiya, Kalyan Das, Eddy Arnold, Carole McArthur, Michael Parniak, and Stefan G. Sarafianos. Drug resistance in non-B subtype HIV-1: Impact of HIV-1 reverse transcriptase inhibitors. *Viruses*, 6(9):3535–3562, 2014.

[168] Henry Sunpath, Baohua Wu, Michelle Gordon, Jane Hampton, Brent Johnson, Mahomed-Yunus S. Yunus S. Moosa, Claudia Ordonez, Daniel R. Kuritzkes, and Vincent C. Marconi. High rate of K65R for antiretroviral therapy-naive patients with subtype C HIV infection failing a tenofovir-containing first-line regimen. *AIDS*, 26(13):1679–1684, aug 2012.

[169] Jessica H. Brehm, Dianna L. Koontz, Carole L. Wallis, Kathleen A. Shutt, Ian Sanne, Robin Wood, James A. McIntyre, Wendy S. Stevens, Nicolas Sluis-Cremer, and John W. Mellors. Frequent emergence of N348I in HIV-1 subtype c reverse transcriptase with failure of initial therapy reduces susceptibility to reverse-transcriptase inhibitors. *Clinical Infectious Diseases*, 55(5):737–745, 2012.

[170] Anne Derache, Carole L. Wallis, Saran Vardhanabhuti, John Bartlett, Nagalingeswaran Kumarasamy, and David Katzenstein. Phenotype, genotype, and drug resistance in Subtype C HIV-1 infection. *Journal of Infectious Diseases*, 213(2):250–256, 2016.

[171] Cissy Kityo, Jennifer Thompson, Immaculate Nankya, Anne Hoppe, Emmanuel Ndashimye, Colin Warambwa, Ivan Mambule, Joep J. Van Oosterhout, Kara Wools-Kaloustian, Silvia

Bertagnolio, Philippa J. Easterbrook, Peter Mugyenyi, A. Sarah Walker, and Nicholas I. Paton. HIV Drug Resistance Mutations in Non-B Subtypes after Prolonged Virological Failure on NNRTI-Based First-Line Regimens in Sub-Saharan Africa. *Journal of Acquired Immune Deficiency Syndromes*, 75(2):e45–e54, 2017.

[172] Samuel Genheden and Ulf Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5):449–461, may 2015.

[173] Georgios Leonis, Thomas Steinbrecher, and Manthos G. Papadopoulos. A contribution to the drug resistance mechanism of darunavir, amprenavir, indinavir, and saquinavir complexes with HIV-1 protease due to flap mutation I50V: A systematic MM-PBSA and thermodynamic integration study. *Journal of Chemical Information and Modeling*, 53(8):2141–2153, aug 2013.

[174] Hugo Gutiérrez-de Terán and Johan Åqvist. Linear Interaction Energy: Method and Applications in Drug Design. In Riccardo Baron, editor, *Computational Drug Discovery and Design*, volume 819, chapter 20, pages 305–323. Springer Science+Business Media, 2012.

[175] Christophe Chipot. Frontiers in free-energy calculations of biological systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(1):71–89, jan 2014.

[176] Zoe Cournia, Bryce Allen, and Woody Sherman. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling*, 57(12):2911–2937, 2017.

[177] Ignasi Buch, S Kashif Sadiq, and Gianni De Fabritiis. Optimized potential of mean force calculations for standard binding free energies. *Journal of Chemical Theory and Computation*, 7(6):1765–1772, 2011.

[178] Qurrat Ul Ain, Antoniya Aleksandrova, Florian D. Roessler, and Pedro J. Ballester. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 5(6):405–424, 2015.

[179] Ekachai Jenwitheesuk and Ram Samudrala. Prediction of HIV-1 protease inhibitor resistance using a protein inhibitor flexible docking approach. *Antiviral Therapy*, 10:157–166, 2005.

[180] Jaideep S. Toor, Aman Sharma, Rajender Kumar, Pawan Gupta, Prabha Garg, and Sunil K. Arora. Prediction of drug-resistance in HIV-1 subtype C based on protease sequences from ART naive and first-line treatment failures in North India using genotypic and docking analysis. *Antiviral Research*, 92(2):213–218, 2011.

[181] Zaheer Ul-Haq, Saman Usmani, Hina Shamshad, Uzma Mahmood, and Sobia Ahsan Halim. A combined 3D-QSAR and docking studies for the In-silico prediction of HIV-protease inhibitors. *Chemistry Central Journal*, 7(1):1, 2013.

[182] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. ProteinLigand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 57(4):942–957, apr 2017.

[183] Cheng Wang and Yingkai Zhang. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. *Journal of Computational Chemistry*, 38(3):169–177, jan 2017.

[184] Jooyoung Lee, Peter L Freddolino, and Yang Zhang. Ab Initio Protein Structure Prediction. In *From Protein Structure to Function with Bioinformatics*, chapter 1, pages 3–35. Springer Netherlands, Dordrecht, 2017.

[185] A Šali and T L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3):779–815, dec 1993.

[186] Andrej Šali. Modelling mutations and homologous proteins. *Current Opinion in Biotechnology*, 6(4):437–451, 1995.

[187] C M Topham, P Thomas, J P Overington, M S Johnson, F Eisenmenger, and T L Blundell. An assessment of COMPOSER: a rule-based approach to modelling protein structure. *Biochem.Soc.Symp.*, 1990.

[188] R Sánchez and a Sali. Evaluation of comparative protein structure modeling by MODELLER-3., 1997.

[189] Andrej Sali. MODELLER: A Program for Protein Structure Modeling Release 9.12, r9480, 2013.

[190] Mark J. Forster. Molecular modelling in structural biology. *Micron*, 33(4):365–384, jan 2002.

[191] Eric Feyfant, Andrej Sali, and András Fiser. Modeling mutations in protein structures. *Protein science : a publication of the Protein Society*, 16(9):2030–41, 2007.

[192] Mark Abraham, Berk Hess, David van der Spoel, and Erik Lindahl. GROMACS Reference Manual 2016, 2016.

[193] Sheng Tian, Huiyong Sun, Peichen Pan, Dan Li, Xuechu Zhen, Youyong Li, and Tingjun Hou. Assessing ensemble docking-based virtual screening strategy for kinase targets by considering protein flexibility Assessing ensemble docking-based virtual screening strategy for kinase targets by considering protein flexibility. *Journal of Chemical information and modeling*, 2014.

[194] Kanagarajan Surekha, Mutharasappan Nachiappan, Dhamodharan Prabhu, Sanjay Kumar Choubey, Jayashree Biswal, and Jeyaraman Jeyakanthan. Identification of potential inhibitors for oncogenic target of dihydroorotate dehydrogenase using in silico approaches. *Journal of Molecular Structure*, 1127:675–688, 2017.

[195] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–57, jun 2011.

[196] Douglas B. Kitchen, Hélène Decornez, John R. Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949, 2004.

[197] Ruth Huey, Garrett M. Morris, Arthur J. Olson, and David S. Goodsell. A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry*, 28(6):1145–1152, apr 2007.

[198] G Morris, D Goodsell, M Pique, W Lindstrom, R Huey, S Forli, W E Hart, S Halliday, R Belew, and A J Olson. User Guide AutoDock Version 4.2, 2012.

[199] L Wesson, D Eisenberg, Laura Wesson, and David Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science*, pages 227–235, 1992.

[200] Oleg Trott and Arthur J Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2009.

[201] Gabriela Bitencourt-Ferreira and Walter Filgueira de Azevedo. Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes. *Biophysical Chemistry*, 240(June):63–69, 2018.

[202] Renxiao Wang, Luhua Lai, and Shaomeng Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design*, 16(1):11–26, 2002.

[203] S. Bowerman and J. Wereszczynski. Detecting Allosteric Networks Using Molecular Dynamics Simulation, jan 2016.

[204] David K. Brown, Olivier Sheik Amamuddy, and Özlem Tastan Bishop. Structure-Based Analysis of Single Nucleotide Variants in the Renin-Angiotensinogen Complex. *Global Heart*, 2017.

[205] Yong Duan, Chun Wu, Shibasish Shibashish Chowdhury, Mathew C. Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, James Caldwell, Junmei Wang, and Peter Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24(16):1999–2012, 2003.

[206] Jay W. Ponder and David A. Case. Force Fields for Protein Simulations. *Advances in Protein Chemistry*, 66:27–85, 2003.

[207] H J C Berendsen, J P M Postma, W F van Gunsteren, a DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.

[208] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[209] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[210] Jean Paul Ryckaert, Giovanni Ciccotti, and Herman J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 1977.

[211] Patrick Guillaume. Modal Analysis. Technical report, Department of Mechanical Engineering,Vrije Universiteit Brussel, Brussel, Belgium, 2000.

[212] Eric C. Dykeman and Otto F. Sankey. Normal mode analysis and applications in biological physics. *Journal of Physics Condensed Matter*, 22(42), 2010.

[213] Guang Hu, Luisa Di Paola, Zhongjie Liang, and Alessandro Giuliani. Comparative Study of Elastic Network Model and Protein Contact Network for Protein Complexes: The Hemoglobin Case. *BioMed Research International*, 2017:1–15, jan 2017.

[214] Shin Nakamura and Takumi Noguchi. Quantum mechanics/molecular mechanics simulation of the ligand vibrations of the water-oxidizing Mn <sub>4</sub> CaO <sub>5</sub> cluster in photosystem II. *Proceedings of the National Academy of Sciences*, 113(45):12727–12732, 2016.

[215] Michael Levitt, Christian Sander, and Peter S. Stern. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *International Journal of Quantum Chemistry*, 24(S10):181–199, 2009.

[216] Monique M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters*, 77(9):1905–1908, 1996.

[217] Charles C David and Donald J Jacobs. Principal component analysis: a method for determining the essential dynamics of proteins. *Methods in molecular biology*, 1084:193–226, 2014.

[218] Ataur R Katebi, Kannan Sankar, Kejue Jia, and Robert L Jernigan. The Use of Experimental Structures to Model Protein Dynamics. In *Methods in molecular biology (Clifton, N.J.)*, volume 1215, pages 213–236. NIH Public Access, 2015.

[219] Ivet Bahar, Timothy R Lezon, Ahmet Bakan, and Indira H Shrivastava. Normal Mode Analysis of Biomolecular Structures: Functional Mech Membrane Proteins. *Chemical reviews*, 110(3):1463–1497, 2010.

[220] Konrad Hinsen. Normal Mode Theory and Harmonic Potential Approximations. In *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, pages 1–16. Boca Raton, FL: Chapman & Hall/CRC, dec 2005.

[221] Jianpeng Ma. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13(3):373–380, 2005.

[222] B. E. Eichinger. Elasticity Theory. I. Distribution Functions for Perfect Phantom Networks. *Macromolecules*, 5(4):496–505, 1972.

[223] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[224] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, 80(1):505–515, 2001.

[225] J A McCammon. Protein dynamics. *Reports on Progress in Physics*, 47(1):1–46, jan 1984.

[226] Kota Kasahara, Ikuo Fukuda, and Haruki Nakamura. A novel approach of dynamic cross correlation analysis on molecular dynamics simulations and its application to Ets1 dimer-DNA complex. *PLoS ONE*, 9(11), 2014.

[227] S Swaminathan. Investigation of domain structure in proteins via molecular dynamics simulation: application to HIV-1 protease dimer. *Journal of the American ...*, 11351(36):2717–2721, 1991.

[228] Akio Kitao and Nobuhiro Go. Investigating protein dynamics in collective coordinate space. *Current Opinion in Structural Biology*, 9(2):164–169, 1999.

[229] P.H. Hünenberger, A.E. Mark, and W.F. van Gunsteren. Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *Journal of Molecular Biology*, 252(4):492–503, 1995.

[230] Maria Batool, Masaud Shah, Mahesh Chandra Patra, Dhanusha Yesudhas, and Sangdun Choi. Structural insights into the Middle East respiratory syndrome coronavirus 4a protein and its dsRNA binding mechanism. *Scientific Reports*, 7(1), 2017.

[231] Duan Ni, Kun Song, Jian Zhang, and Shaoyong Lu. Molecular dynamics simulations and dynamic network analysis reveal the allosteric unbinding of monobody to H-Ras triggered by R135K mutation. *International Journal of Molecular Sciences*, 18(11), 2017.

[232] Huimin Zhang, Tianqing Song, Yizhao Yang, Chenggong Fu, and Jiazhong Li. Exploring the Interaction Mechanism Between Cyclopeptide DC3 and Androgen Receptor Using Molecular Dynamics Simulations and Free Energy Calculations. *Frontiers in Chemistry*, 6(April):1–15, apr 2018.

[233] Canan Atilgan and Ali Rana Atilgan. Perturbation-Response Scanning Reveals Ligand Entry-Exit Mechanisms of Ferric Binding Protein. *PLoS computational biology*, 5(10), 2009.

[234] C. Atilgan, Z. N. Gerek, S. B. Ozkan, and A. R. Atilgan. Manipulation of conformational change in proteins by single-residue perturbations. *Biophysical Journal*, 99(3):933–943, 2010.

[235] Z. Nevin Gerek and S. Banu Ozkan. Change in allosteric network affects binding affinities of PDZ domains: Analysis through perturbation response scanning. *PLoS Computational Biology*, 7(10):18–25, 2011.

[236] Haleh Abdizadeh, Gokce Guven, Ali Rana Atilgan, and Canan Atilgan. Perturbation response scanning specifies key regions in subtilisin serine protease for both function and stability. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 30(6):867–873, 2015.

[237] David K. Brown, David L. Penkler, Olivier Sheik Amamuddy, Caroline Ross, Ali Rana Atilgan, Canan Atilgan, and Özlem Tastan Bishop. MD-TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics*, 33(May):2768–2771, 2017.

[238] David Penkler, Özge Sensoy, Canan Atilgan, and Özlem Tastan Bishop. Perturbation-Response Scanning Reveals Key Residues for Allosteric Control in Hsp70. *Journal of Chemical Information and Modeling*, 57(6):1359–1374, 2017.

[239] Luca Ponzoni and Ivet Bahar. Structural dynamics is a determinant of the functional significance of missense variants. *Proceedings of the National Academy of Sciences*, 115(16):4164–4169, 2018.

[240] M. Salter-Townshend, A. White, I. Gollini, and T. B. Murphy. Review of statistical network analysis: Models, algorithms, and software. *Statistical Analysis and Data Mining*, 5(4):243–264, aug 2012.

[241] Steve Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology.* Springer, 2011.

[242] Albert-László Barabási and Eric Bonabeau. Scale-Free Networks. *Scientific American*, 288(5):60–69, may 2003.

[243] Giorgio Ausiello, Donatella Firmani, and Luigi Laura. The (betweenness) centrality of critical nodes and network cores. In *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 90–95. IEEE, jul 2013.

[244] Linton C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35, mar 1977.

[245] Pew Thian Yap, Yong Fan, Yasheng Chen, John H. Gilmore, Weili Lin, and Dinggang Shen. Development trends of white matter connectivity in the first years of life. *PLoS ONE*, 6(9), 2011.

[246] T. Maekawa. Computation of Shortest Paths on Free-Form Parametric Surfaces. *Journal of Mechanical Design*, 118(4):499, 1996.

[247] Rajdeep K. Grewal Roy and Soumen. Modeling proteins as residue interaction networks. *Protein & Peptide Letters*, 22:923–933, 2015.

[248] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1959.

[249] Yefim Dinitz and Rotem Itzhak. Hybrid BellmanFordDijkstra algorithm. *Journal of Discrete Algorithms*, 42:35–44, 2017.

[250] W. Zeng and R. L. Church. Finding shortest paths on real road networks: the case for A*. *International Journal of Geographical Information Science*, 23(4):531–543, apr 2009.

[251] Donald B. Johnson. Efficient Algorithms for Shortest Paths in Sparse Networks. *Journal of the ACM*, 1977.

[252] Yijie Han and Tadao Takaoka. An O(n3loglogn/log2n) time algorithm for all pairs shortest paths. *Journal of Discrete Algorithms*, 38-41:9–19, 2016.

[253] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10: proteinprotein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, jan 2015.

[254] Khalid Zuberi, Max Franz, Harold Rodriguez, Jason Montojo, Christian Tannus Lopes, Gary D. Bader, and Quaid Morris. GeneMANIA prediction server 2013 update. *Nucleic acids research*, 2013.

[255] Ignacio Riquelme Medina and Zelmina Lubovac-Pilav. Gene Co-Expression Network Analysis for Identifying Modules and Functionally Enriched Pathways in Type 1 Diabetes. *PLOS ONE*, 11(6), jun 2016.

[256] Julia C. Fitzgerald and Helene Plun-Favreau. Emerging pathways in genetic Parkinson's disease: Autosomal-recessive genes in Parkinson's disease - A common pathway?, 2008.

[257] Paulami Chatterjee, Debjani Roy, Malay Bhattacharyya, and Sanghamitra Bandyopadhyay. Biological networks in Parkinson's disease: An insight into the epigenetic mechanisms associated with this disease. *BMC Genomics*, 18(1):721, sep 2017.

[258] Alex Fornito, Andrew Zalesky, and Michael Breakspear. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172, 2015.

[259] Guang Hu, Wenying Yan, Jianhong Zhou, and Bairong Shen. Residue interaction network analysis of Dronpa and a DNA clamp. *Journal of Theoretical Biology*, 348:55–64, may 2014.

[260] Damiano Piovesan, Giovanni Minervini, and SilvioC.E. Tosatto. The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Research*, page gkw315, 2016.

[261] Gizem Ozbaykal, Ali Rana Atilgan, and Canan Atilgan. In silico mutational studies of Hsp70 disclose sites with distinct functional attributes. *Proteins: Structure, Function and Bioinformatics*, 83(11):2077–2090, 2015.

[262] Urmi Doshi, Michael J Holliday, Elan Z Eisenmesser, and Donald Hamelberg. Dynamical network of residue-residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 113(17):4735–4740, 2016.

[263] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindah. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[264] A W Sousa da Silva and W F Vranken. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC research notes*, 5(1):367, 2012.

[265] Nancy M. King, Moses Prabu-Jeyabalan, Rajintha M. Bandaranayake, Madhavi N. L. Nalam, Ellen A Nalivaika, Ayegül Özen, Türkan Halilolu, Nee Kurt Ylmaz, and Celia A. Schiffer. Extreme EntropyEnthalpy Compensation in a Drug-Resistant Variant of HIV-1 Protease. *ACS Chemical Biology*, 7(9):1536–1546, sep 2012.

[266] Andrey Y Kovalevsky, Fengling Liu, Sofiya Leshchenko, Arun K Ghosh, John M Louis, Robert W Harrison, and Irene T Weber. Ultra-high resolution crystal structure of HIV-1 protease mutant reveals two binding sites for clinical inhibitor TMC114. *Journal of molecular biology*, 363(1):161–73, oct 2006.

[267] Chen-Hsiang Shen, Yuan-Fang Wang, Andrey Y Kovalevsky, Robert W Harrison, and Irene T Weber. Amprenavir complexes with HIV-1 protease and its drug-resistant mutants altering hydrophobic clusters. *The FEBS journal*, 277(18):3699–714, sep 2010.

[268] Fengling Liu, Peter I Boross, Yuan-Fang Wang, Jozsef Tozser, John M Louis, Robert W Harrison, and Irene T Weber. Kinetic, stability, and structural changes in high-resolution crystal structures of HIV-1 protease with drug-resistant mutations L24I, I50V, and G73S. *Journal of molecular biology*, 354(4):789–800, dec 2005.

[269] S. Muzammil, A. A. Armstrong, L. W. Kang, A. Jakalian, P. R. Bonneau, V. Schmelmer, L. M. Amzel, and E. Freire. Unique Thermodynamic Response of Tipranavir to Human Immunodeficiency Virus Type 1 Protease Drug Resistance Mutations. *Journal of Virology*, 81(10):5144–5154, 2007.

[270] Yunfeng Tie, Andrey Y. Kovalevsky, Peter Boross, Yuan-Fang Wang, Arun K. Ghosh, Jozsef Tozser, Robert W. Harrison, and Irene T. Weber. Atomic resolution crystal structures of HIV-1 protease and mutants V82A and I84V with saquinavir. *Proteins: Structure, Function, and Bioinformatics*, 67(1):232–242, jan 2007.

[271] Ahmet Bakan, Lidio M. Meireles, and Ivet Bahar. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577, 2011.

[272] Todd J. Dolinsky, Jens E. Nielsen, J. Andrew McCammon, and Nathan A. Baker. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Research*, 32(WEB SERVER ISS.), 2004.

[273] Garrett M. Morris, Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, 2009.

[274] Ole Tange. GNU Parallel: the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47, 2011.

[275] Ekachai Jenwitheesuk and Ram Samudrala. Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations. *BMC Structural Biology*, 3(1):2, 2003.

[276] A. Pedretti, L. Villa, and Giulio Vistoli. Atom-type description language: A universal language to recognize atom types implemented in the VEGA program. *Theoretical Chemistry Accounts*, 109(4):229–232, 2003.

[277] Bradley C Logsdon, John F Vickrey, Philip Martin, Gheorghe Proteasa, Jay I Koepke, Stanley R Terlecky, Zdzislaw Wawrzak, Mark A Winters, Thomas C Merigan, and Ladislau C Kovari. Crystal Structures of a Multidrug-Resistant Human Immunodeficiency Virus Type 1 Protease Reveal an Expanded Active-Site Cavity. *Journal of Virology*, 78(6):3123–3132, mar 2004.

[278] Andrej Sali. MODELLER A Program for Protein Structure Modeling Release 9v4, r6262, 2008.

[279] Hongjian Li, Kwong Sak Leung, and Man Hon Wong. Idock: A multithreaded virtual screening tool for flexible ligand docking. *2012 IEEE Symposium on Computational Intelligence and Computational Biology, CIBCB 2012*, pages 77–84, 2012.

[280] David A. Pearlman. Evaluating the molecular mechanics Poisson-Boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP Kinase. *Journal of Medicinal Chemistry*, 48(24):7796–7807, 2005.

[281] Nese Kurt, Walter R.P. Scott, Celia A. Schiffer, and Turkan Haliloglu. Cooperative fluctuations of unliganded and substrate-bound HIV-1 protease: A structure-based analysis on a variety of conformations from crystallography and molecular dynamics simulations. *Proteins: Structure, Function and Genetics*, 2003.

[282] Shu-Wen W Chen and Jean-Luc Pellequer. Identification of functionally important residues in proteins using comparative models. *Current medicinal chemistry*, 11(5):595–605, 2004.

[283] Jenny Gu and Philip E Bourne. *Structural bioinformatics.* John Wiley & Sons, 2009.

[284] J. Rydzewski, R. Jakubowski, and W. Nowak. Communication: Entropic measure to prevent energy over-minimization in molecular dynamics simulations. *Journal of Chemical Physics*, 143(17), 2015.

[285] Jiapu Zhang. The Hybrid Idea of (Energy Minimization) Optimization Methods Applied to Study PrionProtein Structures Focusing on the beta2-alpha2 Loop. *Biochemistry & Pharmacology: Open Access*, 04(04):1–23, 2015.

[286] Jiapu Zhang. *Molecular Dynamics Analyses of Prion Protein Structures: The Resistance to Prion Diseases Down Under.* Springer, 2018.

[287] Jennifer C Brookes. Quantum effects in biology: golden rule in enzymes, olfaction, photosynthesis and magnetodetection. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 473(2201):20160822, may 2017.

[288] Jennifer E Foulkes-Murzycki, Walter Scott Peter Robert, and Celia A Schiffer. Hydrophobic Sliding : A Possible Mechanism for Drug Resistance in Human Immunodeficiency Virus Type 1 Protease. *Structure*, 15(February):225–233, feb 2007.

[289] Rieko Ishima, Qingguo Gong, Yunfeng Tie, Irene T. Weber, and John M. Louis. Highly conserved glycine 86 and arginine 87 residues contribute differently to the structure and activity of the mature HIV-1 protease. *Proteins: Structure, Function, and Bioinformatics*, 78(4):1015–1025, mar 2010.

[290] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In Gäel Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference (SciPy2008)*, volume 836, pages 11–15, Pasadena, CA USA, 2008.

[291] Kazuya Okamoto, Wei Chen, and Xiang-yang Li. Ranking of Closeness Centrality for Large-Scale Social Networks. In Franco P Preparata, Xiaodong Wu, and Jianping Yin, editors, *Frontiers in Algorithmics*, pages 186–195. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[292] Debra a Ragland, Ellen a Nalivaika, Madhavi N L Nalam, Kristina Parachanronarong, Hong Cao, Rajintha M Bandaranayake, Yufeng Cai, Nese Kurt-Yilmaz, and Celia a Schiffer. Drug Resistance Conferred by Mutations Outside the Active Site through Alterations in the Dynamic and Structural Ensemble of HIV-1 Protease. *Journal of the American Chemical Society*, 2014.

[293] Hiroyasu Ohtaka, Arne Schön, and Ernesto Freire. Multidrug Resistance to HIV-1 Protease Inhibition Requires Cooperative Coupling between Distal Mutations. *Biochemistry*, 42(46):13659–13666, 2003.

[294] Andrew H Kaplan, Scott F Michael, Robert S Wehbie, Mark F Knigge, Deborah A Paul, Lorraine Everitt, Dale J Kempfii, Daniel W Norbeckii, John W Erickson, and Ronald Swanstrom. Selection of multiple human immunodeficiency virus type 1 variants that encode viral proteases with decreased sensitivity to an inhibitor of the viral protease. *Proceedings of the National Academy of Sciences*, 91(June):5597–5601, 1994.

[295] Monique Nijhuis, Rob Schuurman, Dorien De Jong, John Erickson, Elena Gustchina, Jan Albert, Pauline Schipper, Sergei Gulnik, and Charles A B Boucher. Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. *AIDS*, 13:2349–2359 Keywords:, 1999.

[296] Hilary B Schock, Victor M Garsky, and Lawrence C Kuo. Mutational Anatomy of an HIV-1 Protease Variant Conferring Cross-resistance to Protease Inhibitors in Clinical Trials. *The Journal of biological chemistry*, 271(50):31957–31963, 1996.

[297] John M Louis, Ying Zhang, Jane M Sayer, Yuan Fang Wang, Robert W Harrison, and Irene T Weber. The L76V drug resistance mutation decreases the dimer stability and rate of autoprocessing of HIV-1 protease by reducing internal hydrophobic contacts. *Biochemistry*, 50(21):4786–4795, may 2011.

[298] Nathan E. Goldfarb, Meray Ohanessian, Shyamasri Biswas, T. Dwight McGee, Brian P. Mahon, David A. Ostrov, Jose Garcia, Yan Tang, Robert McKenna, Adrian Roitberg, and Ben M. Dunn. Defective Hydrophobic Sliding Mechanism and Active Site Expansion in HIV-1 Protease Drug Resistant Variant Gly48Thr/Leu89Met: Mechanisms for the Loss of Saquinavir Binding Potency. *Biochemistry*, 54(2):422–433, jan 2015.

[299] Viktor Hornak, Asim Okur, Robert C Rizzo, and Carlos Simmerling. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 103(4):915–920, jan 2006.

[300] Hai Nguyen, David A Case, and Alexander S Rose. NGLviewinteractive molecular graphics for Jupyter notebooks. *Bioinformatics*, dec 2017.

[301] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee Ping Wang, Thomas J. Lane, and Vijay S. Pande. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*, 109(8):1528–1532, 2015.

[302] Stéfan Van Der Walt, S. Chris Colbert, and Gaël Varoquaux. The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30, 2011.

[303] Olivier Sheik Amamuddy, Nigel T Bishop, and Tastan Bishop. Characterizing drug resistance using geometric ensembles from HIV protease dynamics, 2018.

[304] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286, oct 1999.

[305] UNAIDS. Fact sheet - Latest global and regional statistics on the status of the AIDS epidemic. Technical Report June, UNAIDS, 2017.

[306] European AIDS Clinical Society. European AIDS Clinical Society (EACS) Guidelines, 2017.

[307] Jeanette M. Wood, Jürgen Maibaum, Joseph Rahuel, Markus G. Grütter, Nissim Claude Cohen, Vittorio Rasetti, Heinrich Rüger, Richard Göschke, Stefan Stutz, Walter Fuhrer, Walter Schilling, Pascal Rigollier, Yasuchika Yamaguchi, Frederic Cumin, Hans Peter Baum, Christian R. Schnell, Peter Herold, Robert Mah, Chris Jensen, Eoin O'Brien, Alice Stanton, and Martin P. Bedigian. Structure-based design of aliskiren, a novel orally effective renin inhibitor. *Biochemical and Biophysical Research Communications*, 308(4):698–705, 2003.

[308] Mariela Bollini, Robert A. Domaoal, Vinay V. Thakur, Ricardo Gallardo-Macias, Krasimir A. Spasov, Karen S. Anderson, and William L. Jorgensen. Computationally-Guided Optimization of a Docking Hit to Yield Catechol Diethers as Potent Anti-HIV Agents. *Journal of Medicinal Chemistry*, 54(24):8582–8591, dec 2011.

[309] William T. Gray, Kathleen M. Frey, Sarah B. Laskey, Andrea C. Mislak, Krasimir A. Spasov, Won-Gil Lee, Mariela Bollini, Robert F. Siliciano, William L. Jorgensen, and Karen S. Anderson. Potent Inhibitors Active against HIV Reverse Transcriptase with K101P, a Mutation Conferring Rilpivirine Resistance. *ACS Medicinal Chemistry Letters*, 6(10):1075–1079, oct 2015.

[310] Mingkun Jiao, Gang Liu, Yu Xue, and Chunyong Ding. Computational Drug Repositioning for Cancer Therapeutics. *Current Topics in Medicinal Chemistry*, 15(8):767–775, 2015.

[311] L. Patrick Graham. *Introduction to medicinal chemistry.* Oxford University Press, United Kingdom, Oxford, 5th edition, 2013.

[312] Zhengtong Lv, Yuan Chu, and Yong Wang. HIV protease inhibitors: a review of molecular selectivity and toxicity. *HIV/AIDS - Research and Palliative Care*, 7:95–104, 2015.

[313] Ruy M Ribeiro, Li Qin, Leslie L Chavez, Dongfeng Li, Steven G Self, and Alan S Perelson. Estimation of the Initial Viral Growth Rate and Basic Reproductive Number during Acute HIV-1 Infection. *Journal of Virology*, 84(12):6096–6102, jun 2010.

[314] Raphael W. Lihana, Deogratius Ssemwanga, Alash'le Abimiku, and Nicaise Ndembi. Update on HIV-1 diversity in Africa: A decade in review. *AIDS Reviews*, 14:83–100, 2012.

[315] Max W. Chang and Bruce E. Torbett. Accessory mutations maintain stability in drug-resistant HIV-1 protease. *Journal of Molecular Biology*, 410(4):756–760, 2011.

[316] Adrian Curran and Esteve Ribera Pascuet. [Darunavir as first-line therapy. The TITAN study]. *Enfermedades infecciosas y microbiologia clinica*, 26 Suppl 1:14–22, 2008.

[317] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–7, 2004.

[318] Marcus D. Hanwell, Donald E. Curtis, David C. Lonie, Tim Vandermeerschd, Eva Zurek, and Geoffrey R. Hutchison. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 2012.

[319] Seema Patel, Abdur Rauf, Haroon Khan, and Tareq Abu-Izneid. Renin-angiotensin-aldosterone (RAAS): The ubiquitous system for homeostasis and pathologies. *Biomedicine & Pharmacotherapy*, 94:317–325, oct 2017.

[320] Shahriar Iravanian and Samuel C. Dudley. The renin-angiotensin-aldosterone system (RAAS) and cardiac arrhythmias. *Heart Rhythm*, 5(6):S12–S17, jun 2008.

[321] Alex Bateman, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, Mark Bingley, Carlos Bonilla, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Andrew Cowley, Alan Da Silva, Maurizio De Giorgi, Tunca Dogan, Francesco Fazzini, Leyla Garcia Castro, Luis Figueira, Penelope Garmiri, George Georghiou, Daniel Gonzalez, Emma Hatton-Ellis, Weizhong Li, Wudong Liu, Rodrigo Lopez, Jie Luo, Yvonne Lussi, Alistair MacDougall, Andrew Nightingale, Barbara Palka, Klemens Pichler, Diego Poggioli, Sangya Pundir, Luis Pureza, Guoying Qi, Alexandre Renaux, Steven Rosanoff, Rabie Saidi, Tony Sawford, Aleksandra Shypitsyna, Elena Speretta, Edward Turner, Nidhi Tyagi, Vladimir Volynkin, Tony Wardell, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casal-Casas, Edouard de Castro, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Florence Jungo, Guillaume Keller, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Neto, Nevila Nouspikel, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey, Cathy H Wu, Cecilia N Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Qinghua Wang, Yuqi Wang, Lai-Su Yeh, and Jian Zhang. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, jan 2017.

[322] J. Bouhnik, F.-X. Galen, J. Menard, P. Corvol, R. Seyer, J.A. Fehrentz, D.L. Nguyen, P. Fulcrand, and B. Castro. Production and characterization of human renin antibodies with region-oriented synthetic peptides. *Journal of Biological Chemistry*, 1987.

[323] Gerhard (Editor) Klebe. *Drug Design: Methodology, Concepts and Mode-of-Action.* Springer Berlin Heidelberg, 2013.

[324] John M. Eisenberg Center for Clinical Decisions John M. Eisenberg Center for Clinical Decisions and Communications Science and Communications. *ACEIs, ARBs, or DRI for Adults With Hypertension.* Agency for Healthcare Research and Quality (US), oct 2007.

[325] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, apr 2010.

[326] Yongwook Choi, Gregory E. Sims, Sean Murphy, Jason R. Miller, and Agnes P. Chan. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, 7(10):e46688, oct 2012.

[327] E. Capriotti, R. Calabrese, and R. Casadio. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22):2729–2734, nov 2006.

[328] Huaiyu Mi, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D. Thomas. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189, jan 2017.

[329] Hashem A. Shihab, Julian Gough, David N. Cooper, Peter D. Stenson, Gary L A Barker, Keith J. Edwards, Ian N M Day, and Tom R. Gaunt. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, 34(1):57–65, 2013.

[330] G Csárdi and T Nepusz. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695:1695, 2006.

[331] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[332] Arun K. Ghosh, Jun Takayama, Luke A. Kassekert, Jean-Rene Ella-Menye, Sofiya Yashchuk, Johnson Agniswamy, Yuan-Fang Wang, Manabu Aoki, Masayuki Amano, Irene T. Weber, and Hiroaki Mitsuya. Structure-based design, synthesis, X-ray studies, and biological evaluation of novel HIV-1 protease inhibitors containing isophthalamide-derived P2-ligands. *Bioorganic & Medicinal Chemistry Letters*, 25(21):4903–4909, nov 2015.

[333] Yunfeng Tie, Yuan-Fang Wang, Peter I. Boross, Ting-Yi Chiu, Arun K. Ghosh, Jozsef Tozser, John M. Louis, Robert W. Harrison, and Irene T. Weber. Critical differences in HIV-1 and HIV-2 protease specificity for clinical inhibitors. *Protein Science*, 21(3):339–350, mar 2012.

[334] Caroline Ross, Bilal Nizami, Michael Glenister, Olivier Sheik Amamuddy, Ali Rana Atilgan, Canan Atilgan, and Özlem Tastan Bishop. MODE-TASK: large-scale protein motion tools. *Bioinformatics*, may 2018.

[335] Jingshan Ren, Xiangxi Wang, Ling Zhu, Zhongyu Hu, Qiang Gao, Pan Yang, Xuemei Li, Junzhi Wang, Xinliang Shen, Elizabeth E. Fry, Zihe Rao, and David I. Stuart. Structures of Coxsackievirus A16 Capsids with Native Antigenicity: Implications for Particle Expansion, Receptor Binding, and Immunogenicity. *Journal of Virology*, 89(20):10500–10511, 2015.