Analysis of the Protein–Ligand and Protein–Peptide Interactions Using a Combined Sequence- and Structure-Based Approach

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

von
Tomas Bastys

Saarbrücken, 2018

Tag des Kolloquiums:	17. April 2019
Dekan:	Prof. Dr. Sebastian Hack
Prüfungsausschuss:	
Vorsitzender:	Prof. Dr. Dr. h.c. mult. Kurt Mehlhorn
Berichterstatter:	Prof. Dr. Olga V. Kalinina Prof. Dr. Volkhard Helms
Akademischer Mitarbeiter:	Dr. Peter Ebert

Mano tėvui

Abstract

Proteins participate in most of the important processes in cells, and their ability to perform their function ultimately depends on their three-dimensional structure. They usually act in these processes through interactions with other molecules. Because of the importance of their role, proteins are also the common target for small molecule drugs that inhibit their activity, which may include targeting protein interactions. Understanding protein interactions and how they are affected by mutations is thus crucial for combating drug resistance and aiding drug design.

This dissertation combines bioinformatics studies of protein interactions at both primary sequence and structural level. We analyse protein–protein interactions through linear motifs, as well as protein–small molecule interactions, and study how mutations affect them. This is done in the context of two systems. In the first study of drug resistance mutations in the protease of the human immunodeficiency virus type 1, we successfully apply molecular dynamics simulations to estimate the effects of known resistance-associated mutations on the free binding energy, also revealing molecular mechanisms of resistance. In the second study, we analyse consensus profiles of linear motifs that mediate the recognition by the mitogen-activated protein kinases of their target proteins. We thus gain insights into the cellular processes these proteins are involved in.

Kurzfassung

Proteine sind an den meisten wichtigen Prozessen in Zellen beteiligt, und ihre Fähigkeit, ihre Funktion zu erfüllen, hängt letztlich von ihrer dreidimensionalen Struktur ab. In diesen Prozessen wirken sie normalerweise durch Wechselwirkungen mit anderen Molekülen. Aufgrund der Bedeutung ihrer Rolle sind Proteine auch die häufigsten Angriffspunkte für niedermolekulare Wirkstoffe, die ihre Aktivität hemmen. Dies kann das Targeting von Proteinwechselwirkungen umfassen. Um Wechselwirkungen mit Medikamenten zu bekämpfen und das Wirkstoffdesign zu unterstützen, ist es wichtig, die Wechselwirkungen zwischen Proteinen und deren Einfluss auf Mutationen zu verstehen.

Diese Dissertation kombiniert bioinformatische Studien zu Proteinwechselwirkungen sowohl auf primärer als auch auf struktureller Ebene. Wir analysieren Protein-Protein-Wechselwirkungen anhand linearer Motive sowie Protein-Kleinmolekül-Wechselwirkungen und untersuchen, wie sich Mutationen auf sie auswirken. Dies wird untersucht im Kontext von zwei Systemen. In der ersten Studie zu Resistenzmutationen in der Protease des humanen Immundefizienzvirus Typ 1 haben wir molekulardynamische Simulationen erfolgreich eingesetzt, um die Auswirkungen bekannter Resistenz-assoziierter Mutationen auf die freie Bindungsenergie abzuschätzen und molekulare Resistenzmechanismen aufzuzeigen. In der zweiten Studie analysieren wir Konsensusprofile von linearen Motiven, die die Erkennung der Zielproteine durch die Mitogen-aktivierten Proteinkinasen vermitteln. So gewinnen wir Einblick in die zellulären Prozesse, an denen diese Proteine beteiligt sind.

Acknowledgements

There are many people from whose support this dissertation has benefited greatly and who deserve special acknowledgement here.

First and foremost, I would like to express my utmost gratitude to my supervisor Olga Kalinina for the guidance, support, trust, and the liberty in the means of pursuing research projects. Doctorate is a long road that can end in many places and I am happy where I've arrived thanks to you.

My special thanks go to Volkhard Helms for *yet again* agreeing to review a thesis of mine.

I am very thankful to my collaborators András, Attila, Bert, Hauke, Rolf, and Vytautas - the discussions we've had have been very stimulating, they've helped me advance my knowledge, as well as my curiosity. Very special thanks in this category go to András, who proofread parts of this dissertation, and to Vytautas, thanks to whom I've learned immensely on the subject, as well became a better scientist, and also for proofreading parts of this dissertation.

I am grateful to my colleagues at Max Planck Institute for Informatics and Saarland University for making my journey a pleasant one. First and foremost thank you Mazen, who originally introduced me to molecular dynamics, for the scientific advice, for the countless interesting chats we had over cups of tea, and for all the various help you've provided. Secondly, Peter, whom I rightfully refer to as my "lawyer" - I put my full trust in you on advice in various scientific as well as all sorts of "what do I do now?" matters, and was never disappointed. For long scientific discussions to my long-time office mate Olga, as well as Prabhav and Sarvesh, to both of whom I am also indebted for proofreading parts of this dissertation. My further thanks for advice extend to Nadezhda, Nico, Adrin, Alexander, Sebastian, Alejandro, Anna, Matthias, and Michael. Finally I thank for the various interesting discussions Dilip, Florian, Glenn, Lara, Lisa, Markus, Sivarajan, and Fabian, who also cordially provided me with this dissertation's template.

I am thankful for the opportunity to work in Thomas Lengauer's department, who has been a role model for me with regards to his academic approach and integrity.

My appreciation for all the technical support goes to Achim, Georg, and also Andreas, Maik, Wolfram, and the rest of MPII-IST.

I had the fortune to advise a talented master student Sanjay on his thesis, whom I must thank for teaching me about myself.

I owe my gratitude to my maths teachers from Lithuania Jūratė Bakasėnaitė for encouragement, and the extraordinary Antanas Skūpas for pushing his pupils to gain extra knowledge.

I thank my friends in Saarbrücken Adam, Ashkan, Barbara, Freddie, Markus, Saskia, and Stefan for the support and great times we have had together.

My deep gratitude goes to my family, for all of the unquestionable support over the years. *Babul, ja uže ne student.*

My final thanks goes to the greatest treasure and support gained during my studies in Saarbrücken, the queen of LaTeX and my queen too, Iulia.

Contents

List of Figures

List of Tables

1 Introduction

After observing the consistency of the amount of nuclear DNA in all the cells in all the individuals for a given animal species, it was suggested that DNA and not the protein is the underlying substance of genes (R. Vendrely and C. Vendrely 1948). Lack of correlation between the amount of this DNA (C-value) and the presumed organism complexity was later referred to as C-value paradox (Thomas Jr 1971), which was then resolved by the discovery of the non-coding DNA. Sequencing of human genome, however, indicated an unexpectedly small number of protein-coding genes and domains (International Human Genome Sequencing Consortium 2001; Rubin 2001). This apparent lack of relationship between the genes and organism complexity was then termed the G-value paradox (M. W. Hahn and Wray 2002). Proteins, however, rarely act alone and usually play a role in protein–protein interaction networks or complexes (Cusick *et al.* 2005). Thinking in terms of electric circuits, it is possible to build a large variety of different electric circuits using only a limited number of electronic components; it is in the same way by combining different proteins and/or protein domains that a phenotypic complexity and diversity can arise in biology (Bhattacharyya *et al.* 2006). Such modular approach is particularly suited for protein signalling networks (Cusick *et al.* 2005). Some of the underlying biological techniques for modularity in such networks include docking[1], modular interaction domains, and scaffold or adapter proteins, as well as their combinations (Figure 1.1). All of these protein–protein interactions refer to physical encounters between the underlying three-dimensional structures of these molecules. There is however a wide spectrum of variety of the interfaces facilitating these interactions, ranging from coming together of rigid domains, unstructured linear motifs that may acquire secondary conformation only upon binding to its structured partner, or even cases where both interfaces lack rigid structure (Davey *et al.* 2012). Mitogen-activated protein kinase (MAPK), a broadly used enzyme for signalling in eukaryotes, is an example of a protein that uses docking with an unstructured linear motif to facilitate its interactions.

Due to the central role of proteins in the cellular processes, in the vast majority of the cases they are the target of pharmaceutical drugs (R. Santos *et al.* 2016). Kinases is one class of proteins which are targeted by a rapidly growing number of inhibitors. One of the most widely pharmaceutically targeted pathogens, Human Immunodeficiency Virus type 1 (HIV-1), can be considered both one of the success stories in medical treatment in terms of combating its pathogenicity (Antiretroviral Therapy Cohort Collaboration 2017), and arguably one of the first widely used applications of personalised medicine (Lengauer *et al.* 2014). Both of these factors are the result of HIV-1 being the first viral disease with a large number of different therapy options available (Palella *et al.* 1998). Indeed, multiple classes of drugs against HIV-1 include those targeting different HIV-1 proteins and their different interaction mechanisms with other molecules: binding to the active site, to the allosteric site restricting their conformational flexibility, or even using

[1] As per (Bhattacharyya *et al.* 2006), in the context of this dissertation, docking refers to interaction of a catalytic domain and a partner protein that does not involve the active site.

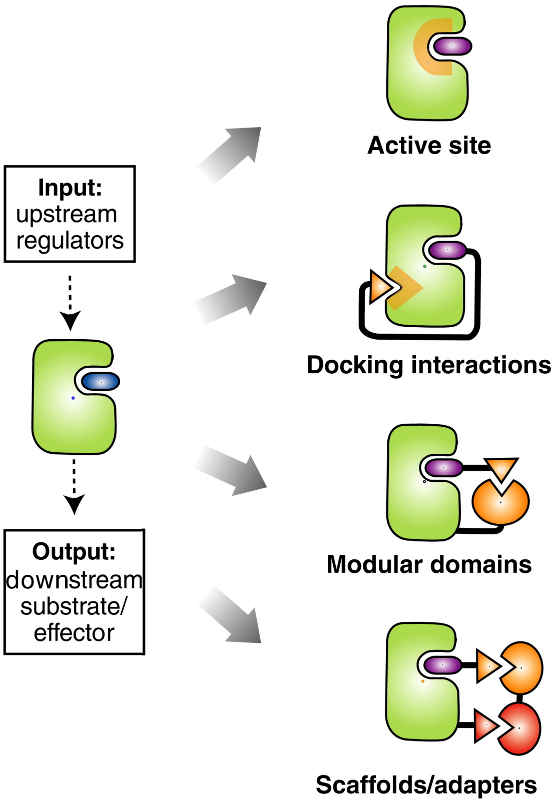


Figure 1.1: Different enzyme-protein interaction modes. Adapted from (Bhattacharyya *et al.* 2006) and reprinted, with permission, from Annual Reviews ©2006.

their catalytic reaction to cause chain termination. However, through mutation and resulting changes in protein, HIV-1 is able to develop resistance towards all treatment options, forcing changes in therapy. While expert knowledge-based mutation tables and algorithms for resistance detection, as well as statistical analysis of clinical data aids optimizing the therapy for individual patients (Lengauer *et al.* 2014), it does not broaden our understanding of resistance in the molecular terms at the protein level. Analysing protein–drug interactions, structural and energetic changes induced by mutations would aid drug development against HIV-1, potentially opening a door to development of new drugs or whole new drug classes (Hwang *et al.* 2017; Markham 2018).

Dissertation Scope and Outline

This dissertation comprises a collection of studies performed using structure- and sequence-based approaches to analyse protein–protein and protein–drug interactions, in particular, a study of HIV-1 protease drug resistance and MAPK docking interactions. Chapter 2 first introduces the chemical theory on enzyme kinetics necessary for understanding the inhibition of HIV-1 protease by antiviral drugs. The background of HIV epidemic is then introduced, the structure of its viral particle, replication, as well as treatment options and resistance, with a focus on the protease in the latter. The second part of Chapter 2 describes phosphorylation and its biological function, with a focus on MAPKs, their role in the signalling networks of the cell, and docking as a mechanism they use for specific recognition of other proteins.

Chapter 3 introduces the experimental techniques used for protein identification, resolving their three-dimensional structures, measuring their enzymatic activity, as well as resistance towards inhibitors, with more emphasis on techniques underlying acquiring of data used in the present study. Computational techniques based on these data used for analysing drug resistance upon mutations in the HIV-1 protease are then introduced. Sequence-based computational techniques used for identifying the interaction partners of MAPK are shortly discussed in Chapter 5 where the reader is referred to relevant literature.

Chapter 4 describes a study we performed on HIV-1 protease drug resistance. For this, computational techniques related to molecular dynamics simulations are used to estimate the effect of various major resistance-associated mutations on the free energy of binding of protease inhibitors. Then these computational estimations are compared to different experimental measurements of the same phenomena. The effect of mutations on the energetics of protein–drug interactions, as well as on the structure of the protein are then analysed to gain insight on the resistance in molecular terms.

Chapter 5 describes our study of MAPKs' interactome based on an analysis of docking interactions through peptides called D-motif. Several different classes of linear motifs mediating these interactions of MAPKs are identified based on a combination of evolutionary and structural considerations. Different methods are applied to train models for different classes of known or experimentally-identified D-motifs, as well as their orthologs, and then used to predict specific interactions in the MAPKs' interactome and gain insight in the associated functional processes and pathways. In a separate study of interactions mediated by linear motifs, peptides targeted by proline-guided kinases, such as MAPK, are grouped in order to find classes of potential downstream interactors.

Finally, Chapter 6 concludes this dissertation by summarizing the results of this work and describing related future directions of research.

2
Biological Background

In this chapter we present the background for the studies reported in this dissertation. Section 2.1 introduces the theory of the enzyme kinetics and inhibition which is necessary to understand the methods used for the resistance factor prediction of HIV type 1 protease. This section also introduces the acid disassociation constant relevant for the protonation state of the active site of HIV-1 protease. The following Sections 2.2 and 2.3 introduce the biological background of HIV and the MAPK signalling.

> *Descriptions of the theoretical foundations follow in argument and in notation (Copeland 2000) and (Copeland 2013) (Section 2.1) and (Schmitt 2005) (Section 2.1.1).*

2.1 Enzyme Kinetics and Inhibition

An enzymatic reaction starts with an encounter between an enzyme E and a substrate S, forming a complex ES (also called Michaelis complex):

$$E + S \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} ES,$$

with the reaction governed by a pseudo-first-order association constant k_{on} and a first-order dissociation constant k_{off}. The equilibrium of the reaction by which complex ES is formed is quantified by the equilibrium disassociation constant K_d (also called K_s for such complex):

$$K_s = \frac{[E]_f[S]}{[ES]} = \frac{k_{off}}{k_{on}}, \tag{2.1}$$

where $[E]_f$ refers to the molar concentration of the unbound enzyme ($[E]_f = [E] - [ES]$), and $[S]$ is considered to approximate the free substrate concentration at saturating substrate concentrations. Rearranging Equation 2.1 gives:

$$[ES] = \frac{[E][S]}{K_s + [S]}. \tag{2.2}$$

Once the complex ES is formed, the substrate is transformed by the enzyme active site to a product P. This reaction is modelled in terms of a catalytic rate (also called turnover or composite rate) constant k_{cat}:

$$E + S \underset{K_s}{\rightleftharpoons} ES \xrightarrow[k_{cat}]{} E + P.$$

The velocity of such reaction, or reaction rate v, is derived in terms of velocity of the conversion of substrate to product:

$$v = -\frac{d[S]}{dt} = \frac{d[P]}{dt} = k_{cat}[ES]. \tag{2.3}$$

At infinite $[S]$, one can define the maximum reaction velocity V_{max}:

$$V_{max} = k_{cat}[E]. \tag{2.4}$$

Combining Equations 2.2, 2.3, and 2.4, one obtains:

$$v = \frac{V_{max}[S]}{K_s + [S]}. \tag{2.5}$$

Under steady-state conditions $\left(\frac{d[ES]}{dt} = 0\right)$, by substituting K_s one arrives at the Henri-Michaelis-Menten equation:

$$v = \frac{V_{max}[S]}{K_m + [S]}, \tag{2.6}$$

where $K_m = \frac{k_{off} + k_{cat}}{k_{on}}$. By setting the system in such a way that $[S]$ equals K_m, one can consider K_m as the substrate concentration that provides a reaction velocity of half of the maximal velocity under conditions when the system is saturated with substrate.

In the presence of an enzyme inhibitor I with disassociation constant K_i, the enzyme turnover is modified in the following way:

$$\begin{tikzcd}
E + S \arrow[r, shift left, "K_s"] & ES \arrow[l, shift left] \arrow[r, "k_{cat}"] & E + P, \\
+ & + & \\
I \arrow[d, shift left] & I \arrow[d, shift left, "\alpha K_i"] & \\
EI \arrow[u, shift left, "K_i"] \arrow[r, shift left, "\alpha K_s"] & ESI \arrow[l, shift left] \arrow[u, shift left] &
\end{tikzcd}$$

where constant α defines the degree to which I modulates the affinity of the enzyme to the substrate. If the inhibitor has no effect on the ES complex formation, then $\alpha = 1$, whereas if it excludes its formation $\alpha = \infty$. The latter class of inhibitors is referred to as competitive inhibitors. The values of α between 1 and ∞ lead to formations in the ESI complex, albeit at a slower rate, a process that is referred to as partial inhibition. While partial inhibitors are known, the majority of therapeutically-used enzyme inhibitors at saturating concentrations disrupt the activity of the enzyme completely.

Screening assays for enzyme targets can be used to measure the potency of different enzyme inhibitors from the so-called progress curves. In such a curve, one measures over time (x axis) the concentration of the product in moles (y axis). In the absence of the inhibitor, the velocity of the enzymatic reaction v_0 is defined by the slope of the progress curve. The velocity of this reaction in the presence of a fixed concentration of inhibitor $[I]$ is denoted by v_i. The remaining enzymatic activity at a fixed inhibitor concentration is then given by the ratio $\frac{v_i}{v_0}$ and is termed the fractional activity. A typically used measure of inhibitor potency, the Inhibitor Concentration required to reduce enzymatic reaction activity by 50% (IC_{50}), is directly related to the fractional activity:

$$\frac{v_i}{v_0} = \frac{1}{1 + ([I]/\text{IC}_{50})}. \tag{2.7}$$

The binding affinity of inhibitor K_i is related to its IC_{50} by the following equation (Cha 1975):

$$\mathrm{IC}_{50} = K_i \left(1 + \frac{[S]}{K_m}\right) + \frac{1}{2}[E]_T, \tag{2.8}$$

where $[S]$ is a fixed substrate concentration and $[E]_T$ is the total enzyme concentration. For most protein–ligand binding interactions, the inhibitor concentration required to achieve 50% inhibition of the enzyme far outstrips the enzyme concentration, meaning that the amount of inhibitor sequestered by forming the EI complex is just a small fraction of the total concentration of I. This renders $[E]_T$ in Equation 2.8 negligible, leading to the Cheng-Prusoff equation for competitive inhibitors (Cheng and Prusoff 1973):

$$\mathrm{IC}_{50} = K_i \left(1 + \frac{[S]}{K_m}\right). \tag{2.9}$$

Importantly in thermodynamic terms, K_i can be converted to the change in Gibbs free energy:

$$\Delta G = RT \ln K_i, \tag{2.10}$$

where R is the ideal gas constant (it can also be expressed as $R = k_\beta N_A$, with k_β denoting the Boltzmann constant and N_A the Avogadro constant) and T is the absolute temperature.

In thermodynamics, the change in Gibbs free energy is typically defined in the following way:

$$\Delta G = \Delta H - T\Delta S, \tag{2.11}$$

where ΔH is the change in enthalpy and ΔS is the change in entropy.

The reaction rate can also be defined in terms of the enthalpy change in the system. Namely, the heat change in system Q during a catalytic reaction is related to the concentration of product $[P]$ generated and the molar enthalpy of the following reaction (Mazzei *et al.* 2014):

$$Q = n\Delta H = V\Delta H[P], \tag{2.12}$$

where n is the number of moles of product generated and V is the total volume. The reaction rate in Equation 2.3 can then be redefined as:

$$v = \frac{d[P]}{dt} = \frac{1}{V\Delta H}\frac{dQ}{dt}. \tag{2.13}$$

This relation is used to determine the kinetic parameters of reaction using Isothermal Titration Calorimetry (ITC) (see Section 3.1.2).

2.1.1 Acid Disassociation Constant

Acids and bases can modulate the equilibrium of water self-ionization:

$$\mathrm{H_2O} + \mathrm{H_2O} \rightleftharpoons \mathrm{H_3O^+} + \mathrm{OH^-}.$$

By Brønsted's definition acids are proton donors, while bases are proton acceptors, giving the following acid-base equilibrium expression:

$$\mathrm{AH} + \mathrm{B} \rightleftharpoons \mathrm{A^-} + \mathrm{BH^+}.$$

Through proton donation acid AH becomes a conjugated base A^- and through proton acceptance base B becomes a conjugated acid BH^+. In case of strong bases, such as sodium hydroxide (NaOH), solvent molecules are forced to donate a proton; in case of strong acids, such as hydrochloric acid (HCl), almost every hydrochloric acid molecule donates to the solution its proton. In case of weak bases and acids a "real" equilibrium exists between acid and base, or conjugated base and conjugated acid.

Functional groups of organic molecules are often weak bases or acids. So carboxyl groups (R-COOH) can donate their proton to make a carboxylate group ($R\text{-}COO^-$), while the base amino group ($R\text{-}NH_2$) can accept a proton to become ammonium ions ($R\text{-}NH_3^+$). For the dissociation constant of acids with H_2O the following equations apply:

$$\begin{aligned} K &= \frac{[H_3O][A^-]}{[AH][H_2O]} \\ K_a &= \frac{[H^+][A^-]}{[AH]}, \end{aligned} \tag{2.14}$$

where in case of K_a the concentration of H_2O is considered to be almost unchanged. The negative logarithm $-\log K_a$ of this dissociation constant is referred to as pK_a. Thus, pK_a measures the strength of the acid, with a low pK_a value corresponding to a strong acid. If the proton concentration in the aqueous solution, pH, equals pK_a, then acid and conjugated base are present in equal concentration. By increasing pH, an acid will turn into a conjugated base.

Description of the biological foundations in Section 2.2 are partly adapted from (Bastys 2012).

2.2 Human Immunodeficiency Virus

2.2.1 General Information

Human Immunodeficiency Virus is a retrovirus that causes Acquired Immunodeficiency Syndrome (AIDS), a condition in which a dysfunctional immune system is unable to cope with opportunistic infections and cancer. The virus was first discovered in the early 1980s and is thought to have originated from a transmission from non-human primates to humans (Sharp and B. H. Hahn 2011; Faria *et al.* 2014). Currently it infects 36.7 million people with 1.8 million new infections per year (Joint United Nations Programme on HIV/AIDS (UNAIDS) 2017b). The infections occur through transmission of bodily fluids, primarily being sexually transmitted (Rom and Markowitz 2007). HIV compromises the human immune system by infecting helper T-cells (cluster of differentiation 4 $(CD4)^+$ type), macrophages, and dendritic cells, eventually leading to their destruction.

2.2.2 Subtypes

Two types of HIV are currently recognised: HIV-1 and HIV-2. The Simian Immunodeficiency Virus (SIV) infecting the chimpanzee subspecies *Pan troglodytes troglodytes* (SIVcpz) is considered to be the source of HIV-1 and gorilla SIV (SIVgor) through cross-species transmission (Keele *et al.* 2006; Van Heuverswyn *et al.* 2006). SIVsmm, infecting sooty mangabey (*Cercocebus atys atys*), is thought to be the source of HIV-2 (Hirsch

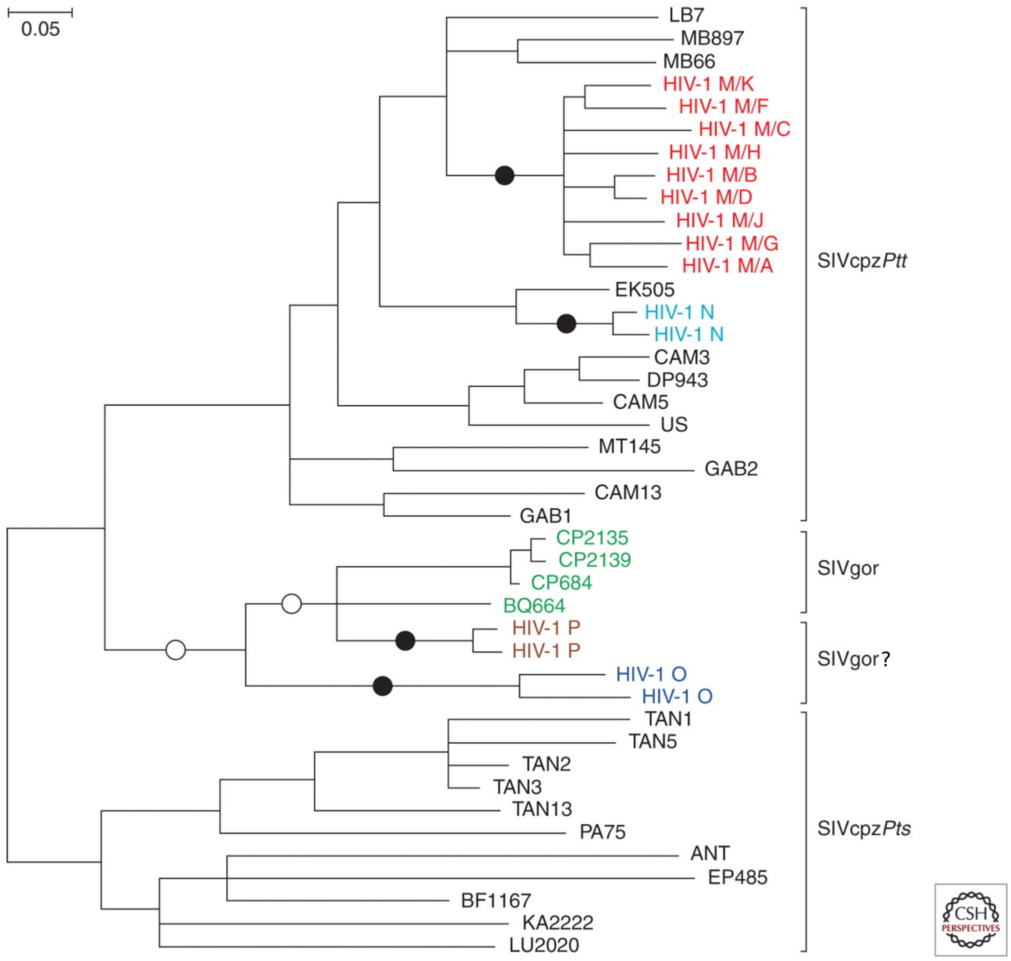


Figure 2.1: A phylogenetic tree showing relationship between SIVcpz, HIV-1, and SIVgor. SIVcpz sequences are depicted in black, SIVgor in green, and HIV-1 in the other colours respectively. Black circles reflect cross-species transmission-to-humans and white circles reflect two possible alternative branches for chimpanzee-to-gorilla transmission. Question mark indicates SIVgor as likely source of HIV-1 groups O and P, with particularly strong evidence for the former (D'arc *et al.* 2015). Adapted from (Sharp and B. H. Hahn 2011) and reprinted, with permission, from Cold Spring Harbor Laboratory Press ©2011.

et al. 1989). Of the two types of HIV, HIV-2 is considered to be a lesser health hazard as it accounts for fewer infections worldwide (World Health Organization 2016), as well as it has lower transmission rate and pathogenicity (Kannangai *et al.* 2012).

HIV-1 is divided into groups M, N, O, and P, all of which are thought to have originated from separate introductions of SIVcpz and SIVgor into the human population (Gao *et al.* 1999; Plantier *et al.* 2009; Vallari *et al.* 2011; D'arc *et al.* 2015) (Figure 2.1). Group M is estimated to account for the vast majority of all HIV infections in humans (A. F. Santos and Soares 2010; Hemelaar 2012). Based on the genetic variation within the M group, the group is further divided into subtypes: A, B, C, D, F, G, H, J, K and their hybrids called Circulating Recombinant Forms (CRFs). A commonly used HIV-1 reference sequence called HXB2 is from group M subtype B. From here on we will refer to HIV-1 as HIV.

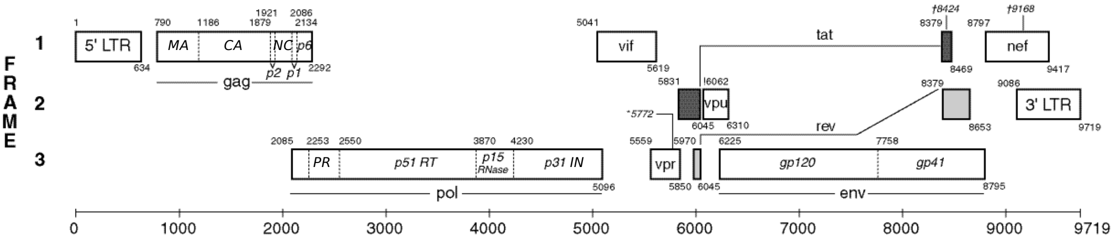


Figure 2.2: HIV genome structure. Adapted from (Los Alamos National Laboratory 2017).

2.2.3 Genome and Viral Particle Structure

The HIV particle is about 120 nm in diameter. It consists of 15 types of proteins and a genome coding Ribonucleic Acid (RNA) of approximatively 9700 bp length (Figure 2.2). HIV proteins can be roughly divided into three groups: structural, enzymatic, and accessory proteins (Frankel and J. A. Young 1998). Some of these proteins are expressed as polyproteins and are cleaved by the HIV protease during the virus maturation. The structural proteins are: (i) four proteins from the group-specific antigen (gag) polyprotein: matrix (MA), capsid (CA), nucleocapsid (NC), and p6; and (ii) the two proteins from the envelope (env) polyprotein: glycoprotein 120 (gp120) and glycoprotein 41 (gp41). Together these proteins make up the virus capsid and, along with a portion of the lipid membrane borrowed from the host cell, the viral envelope. Proteins from the polymerase (pol) polyprotein, protease (PR), reverse transcriptase (RT), and integrase (IN), are responsible for the enzymatic functions in the virus. The rest of the proteins, negative regulatory factor (Nef), regulator of expression of virion proteins (Rev), trans-activator of transcription (Tat), virion infectivity factor (Vif), viral protein R (Vpr), and viral protein U (Vpu), perform various auxiliary functions throughout the HIV replication cycle. At both ends of the viral RNA, there are the long terminal repeats (LTRs), which include transcription promoters.

2.2.4 Replication Cycle

The first step performed by HIV to enter into its target cell is through interaction with it via binding of the viral protein gp120 to the CD4 receptor protein and coreceptors C-C chemokine receptor type 5 (CCR5) or C-X-C chemokine receptor type 4 (CXCR4) of the target cell. The preference of HIV to bind to a certain type of coreceptor depends on the strain of the virus and is called viral tropism. Subsequently, a conformational change in another envelope protein, gp41, induces fusion of the membranes of the virus particle and host cell. Then the virion core is uncoated, exposing viral RNA. With the help of RT, RNA is reversely transcribed to Deoxyribonucleic Acid (DNA), a double-stranded version of which is transported to the cell nucleus with MA mediating the process (Gallay *et al.* 1995). In the nucleus it is then integrated into the DNA of the host cells by IN.

A copy of the viral genome can reside dormant in a host cell for a long time. At the next active stage of the virus replication cycle, viral genome is transcribed from its integrated copy in the host DNA with the help of promoters residing in LTR, with Tat protein enhancing the process (Kao *et al.* 1987; Roy *et al.* 1990). At this stage, a set of both spliced and full-length RNAs is transported from the host nucleus to the cytoplasm, a development controlled by Rev protein (Hope 1999). The viral Messager RNA (mRNA) is then transcribed in the cytoplasm, and a new viral particle starts to

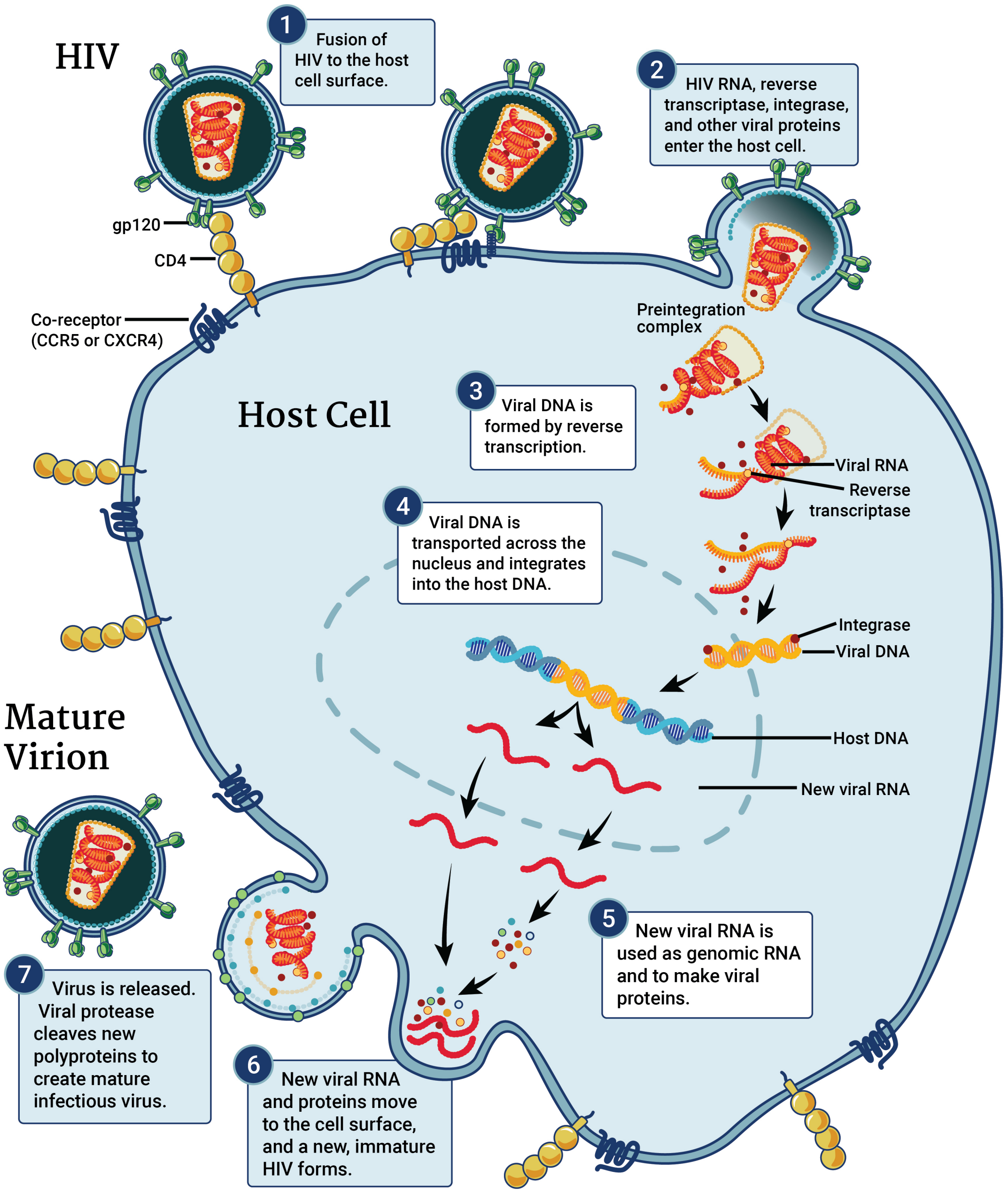

HIV
1
Fusion of HIV to the host cell surface.
2
HIV RNA, reverse transcriptase, integrase, and other viral proteins enter the host cell.
gp120
CD4
Co-receptor (CCR5 or CXCR4)
Preintegration complex
Host Cell
3
Viral DNA is formed by reverse transcription.
Viral RNA
Reverse transcriptase
4
Viral DNA is transported across the nucleus and integrates into the host DNA.
Integrase
Viral DNA
Mature Virion
Host DNA
New viral RNA
5
New viral RNA is used as genomic RNA and to make viral proteins.
7
Virus is released. Viral protease cleaves new polyproteins to create mature infectious virus.
6
New viral RNA and proteins move to the cell surface, and a new, immature HIV forms.

Figure 2.3: HIV life cycle. Taken from (*National Institute of Allergy and Infectious Diseases* 2010).

form: gag polyproteins associate with the inner membrane side and the env polyproteins anchor to the outer membrane and transmembrane regions after the degradation of CD4+ receptors (carried through by proteins Vpu and Nef) (Bour *et al.* 1995). The viral RNA is brought to the assembling virion by NC. After the virus particle has finished forming, it buds from the cell. During the budding process, Vif protein protects the virion by inhibiting the viral genome hypermutation-inducing human protein, apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G (APOBEC3G), from entering the particle (Stanley *et al.* 2008; Henriet *et al.* 2009). At the same time and also later as a free particle, the virus undergoes a maturation process, where gag and pol polyproteins are cleaved by PR. At this stage the capsid forms from CA proteins within the virion protecting its genome.

2.2.5 Treatment and Drug Resistance

Despite some promising recent advances in the development of vaccine against HIV (Barouch *et al.* 2018), there is currently no licensed vaccine for it. Pre-Exposure Prophylaxis (PrEP) and Post-Exposure Prophylaxis (PEP) are recommended for people with substantial risk of HIV infection, and Antiretroviral Therapy (ART) is recommended for every infected individual to suppress the virus (World Health Organization 2016). Antiretroviral (ARV) drugs are divided into several categories roughly based on the stage of replication cycle or the part of virus that they target. Since the drugs have different targets and act at different stages of the HIV replication cycle, it is possible to use them in combinations. Except for select few individual cases (Allers *et al.* 2011; Jessen *et al.* 2014), with no cure available ART remains the only means to combat the virus, leading to near-normal life expectancy of the HIV-infected patients (Antiretroviral Therapy Cohort Collaboration 2017).

In the following, we discuss the action mechanisms of the U.S. Food and Drug Administration (FDA)-approved inhibitors action.

Entry Inhibitors

Entry inhibitors target the interactions of gp120 and gp41 proteins with the cell. The drug maraviroc binds to the hydrophobic pocket of the CCR5 receptor, which is believed to change the conformation of the extracellular loops of CCR5, restricting the interaction of gp120 with this receptor (Tilton and Doms 2010). On the other hand, enfuvirtide, a biomometic peptide, is a competitive binder to the heptad-repeat domain 1 of the gp41 (Tilton and Doms 2010). Disrupting the proper arrangement of the domains of gp41 precludes the fusion of the viral particle with the cell's membrane. The latest drug to be approved for ART, ibalizumab, is from a subcategory of entry inhibitors, namely post-attachment inhibitors (Markham 2018). It binds to the domain 2 of CD4 receptor of the human cell which, while allowing for binding of receptor with gp120, induces structural changes prohibiting the interaction of gp120 with both CCR5 and CXCR4 coreceptors (S. A. Iacob and D. G. Iacob 2017). Thus, the main advantage of ibalizumab in comparison to maraviroc is that the latter is only effective against CCR5-tropic viruses, whereas the former can potentially target both gp120-tropicity variants.

Nucleoside and Nucleotide Reverse-Transcriptase Inhibitors

Nucleotide or Nucleoside Reverse-Transcriptase Inhibitors (NRTIs) are historically the first FDA-approved drugs against HIV (*Antiretroviral Drugs Used in the Treatment of HIV Infection* n.d.). They compete with the natural Deoxynucleoside Triphosphates (dNTPs) for incorporation by RT in the viral DNA (Cihlar and Ray 2010). NRTIs, despite their structural diversity, all mimic the structural contacts of the natural dNTP in the RT active site. Once they are incorporated into the DNA, because they lack 3'-hydroxl, no further nucleotides can be incorporated, and thus the DNA synthesis is terminated.

Non-Nucleoside Reverse-Transcriptase Inhibitors

Like NRTIs, the Non-Nucleoside Reverse-Transcriptase Inhibitors (NNRTIs) also interfere with the reverse transcription process. However, they bind to the RT itself, specifically to the hydrophobic pocket close to the active site of the polymerase (the so-called NNRTIs binding pocket). While the exact molecular mechanism of action is not clear, it is thought that the structural changes induced by this binding reduce the catalytic activity of RT (Sarafianos *et al.* 2009).

Integrase Strand-Transfer Inhibitors

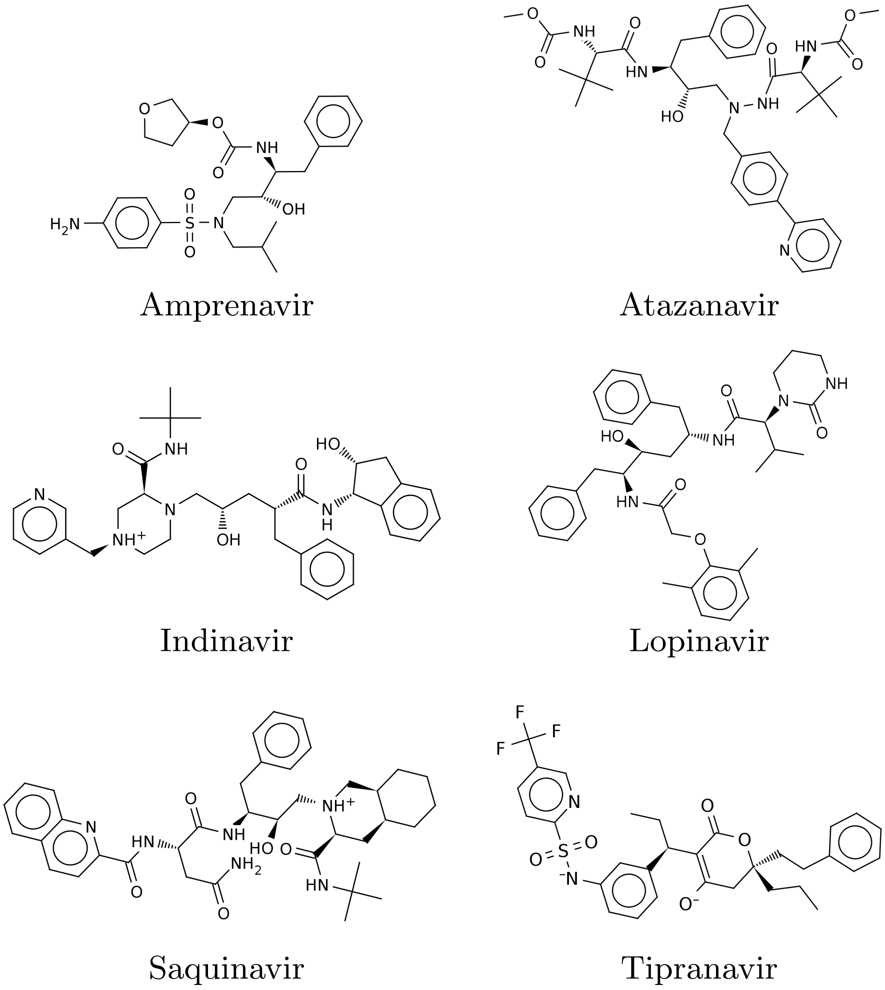
The Integrase Strand-Transfer Inhibitor (INI) targets the next stage of the HIV replication cycle, namely the incorporation of viral DNA into the host cell's genome. Due to the complexity of this reaction, as well the fact that only two such integration events per 50–100 copies of IN present in the host cell after infection are sufficient for production of provirus, INIs were historically the last category of ART drugs to be approved. Indeed, binding of INI to the active site of IN is not well understood, but it has been suggested that viral DNA might be part of the IN binding site (McColl and X. Chen 2010). By chelating the metal ions in the active site of IN INIs inhibit the catalytic activity of the enzyme.

Protease Inhibitors

Protease Inhibitors (PIs) were an early-developed class of ARV drugs for HIV treatment. This class of drugs targets the maturation of the virus, specifically the PR enzyme. The two active site aspartic acids in the substrate binding pocket of PR, which is a symmetric dimer, cleave the gag and gag-pol polyproteins of the virus. These polyproteins have different residues around the cleavage sites, so it has been suggested that PR recognises the asymmetric shape of the peptide rather than the exact sequence itself (Prabu-Jeyabalan *et al.* 2002). All PIs competitively bind in the substrate binding pocket of PR. With the exception of Tiprananavir (TPV), which has a dihydropyrone ring at its core, they are all peptidomimetic and have a hydroxyethylene core to protect them against the cleavage by PR (Wensing *et al.* 2010) (Figure 2.4). Inhibitor Amprenavir (APV), which has previously been aproved by FDA, is no longer marketed as it has been replaced by its prodrug Fosamprenavir (FPV), meaning FPV gets metabolized in the body to APV.

Drug Resistance

Due to the high mutation rate of HIV, reported to be in the range of $1.2 - 3.4 \times 10^{-5}$ (Mansky and Temin 1995; Abram *et al.* 2010), resistance due to mutation in the HIV genome is one of the major issues in HIV treatment issues. Human Immunodeficiency Virus Drug Resistance (HIVDR) mutations can occur *de novo* or they can be transmitted during the infection. With over half of the infected individuals receiving ART (Joint United Nations Programme on HIV/AIDS (UNAIDS) 2017a), the latter source for HIVDR is becoming more prevalent with a recent World Health Organization (WHO) report suggesting over 10% of pretreatment HIVDR to NNRTIs in 6 out of 11 surveyed countries (World Health Organization 2017). Drug resistance is a relevant issue for all ART drugs classes. For the competitive inhibitors, Resistance-Associated Mutations (RAMs) can either unfavourably affect the binding affinity of the inhibitors, or improve the fitness of the virus. In case of PIs, the first category has been referred to as primary



Amprenavir

Atazanavir

Indinavir

Lopinavir

Saquinavir

Tipranavir

Figure 2.4: HIV PIs analysed in this dissertation. Figure taken from (Bastys *et al.* 2018).

mutations, and the second category of mutations is called compensatory mutations, as they act to compensate the negative effect of primary mutations on the catalytic activity of the protein (Henderson *et al.* 2012).

In line with PIs binding in the substrate binding pocket of protease (further referred to as *protease binding pocket*), most of the major RAMs towards PIs appear in the different structural elements composing the binding pocket (Figure 2.5). This includes the active site loop (residues D30, V32, and L33), the so-called 80s loop (residues V82 and I84) — both forming the sides of the pocket, or the flap region of the protease (residues M46, I47, G48, I50, and I54). Yet several RAMs are also found at sites that are distant to the binding pocket sites, e.g. residues N88 and L90 in the protease's α-helix or residue L76 in the hydrophobic core. Hence, the effect of these mutations on inhibitor binding is likely to be not through direct interactions.

Descriptions of the biological foundations in Section 2.3 are partly adapted from (Bastys 2012).

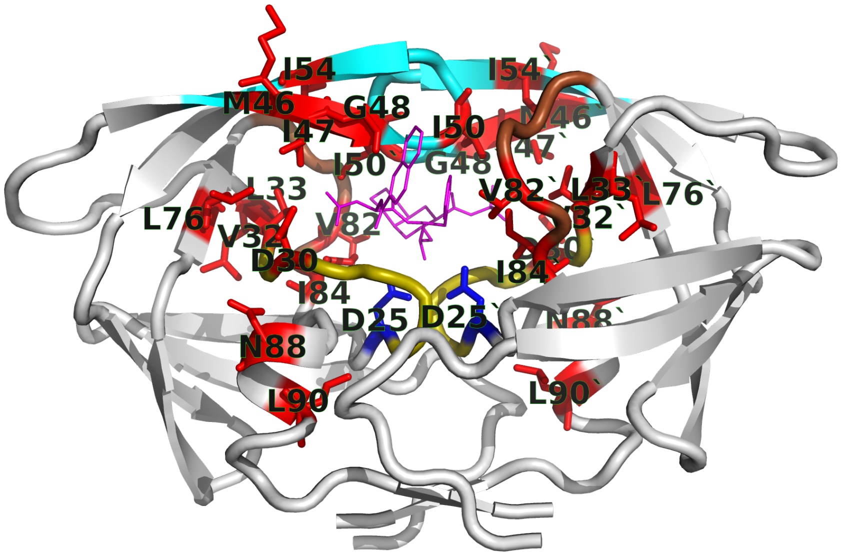


Figure 2.5: HIV protease structure. The flap region is depicted in cyan, 80s loop in brown, active-site proximate loop in olive colours. Major resistance-associated mutation sites (red), catalytic site residues (blue), and bound inhibitor (magenta) are shown in sticks model.

2.3 Mitogen-Activated Protein Kinases

2.3.1 Protein Phosphorylation

Post-Translational Modification (PTM) is a modification of a standard amino acid by a covalently bound chemical group. Although present in most organisms, it is by far more wide-spread in higher eukaryotes, in which estimated 5% of the genome is dedicated to enzymes carrying out PTMs (Walsh *et al.* 2005). PTMs are, for example, employed in signalling cascades that transmit signals from extracellular stimuli to the nucleus, thus contributing to a higher complexity of these organisms (Deribe *et al.* 2010).

The most prevalent PTM is phosphorylation (Khoury *et al.* 2011). It is estimated that between one- and two-thirds of all proteins in a typical mammalian cell may be phosphorylated (S. A. Johnson and Hunter 2005; Vlastaridis *et al.* 2017). During phosphorylation, a phosphate group (PO_4) from a high energy donor molecule (such as Adenosine Triphosphate (ATP)) is transferred to a substrate protein and bound covalently to an amino acid. In case of eukaryotes, primarily serine, threonine, or tyrosine can be phosphorylated, typically by a dedicated enzyme called kinase, whereas in prokaryotes a two-component signalling, involving histidine kinase and regulator protein, is most common (Stock *et al.* 2000). Protein phosphorylation has been suggested to have diverse effects on the protein structure. An early comparative study on the effect of phosphorylation for more than a dozen different proteins suggested as the main effect of phosphorylation to be a change in protein conformation to accommodate the electrostatic effects of the phosphate (L. N. Johnson and Lewis 2001). However, a later large scale study on phosphorylation effects suggested that in most cases changes in the protein structure upon phosphorylation are modest, as in only 13% of the cases the comparison of phosphorylated and unphosphorylated forms of protein exhibited changes > 2 Å in root mean square deviation (Xin and Radivojac 2012). Phosphorylation can also induce large structural effects indirectly, such as the phosphorylation site being recognized by an enzyme which then triggers a conformational switch (Lu *et al.* 2002) or phosphorylation acting as an allosteric effector through which local changes propagate to larger tertiary ones (Nussinov *et al.* 2012). Another observed effect of phosphorylation was modulating the disorder-to-order transitions (Nelson *et al.* 2005).

Among other roles of phosphorylation, its importance in affecting protein–protein interactions is corroborated by the tendency of the phosphorylated sites to be found on the protein–protein interaction interfaces (Nishi *et al.* 2011). The conservation of phosphorylated residues in these interfaces is also higher than that of other residues. It has also been reported that of the residues found on the binding interfaces, serine, threonine, and tyrosine from the disordered regions of the interface tend to be more frequently phosphorylated compared to the same residues from the ordered interfaces (Nishi *et al.* 2013). This is in line with the overall tendency of disordered regions of proteins to be enriched in phosphorylation sites (Iakoucheva *et al.* 2004; Gsponer *et al.* 2008; Singh 2015). If multiple phosphorylation sites are found in a protein, they tend to be clustered in the same region of the protein (H. Li *et al.* 2009; Schweiger and Linial 2010; Freschi *et al.* 2014), with the majority of the serine and threonine phosphosites (phosphorylation consensus sequences) found within four residues of each other (Schweiger and Linial 2010). Co-occurance of multiple phosphorylation sites can play a role in increasing the binding affinity of the protein interactions (Ferreon *et al.* 2009) and the accuracy of the conformational change modulation (Kumar *et al.* 2012).

2.3.2 Mitogen-Activated Protein Kinases Signalling

Over 500 kinases have been identified in the human genome (Manning *et al.* 2002). With a few exceptions (Dhanasekaran and E. P. Reddy 1998; Fuhs and Hunter 2017), mammalian kinases are split into serine/threonine kinases and tyrosine kinases groups. Mitogen-activated protein kinases (MAPKs) are a family of serine/threonine kinases that are broadly spread and well conserved in eukaryotes. The divergence of two of the MAPK subgroups, extracellular signal-regulated kinase (ERK)-like and p38 mitogen-activated protein kinase (p38)-like, pre-dates the divergence of animals and fungi as proven by the existence of Fus3 (ERK-like) and Hog1 (p38-like) yeast kinases (Good *et al.* 2009; Duch *et al.* 2012). The animal-specific MAPK subgroups—mitogen-activated protein kinase 3 (ERK1)/mitogen-activated protein kinase 1 (ERK2), mitogen-activated protein kinase 7 (ERK5), c-Jun N-terminal kinases (JNKs), p38s—appear early in the metazoan evolution, as these kinases can be found in choanoflagellates, sponges, and cnidarians (King *et al.* 2008; Chera *et al.* 2011).

With the exception of the "atypical" MAPKs, such as mitogen-activated protein kinase 6 (ERK3), mitogen-activated protein kinase 4 (ERK4), mitogen-activated protein kinase 15 (ERK7), and nemo-like kinase (NLK) (Coulombe and Meloche 2007), MAPKs usually form a three-tier cascade with two phosphorylation events, i.e. MAP3K–MAP2K–MAPK. Here a mitogen-activated protein kinase kinase kinase (MAP3K) phosphorylates a mitogen-activated protein kinase kinase (MAP2K), which eventually phosphorylates a MAPK. These cascades can be initiated by diverse extracellular stimuli and result in various cellular responses (Figure 2.6). ERK1 and ERK2 are mostly involved in cell division (meiosis and mitosis) and cell differentiation. They are activated by a number of diverse stimuli, which include growth factors, cytokines, viral infections, carcinogens, ligands for G protein coupled receptors, and transforming agents (G. L. Johnson and Lapadat 2002). Mitogen-activated protein kinase 8 (JNK1), mitogen-activated protein kinase 9 (JNK2) and mitogen-activated protein kinase 10 (JNK3) are stress-activated

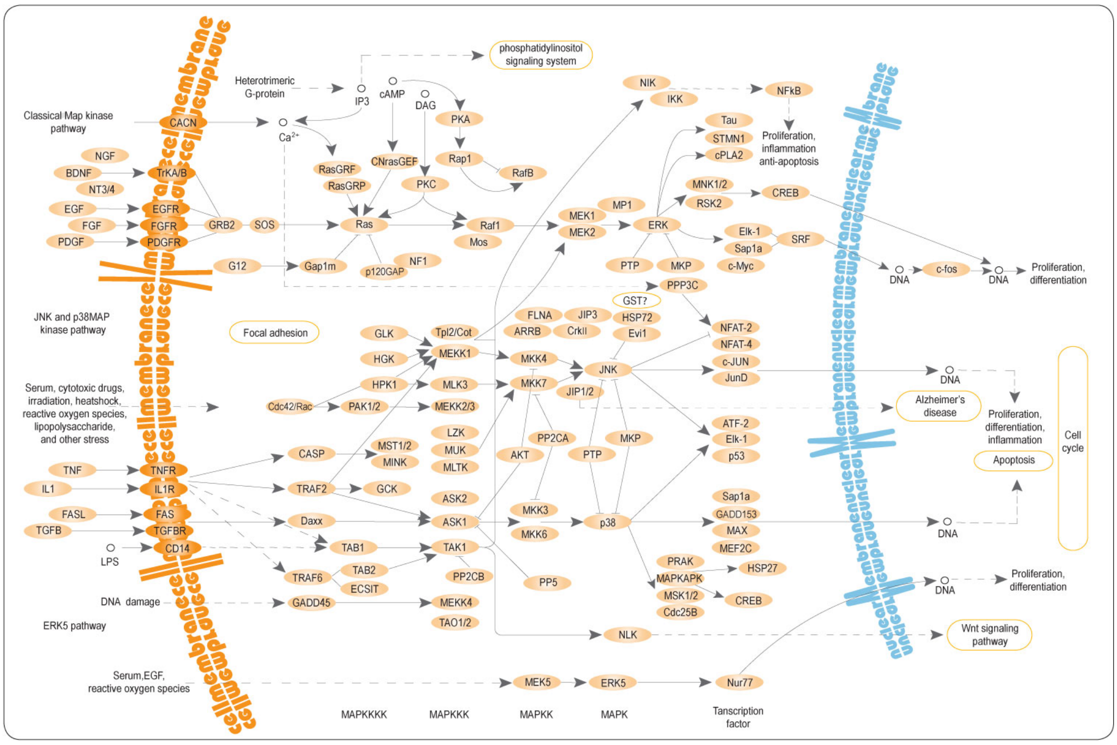


Figure 2.6: MAPK activation cascades. Taken from (Wikipedia contributors 2018).

protein kinases. They were originally identified as phosphorylating the DNA-binding protein c-Jun. JNKs have also been shown to participate in cell death (apoptosis) and immune response by mediating T-cell differentiation, processes in which JNKs' targets include p53, Elk-1, and nuclear factor of activated T-cells 4 (NFAT4) (Vlahopoulos and Zoumpourlis 2004). The p38 kinase group includes α, β, γ, and δ kinases, of which mitogen-activated protein kinase 14 (p38α) is best characterised. Similarly to JNKs, p38s are involved in apoptosis and immune response and their activating stimuli include cytokines, hormones, ligands for G protein-coupled receptors, osmotic shock, and heat shock (G. L. Johnson and Lapadat 2002). Studies of upstream JNK- and p38-activating kinases have indicated an important role of both of these families in development of various organs in animals (Ganiatsas *et al.* 1998; A. Pearlman *et al.* 2010; Le Goff *et al.* 2016; Spielmann *et al.* 2016; Wade *et al.* 2016). Because of similarity in the functions of their pathways, some proteins (e.g. MKK4) have been suggested to directly interact with both JNKs and p38s (Derijard *et al.* 1995). On the other hand, the protein deleted in colorectal carcinoma (DCC) has been suggested to be involved in both JNK and ERK pathways, albeit directly interacting only with the ERK (Arakawa 2004; Qu *et al.* 2013; Ma *et al.* 2010). *In vitro* experiments have also indicated promiscuous interaction between peptides of MAP2Ks and both p38α and ERK2 (Garai *et al.* 2012) (see below). Another ERK, ERK5, has been suggested to be involved in heart and vessel development (Regan *et al.* 2002), with Mef2 proteins (e.g. myocyte-specific enhancer factor 2A (MEF2A)) as some of its reported targets (Yang and Gabuzda 1998).

Target Recognition

Just like cyclin-dependant kinases (CDKs) and glycogen synthase kinase 3 (GSK3), MAPKs are referred to as proline-directed kinases, since they have a preference for the

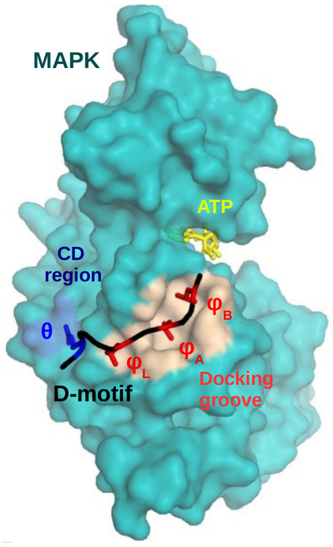


Figure 2.7: A JNK1 structure (cyan) with two distinct docking groove regions, the CD groove (blue) and the hydrophobic docking groove (light-brown), with a bound D-motif peptide (black) and an ATP molecule (yellow) (Garai *et al.* 2012). Taken from (Zeke *et al.* 2015).

phosphorylated residue to be followed by a proline ([ST]P) (Lu *et al.* 2002). Due to the frequency of such short consensus sequence in the proteome, additional mechanisms have been suggested to ensure signal fidelity. In the case of ERK5, a mediation of the interaction with its target protein by a separate domain has been observed (Glatz *et al.* 2013). Different sites on MAPKs and their targets, distant to both the active site and the phosphoacceptor site respectively, were suggested to regulate the interaction (J. A. Smith *et al.* 2000; Tanoue *et al.* 2001).

However, the best studied mechanism facilitating MAPK phosphorylation is through interactions with peptide sequence on the target called D-motif (Sharrocks *et al.* 2000; C.-I. Chang *et al.* 2002; Mooney and Whitmarsh 2003; Reményi *et al.* 2005; Ma *et al.* 2010; Garai *et al.* 2012; Glatz *et al.* 2013). Similar kind of interactions, referred to as docking interactions or docking, have been reported for various other serine/threonine kinases, such as CDKs, GSK3, and PDK1 (for review see (Biondi and Nebreda 2003)) but also for serine/arginine kinases, such as SRPKs (J. C. K. Ngo *et al.* 2005; Long *et al.* 2018). D-motif, which typically resides in intrinsically disordered regions of the target proteins (Neduva and Russell 2005; Garai *et al.* 2012), binds to the site of MAPK composed of the so-called hydrophobic docking groove and CD groove, improving the interaction between the two proteins (Figure 2.7). D-motifs could be described with a common loosely-defined consensus: $\theta_{1-2}x_{0-5}\phi_L x_{1-2}\phi_A x\phi_B$, where ϕ_L, ϕ_A, and ϕ_B denote positions that are typically filled by hydrophobic amino acids, with L, A, and B referring to the lower pocket, and pockets A and B respectively, while θ denotes positively charged (arginine or lysine) and x denotes any amino acid. During the interaction between the D-motif and MAPK, the hydrophobic and charged part of D-motif bind to the hydrophobic docking groove and CD groove, respectively.

"Reverse D-motifs" have also been suggested as binding mediators. They are similar to D-motifs in that they use the same mechanism of interaction, but in them the motif sequence is reversed compared to standard D-motif that runs from the N- to C-terminus (Garai *et al.* 2012). However, only few such examples are known so far.

Unlike for other motifs, such as FxFP which mediates binding with some MAPKs in the so-called DEF groove and is typically located 10 residues away from the phosphoacceptor site towards the C-terminus, the location of D-motifs is variable (Bhattacharyya *et al.* 2006). Employment of D-motifs for interaction is not limited to MAPK targets, but D-motifs have also been suggested to mediate the interaction with MAPK regulators, such as phosphatases (e.g. protein tyrosine phosphatase non-receptor type 7 (HePTP) (Zhou *et al.* 2006)), and MAP2Ks (Enslen *et al.* 2000; Garai *et al.* 2012), as well as mediating assemblies of protein scaffold complexes (e.g. c-Jun-amino-terminal kinase-interacting protein 1, also called MAPK8IP1 or JNK1 interacting protein 1 (JIP1) (Sharrocks *et al.* 2000)).

3 Experimental and Computational Techniques

This chapter introduces some experimental and computational techniques relevant for the work presented in this dissertation. Section 3.1.1 discusses methods used for protein structure identification, whereas Section 3.1.4 discusses the underlying assay used in performing the experiments described in Chapter 5. Relevant to Chapter 4, the methods for determining protein enzymatic activity, inhibition, and its drug resistance are discussed in Sections 3.1.2 and 3.1.3, respectively. Finally, aspects relevant to the mutation modelling in protease (Section 3.2.1), the determination of the reference protonation state of its active site (Section 3.2.5), its simulation (Section 3.2.4), and the theory behind free energy differences and their calculation (Sections 3.2.2 and 3.2.3) later used in Chapter 4 are presented.

> *Descriptions of experimental techniques foundations of X-ray Crystallography follow in argument and in notation (Ilari and Savino 2008), Nuclear Magnetic Resonance Spectroscopy follows (Markley et al. 2009), Cryo-TEM and Cryo-ET follows (Volkmann and Hanein 2009), and Section 3.1.2 follows (Copeland 2000), respectively.*

3.1 Experimental Methods for Data Acquisition

3.1.1 Protein Structure Determination

X-ray Crystallography

Since resolving the crystal structure of myoglobin in 1957 (Kendrew *et al.* 1958), X-ray crystallography has played the central role in structural biology. Close to 90% of all the structures in the Protein Data Bank (PDB) have been resolved using this method (Berman *et al.* 2000; *PDB Current Holdings Breakdown* n.d.), which is based on the observation that atoms form a regular, repeating pattern within the crystals. When stricken by X-rays, this allows for reading the diffraction pattern from the resulting waves.

Protein sample preparation in X-ray crystallography requires the protein to be crystallized, which is mainly a trial-and-error procedure. A highly purified protein is dissolved in a solvent and the solution is then brought to supersaturation, during which time the crystal nuclei can start to form, creating a basis from which the crystal can grow. When X-rays are shone on the crystal, they are scattered by the crystal planes. Most of these rays cancel out through destructive interference, but some of them add constructively as defined by the Bragg's law (W. H. Bragg and W. L. Bragg 1913). As a consequence of this law, the wavelength of radiation must be similar to interatomic distances. The diffraction pattern of the crystal can be recorded on a photographic film or Charged-Coupled Device (CCD) image sensor. To measure all the diffracted reflection intensities of the crystal, the crystal is rotated during the image recording procedure. Then the atom positions may be identified based on the distribution of electron density in the diffraction pattern. This involves estimating the phase of the waves recorded in the pattern (phase problem), which, under certain conditions, can be performed directly from the recorded data or by fitting to a related structure.

X-ray crystallography suffers from a number of difficulties, primarily related to the crystal preparation, but also their handling, collecting the diffraction patterns and their interpretation. Because hydrogen atoms only have one electron, it is difficult to capture its density, thus hydrogen typically cannot be resolved by this technique. Neutron diffraction, a related technique where neutrons are shot instead of X-rays, can alleviate this, as they scatter from the nuclei, enabling to record the positions of hydrogen atoms as well.

Nuclear Magnetic Resonance Spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy, at approximately 9%, is historically the second most used method for macromolecular structure determination. It can be considered complementary to X-ray crystallography in that it can resolve structures which are partially disordered, have multiple stable conformations or are not amenable to crystallization. In NMR, transitions between the spin states of nuclei with magnetic activity, e.g. (^{1}H), (^{13}C), and (^{15}N), when placed in an NMR tube of an NMR spectrometer, are investigated. These stable isotopes have a nuclear spin of $\frac{1}{2}$ and thus two spin states: one paired with the external magnetic field and one opposing the external field. The magnetic moment of each nuclei orients itself around the external magnetic field, being also influenced by other magnetic fields, such as neighbouring spins in the molecule and preprogrammed radio-frequency pulses of spectrometer. Applying these pulses flips the magnetic spin of the nuclei, which induces an oscillating current, called the free induction decay, in the coil that detects the signal (resonance). Fourier transformation of this time-domain signal converts it to frequency-domain signal, called NMR peak. More complicated pulse sequences, i.e. from multiple orthogonal frequency axes, allow for reading out multidimensional NMR spectra of the nuclei. Due to the spinning effects of the neighbouring nuclei, the resonance of the nuclei deviates from a reference frequency or a standard compound (e.g. tetramethylsilane, trimethylsilylpropanoic acid, or 4,4-dimethyl-4-silapentane-1-sulfonic acid), called chemical shift. For proteins, this is sufficient to provide reliable information about their secondary structure. Additional structural restraints in NMR can be obtained from the so-called nuclear Overhauser effect (nOe) spectrum (Overhauser 1953; Anderson and Freeman 1962), which provides $^{1}\mathrm{H} - {}^{1}\mathrm{H}$ distance constraints, as well as empirical torsion angle constraints based on chemical shifts.

NMR is considered to be applicable for high-throughput protein structure determination of proteins up to 25 kDa, as long as stable, soluble isotope-labelled samples can be prepared. Because of the peak overlap that happens when different atoms have very similar chemical shifts, as well as signal degeneracy, it is challenging to solve structures of proteins larger than 40 kDa.

Since spin relaxation rates of nuclei are sensitive to protein motions, changes in protein structure can also be captured in the NMR experiment. Hence, this allows for observing the protein dynamics on a wide range of time scale.

Cryo-TEM and Cryo-ET

Cryo Transmission Electron Cryomicroscopy (cryo-TEM) and cryo Electron Cryotomography (cryo-ET) are related emerging techniques for protein structure solving. Some of the restrictions on samples of X-ray crystallography and NMR techniques do not apply to these methods: the crystalline structure is not necessary, there is no upper size limit, and small quantities are sufficient. The technique itself is very similar to standard light microscopy, but here electrons are used which are of lower wavelength than light, and thus enable for atomic resolution. Briefly, the beam of electrons is directed to the specimen, which has been frozen to protect it from evaporation and radiation damage, scattering some of the passing electrons which are then collected by a lens system and focused to form first a diffraction pattern and then a photographic or CCD image. To acquire a three-dimensional structure for a homogeneous specimen, averaging of thousands of cryo-TEM images of different particles at different orientations is performed. If a single structure is available, cryo-ET can be employed, where multiple images are taken of the specimen as it is tilted over a wide range of angles, from which a structure can be reconstructed. Due to the previously mentioned advantages of cryo-TEM and cryo-ET, and because of their limited resolution, the greatest advantage of these techniques was observed when used in combination with X-ray crystallography and NMR. However, the resolution of structures resolved in these methods keeps improving, with a 1.8 Å glutamate dehydrogenase, a hexamer, reported in recent studies (Merk *et al.* 2016). However, with these methods the higher resolution of the structures, like in the case of the mentioned hexamer, is typically achieved by using the internal symmetry of the underlying structure.

3.1.2 Enzymatic Activity Measuring

Most of the assays used to measure the velocity of the enzymatic reactions are performed by one or more of the following methods: spectroscopy, calorimetry, polarography, radioactive decay, electrophoretic separation, chromatographic separation, or immunological reactivity. These assays directly measure the substrate or product concentration in the reaction as a function of time and thus can be referred to as direct assays. If a distinct signal from a reaction is not available, other, nonenzymatic reactions are coupled to the product generation (indirect assays), or different enzymatic assays are combined to create a coupled assay.

Assays Based on Optical Spectroscopy

Two of the most common means to measure enzymatic reaction are absorption spectroscopy and fluorescence spectroscopy. Changes in the electronic configuration of the molecules as a result of them absorbing light energy of specific wavelengths is the underlying principle of these methods. When irradiating a molecule with varying wavelengths of light, the wavelengths whose energy matches the energy gap between two electronic states of the molecule will be strongly absorbed, inducing a transition between those two states. Beer's law (Beer 1852) relates this absorption A to the concentration of the sample in molar units c, the length of the sample the light traverses l, and the so-called molar absorptivity of the molecule ϵ. Thus, if ϵ is known, A can be determined from the absorption spectroscopy experiment. When such an absorption causes the transition between two electronic states of the molecule, the excited state is short-lived and the molecule returns to the ground state after emitting the excess energy through heat dissipation or, in some cases, emitting a photon. This fluorescence maximum is of longer wavelength (lower energy) than the absorption maximum, which is referred to as Stokes shift. If the substrate–product pair is not naturally fluorescent, it is often possible to covalently attach a fluorescent group without affecting the pair's interaction. After measuring at the same time fluorescence of both the target sample and a standard solution (to avoid variance due to lamp intensity), comparing the two fluorescence curves provides sensitive means to evaluate the concentration of the molecule.

Isothermal Titration Calorimetry

ITC is performed by a homonym device, whose main component is an adiabatic jacket containing two cells connected to the outside by narrow tubes. One of these cells, the sample cell, contains the solution, while the other, the reference cell, contains buffer or water. A thermoelectric device measures the temperature of both cells and using a cell feedback heater maintains them at equal temperature. By injecting the substrate into the sample cell, the amount of heat released or absorbed is measured. By (i) injecting the complete substrate to measure the total molar enthalpy ΔH and (ii) doing multiple substrate injections at different concentrations the heat flow dQ/dt is measured in two separate experiments, which allow to derive K_m and K_{cat}.

The same procedure can be used to measure the binding affinities between different molecules. The advantage of ITC is that by providing both K_d and ΔH measurements, it enables the estimation of all the thermodynamic parameters of the Gibbs free energy from Equation 2.11.

3.1.3 Phenotypic Viral Drug Resistance Testing

Historically, the first widely used HIV drug resistance tests employed Peripheral Blood Mononuclear Cells (PBMCs) isolated from the patients. These are mixed with donor PBMCs for cultivation and, after drug exposure, the antigen (typically Capsid protein 2.2.3) is measured (Mayers *et al.* 1992; Japour *et al.* 1993). This approach had several disadvantages, being time consuming, difficult in terms of virus extraction and cultivation (Schutten 2006), and exhibiting result variability for different donors (Mayers *et al.* 1992; Japour *et al.* 1993). In the late 1990s, two commercial assays were developed, Antivirogram® and Phenosense®. Both of these recombinant virus assays use a molecular HIV clone with *pol* gene deficient sequences (Subsection 2.2.3), the missing sequences added from the patient-derived HIV. After transfecting the HIV permissive cells with this clone, the newly created viral vectors are cultured *in vitro* at different drug concentrations. The main difference between these methods is that Phenosense® assay uses vectors which, apart from RT and PR, are also deficient in the *env* gene, which is replaced by luciferase reporter gene. These vectors are transfected to cells which additionally contain an *env* gene-expressing plasmid, ensuring that proteins coded by this gene are incorporated into the viral particles formed in that cell. However, these particles can only perform one cycle of infecting new cells as they do not have *env* in their RNA (Subsection 2.2.4), reducing the turnaround time and reproducibility of this assay (Petropoulos *et al.* 2000). Another difference between the two assays is that the viral replication in Phenosense® is quantified by luciferase activity, while in Antivirogram® the cytopathic effect (cell killing) is measured.

3.1.4 Blot-Based Protein Detection

Western blot is a blot-based technique for separating and identifying proteins. In this technique gel electrophoresis is used to separate a mixture of proteins based on their molecular weights. These gels are then transferred to a membrane, which produces visible bands for each protein. In a typical two-step procedure, first primary antibodies specific to the protein of interest are incubated together with the membrane. After washing the membrane, only the antibody bound to the protein of interest remains. A secondary antibody is then added which binds to the primary antibody, and it is this antibody that is used for signal detection. There are different options for performing signal detection. One of them is colorimetric detection, where a reporter enzyme bound to the secondary antibody stains the membrane by converting a soluble dye to one in a solid state. In the chemiluminescence detection approach, a substrate is used which luminesces when in contact with the reporter on the secondary antibody, which is captured by CCD cameras. In another approach called fluorescence detection, the probe is excited by light and its emission of excitation can be detected by CCD cameras.

In a simplification of the Western blot, a so-called dot blot procedure, the sample is applied directly on the membrane instead of using electrophoresis. This saves time by eliminating the need to use blotting on gel procedure, but because of this removal it provides no information about the protein size.

Descriptions of the theoretical foundations in Sections 3.2.2 and 3.2.3 follow in argument and in notation (Gapsys et al. 2015), Section 3.2.4 follows (Lindahl 2015), and Section Partial Charges follows (Woods and Chappelle 2000; Schlick 2002).

3.2 Computational Methods

3.2.1 Protein Mutation Modelling

Point mutation modelling in proteins is a problem related to protein homology modelling (Khan *et al.* 2016). Here, instead of aiming to computationally predict the whole structure of an unknown protein based on a homologous protein with a similar primary sequence and a known three-dimensional structure, the task is to predict the changes in the known protein structure upon replacement of a single amino acid. For this purpose, rotamer libraries based on analysis of known protein structures are typically used. These libraries can be split into protein backbone-independent (Ponder and Richards 1987; Tuffery *et al.* 1991; Maeyer *et al.* 1997; Lovell *et al.* 2000) and backbone-dependent (Dunbrack and Karplus 1993; Bower *et al.* 1997). Most of the available tools for introducing point mutations use the latter approach (Chinea *et al.* 1995; Schwede *et al.* 2003; R. E. Smith *et al.* 2007; Feyfant *et al.* 2007), where, given backbone ψ and ϕ angles positions, one aims to fit proper side-chain rotamers. In effect, the methods boil down to the tasks of (i) conformational search of placing the side-chains and (ii) scoring the created models. Historically, newer methods give more emphasis to the latter (Liang and Grishin 2002; Feyfant *et al.* 2007). Overall, different side-chain prediction methods have been suggested to yield similar results (Xiang and Honig 2001). Furthermore, if the models created are used to perform molecular dynamics simulations, different conformations of the protein structure are sampled in the simulation as governed by the underlying force field.

3.2.2 Free Energy

Free energy in this dissertation refers to the thermodynamic free energy, or the amount of work that a system can perform. Based on the second law of thermodynamics, a closed system will minimise its free energy until it reaches equilibrium at the free energy minimum. The thermodynamic free energy can be described as the Helmholtz free energy or Gibbs free energy. While both are applicable at constant temperature conditions, the difference between the two is that the former describes processes under constant volume (isochoric) conditions and the latter at constant pressure (isobaric) conditions. Since in most biochemical processes no changes of pressure occur, processes which minimize the Gibbs free energy are the main driving force.

In statistical mechanics, the Boltzmann distribution describes the probability p to observe a specific state x depending on its free energy:

$$p(x) \propto e^{-\frac{G(x)}{k_\beta T}}, \tag{3.1}$$

where k_β is Boltzmann's constant and T is the absolute temperature.

Equation 3.1 can thus be used to compare probabilities of two states A and B:

$$\frac{p(A)}{p(B)} = e^{-\frac{G(A)-G(B)}{k_\beta T}} = e^{\frac{\Delta G}{k_\beta T}}, \tag{3.2}$$

where ΔG is referred to as the free energy difference.

Intuitively, $p(A)$ here refers to the probability to find the system in a phase space volume A:

$$p(A) = \frac{e^{-\beta G_A}}{e^{-\beta G}}, \tag{3.3}$$

where G_A is the Gibbs free energy of phase space in volume A, G is the Gibbs free energy of the whole phase space of the system, and $\beta = \dfrac{1}{k_\beta T}$.

The free energy difference between two states G_{AB} can also be expressed in the following terms:

$$G_{AB} = G_B - G_A = -\frac{1}{\beta}\ln\frac{p_B}{p_A} = -\frac{1}{\beta}\ln\frac{Q_B}{Q_A}, \tag{3.4}$$

where $Q = Q(N, P, T)$ is referred to as the canonical partition function of the phase space for N number of particles in pressure P.

Replacing G with F in Equations 3.3 and 3.4 results in the Helmholtz free energy formalism, with a difference in the definition of $Q = Q(N, V, T)$ with volume V of the container. The partition function relates to the Helmholtz free energy via $F = -\dfrac{1}{\beta}\ln Q(N, V, T)$ and it is defined as follows:

$$Q(N, V, T) = \frac{1}{h^{3N}N!}\int\cdots\int e^{-\beta H(p_1\cdots p_N, q_1\cdots q_N)} d^3p_1\cdots d^3p_N\, d^3q_1\cdots d^3q_N, \tag{3.5}$$

where $H(p, q)$ is a Hamiltonian, which describes the total energy of the system in terms of coordinates p and momenta q, and h is the Planck's constant.

The partition function can be rewritten for constant pressure P as follows:

$$Q(N, P, T) = \frac{1}{h^{3N}N!}\int\cdots\int e^{-\beta H(p_1\cdots p_N, q_1\cdots q_N)+PV} d^3p_1\cdots d^3p_N\, d^3q_1\cdots d^3q_N. \tag{3.6}$$

Gibbs free energy relates to this partition function as $G = -\dfrac{1}{\beta}\ln Q(N, P, T)$.

3.2.3 Free Energy Estimation

Free Energy Perturbation

The formalism for the Free Energy Perturbation (FEP) method for the free energy difference estimation introduced by Zwanzig (Zwanzig 1954) can be derived from Equation 3.4:

$$\Delta G_{AB} = G_B - G_A = -\frac{1}{\beta}\ln\frac{Q_B}{Q_A} = -\frac{1}{\beta}\ln\langle e^{-\beta(H_B(p,q)-H_A(p,q))}\rangle_A, \tag{3.7}$$

where $\langle\cdots\rangle$ denotes the average of an ensemble. This value can be estimated by equilibrium sampling the system at state A and evaluating the difference of the Hamiltonian of states B and A of the resulting configurations.

Another way of estimating the free energy difference by sampling the end state of a system at equilibrium was introduced by Bennet (Bennett 1976) and is termed Bennet Acceptance Ratio (BAR):

$$\Delta G_{AB} = \frac{1}{\beta}\ln\frac{\langle f(H_A(p,q) - H_B(p,q) + C)\rangle_B}{\langle f(H_B(p,q) - H_A(p,q) - C)\rangle_A} + C, \tag{3.8}$$

where $f(x) = \dfrac{1}{1+e^{\beta x}}$ is the Fermi function and $C = \dfrac{1}{\beta}\ln\left(\dfrac{Q_A}{Q_B}\dfrac{n_A}{n_B}\right)$, with n_A and n_B denoting the number of configurations generated in states A and B. Equation 3.8 can be solved numerically by finding C, such that

$$\sum_B f(H_A(p,q) - H_B(p,q) + C) = \sum_A f(H_B(p,q) - H_A(p,q) - C). \tag{3.9}$$

The resulting free energy difference is then:

$$\Delta G_{AB} = -\frac{1}{\beta}\ln\frac{n_B}{n_A} + C. \tag{3.10}$$

So far Zwanzig's FEP approach has been defined in terms of sampling the end state of a system and BAR in both end states. The accuracy of these methods depends on the phase space overlap between states A and B, since otherwise the system sampled at A (or B) will have high energy state in terms of the Hamiltonian $H_B(p,q)$ (or $H_A(p,q)$), meaning that these configurations will contribute little to the exponential average in Equations 3.7 and 3.8. A coupling parameter λ overcomes this by allowing unphysical intermediate states (also called alchemical) $H_\lambda = (1-\lambda)H_A - \lambda H_B$. By using stratification with $\lambda \in [0..1]$ an alchemical pathway can be created between A and B by ensuring there is phase space overlap between the neighbouring states. Summing up the Zwanzig's FEP or BAR estimates of the free energy difference between neighbouring states results in the estimate of ΔG_{AB}.

Thermodynamic Integration

Another method for free energy change estimation dependent on a coupling parameter λ is called Thermodynamics Integration (TI) (Kirkwood 1935). It is conceptually different from the approaches above in that the ΔG_{AB} is obtained by integrating the average force exerted on the system along λ during an alchemical transition:

$$\Delta G_{AB} = \int_0^1 \langle\frac{\delta H}{\delta\lambda}\rangle_\lambda d\lambda. \tag{3.11}$$

Several different implementations based on TI can be used in simulations for estimating the free energy change. In the case of *slow growth* TI, a transition along λ is performed very slowly to keep the system close to equilibrium at all times, so that the work done along the pathway corresponds to ΔG. In the discrete thermodynamic integration, the path along λ is divided into discrete steps and an equilibrium simulation is performed for evaluating $\langle\frac{\delta H}{\delta\lambda}\rangle_{\lambda_i}$ at each λ_i. The numerical integration of the averages then corresponds to the ΔG_{AB} value. Similarly to FEP and BAR, a large overlap between neighbouring ensembles along the λ coordinate is required for an accurate ΔG estimate.

Non-Equilibrium Methods

The methods for free energy estimation discussed so far depend on the system being at equilibrium with its surroundings at all times. If it is not, friction will result in non-equilibrium work done, which on average will tend to increase the ΔG estimate. In 1997 Jarzynski (Jarzynski 1997) derived an equation which enabled the calculation of ΔG from work W, which is done during a non-equilibrium transition of a system:

$$e^{-\beta\Delta G_{AB}} = \langle e^{-\beta W}\rangle. \tag{3.12}$$

The non-equilibrium transitions are still required to be started from an equilibrium ensemble. The work values can be obtained from an equation similar to Equation 3.11:

$$W = \int_0^1 \frac{\delta H}{\delta\lambda} d\lambda. \tag{3.13}$$

Similarly to FEP, the convergence of this method depends on the occurrence of rare events, during which little work is dissipated. A relation which combines the work value distributions from both forward f and backward r transitions to obtain the Helmholtz free energy has been derived by Crooks (Crooks 1998; Crooks 1999) (it was later shown to apply to NPT ensemble as well (Chelli *et al.* 2007)), and it is called Crooks Fluctuation Theorem (CFT):

$$\frac{P_f(W)}{P_r(-W)} = e^{\beta(W-\Delta G)}. \tag{3.14}$$

The previously introduced BAR estimator can also be used for non-equilibrium simulations (Shirts *et al.* 2003):

$$\sum_{i=1}^{n_f} \frac{1}{1+e^{\ln\frac{n_f}{n_r}+\beta(W_i-\Delta G)}} = \sum_{j=1}^{n_r} \frac{1}{1+e^{\ln\frac{n_f}{n_r}-\beta(W_i-\Delta G)}}, \tag{3.15}$$

giving the maximum likelihood estimator of the free energy difference.

3.2.4 Molecular Dynamics Simulations

Two computational techniques that are prevalent to produce statistical mechanics ensembles are Molecular Dynamics (MD) and Monte Carlo (MC) simulations. Both approaches rely on the systems being ergodic, meaning that the time average (for MD) or the ensemble average (for MC) of a single molecule in a long simulation should correspond to the instantaneous ensemble average over all molecules in an experimental measurement of the same system. While there is disagreement on which technique is more efficient in producing this ensemble (Jorgensen and Tirado-Rives 1996; Yamashita *et al.* 2001), the advantage of MD over MC is that it gives dynamic information about the system.

A set of parameters that underlay an MD simulation of a system of atomic particles is referred to as Molecular Mechanics (MM) *force fields*. They are derived empirically and/or from Quantum Mechanics (QM) (Guvench and MacKerell 2008). In the force fields, typically a distinction is made between bonded interactions and non-bonded interactions. The former covers stretching of covalent bonds, angle-bending, as well as torsion potentials when rotating around the bonds. The latter is described by the attractive-repulsive Lennard-Jones potential and Coulomb electrostatic potentials. The simulation itself proceeds in an iterative manner, where given the initial coordinates and forces of all atoms in the system, the coordinates are updated for the next step.

To remove possible clashes between atoms before the actual MD simulation, typically a steepest descent algorithm is used to move atoms in the direction of decreasing energy. The MD simulation that follows performs the integration of Newton's equations of motion:

$$F_i = \frac{\partial V(r_i, \cdots, r_N)}{\partial r_i} \qquad \text{and} \qquad m_i \frac{\partial^2 r_i}{\partial t^2} = F.$$

MD simulations are carried out in a box of fixed dimensions, and to avoid artefacts at the boundaries of this box so-called periodic boundary conditions are used. Essentially this means an infinite grid of copies of the simulation box, where a molecule exiting the box in simulation in one dimension enters it from the opposite side. For non-bonded interactions this means that the infinite number of interactions should be summed. In practice a cutoff is used for the Lennard-Jones potential, while the Particle-Mesh-Ewald (PME) is used to calculate the infinite electrostatic interactions by dividing the summation into short- and long-range parts (Essmann *et al.* 1995).

When simulating an NPT ensemble, the temperature and pressure have to be controlled. During the simulation, as the potential energy decreases, the kinetic energy, which is related to the temperature, increases. To control this, the system is typically coupled to a thermostat, whose function is to scale the velocities during the integration to maintain the temperature. Similarly, the total pressure of the system is controlled by a barostat which scales the simulation box size.

To access long time-scales of simulation, increasing the time step is desirable. This is however limited by the highest frequency motions in the simulated system, amounting to the fact that errors introduced from the bonds' vibrations manifest themselves already at 1 fs. Bond constraint algorithms such as SHAKE (Ryckaert *et al.* 1977) and LINCS (Hess *et al.* 1997) are thus used to remove these oscillations, allowing the use of 2 fs time steps in simulation.

The Born-Oppenheimer approximation (Born and Oppenheimer 1927), which separates the motion of atomic nuclei and electrons, has an important role in MD. Its manifestation here is that in MD it is nuclei that are the point particles that follow the Newtonian dynamics. Thus, the classical MD is suitable for a wide range of problems, including modelling the conformational changes of organic and inorganic molecules and their non-covalent interactions. For studying important electronic changes such as the bond formation, a QM or hybrid (QM/MM) treatment is needed.

Partial Charges

An important set of parameters in the force fields that allows for modelling of electrostatic energies is that of the partial charges, which, despite being artificial, is particularly relevant for biomolecules. Partial charges are assigned to the atoms by optimising the fit (χ^2_{esp}) between the classical Coulomb model for the electrostatic potential ($\hat{V}$) and the quantum mechanical molecular electrostatic potential (V) at points i around the molecule (called ESP-charges):

$$\chi^2_{esp} = \sum_i (V_i - \hat{V}_i)^2, \tag{3.16}$$

where $\hat{V} = \sum_j \frac{q_j}{r_{ij}}$. The bond polarity overestimation has been suggested to be an issue for these charges, for which reason attenuating with χ_{restr} was recommended in the so-called Restrained Electrostatic Potential (RESP)-charges:

$$\chi^2_{resp} = \chi^2_{esp} + \chi^2_{rstr}, \tag{3.17}$$

where $\chi^2_{rstr} = k_{rstr} \sum_j (\sqrt{q_j^2 + b^2} - b)$. b determines the tightness of the hyperbola around its minimum and k_{rstr} determines the strength of the restraint function.

Ab initio quantum mechanics is one way to acquire V, which implies a non-empirical solution to the Schrödinger equation. In these methods, the molecular orbitals are approximated by a linear combination of atomic orbitals, which are defined for specified basis sets, such as Gaussian functions. The coefficients that describe this linear combination are calculated by minimizing the electronic energy (known as Hartree-Fock energy) of the molecular system for a given set of chosen orbitals.

3.2.5 Prediction of Acid Disassociation Constant

Proteins' ionisable groups have an important function for the intra-protein, protein–solvent, and protein–ligand interactions, which in turn play an important role in protein folding, binding of interaction partners, and catalytic activity. Together with the advancing experimental procedures for pK_a determination, a large body of computational efforts has been invested in developing tools for calculating pK_a (see (Alexov *et al.* 2011) for review). In effect, the computational methods try to predict the change in the model residue's pK_a values in solvent when this residue is moved into protein, or the ΔpK_a. One of the major directions of research for this purpose has been to numerically solve the linearised Poisson–Boltzmann (PB) equation, which describes the electric potential in solution normal to the charged surface (Bashford and Karplus 1990). PB models, which sample protonations at a fixed conformation, have been combined with MM/MD methods, which sample conformations at a fixed protonation (Machuqueiro and Baptista 2006). A much less computationally demanding alternative is implemented in empirical methods such as Propka (H. Li *et al.* 2005), which use simple empirically estimated terms to evaluate the influence of hydrogen bonds, desolvation effects, and charge–charge interactions on the residue's ΔpK_a.

4 Drug Resistance Mutations of HIV-1 Protease

This chapter describes work performed on the drug resistance mutations of the HIV protease. Part of the study which concerns the resistance mutations G48V, I50V, and L90M was done in collaboration with Vytautas Gapsys, Nadezhda T. Doncheva, Rolf Kaiser, Bert L. de Groot, and Olga V. Kalinina and published in [Bastys et al. 2018]. I have written the manuscript for the publication and the corresponding text and figures in this chapter have been adapted from that publication[a]*. The second half of the study has been done in collaboration with the same people and additionally with Hauke Walter providing the experimental resistance factor measurements for the L76V mutation. The manuscript for that study is currently in preparation. I performed all the computational work described in this chapter, except for the implementation of the method for calculating the resistance factor values which was done by Vytautas Gapsys.*

[a] Reproduced in part with permission from T. Bastys *et al.* [2018]. "Consistent Prediction of Mutation Effect on Drug Binding in HIV-1 Protease Using Alchemical Calculations." *Journal of Chemical Theory and Computation* 14.7, pp. 3397–3408. DOI: 10.1021/acs.jctc.7b01109. Copyright 2018 American Chemical Society.

As outlined in Section 2.2, the HIV epidemic, which started over twenty years ago, remains a global health hazard. Despite the many different ART options, drug resistance is still an issue. Current treatment protocols include PIs, nine of which are currently marketable as FDA-approved treatments against the HIV protease (*Antiretroviral Drugs Used in the Treatment of HIV Infection* n.d.). Nevertheless Resistance-Associated Mutations (RAMs) can arise against all of them (*The Stanford HIV Drug Resistance Database* n.d.).

In this chapter a study of the molecular mechanisms of the HIV protease RAMs is presented. The focus of this study is the use of MD simulations to estimate the effect of the mutation on the free binding energy of the inhibitors and to relate these resulting observations to the experimental measurements of the same values, as well as to the Resistance Factor (RF) measurements. The underlying energetic and mechanistic reasons for the mutations' effect on inhibitor binding are also analysed to gain understanding of the underlying molecular reasons for the drug resistance and sensitivity phenomena.

4.1 Introduction

Acquiring resistance against antiviral drugs by HIV-positive patients can have a number of underlying causes, including treatment adherence failure, sub-optimal treatment regimen prescribed, disease stage at the time of treatment initiation, pharmacogenetics, drug–drug interactions, failure to monitor viral load, or interruption of treatment due to medication availability. Some of these and other causes are exacerbated in the countries with resource-limited settings. Treatments which combine multiple drugs, including the ones targeting different HIV proteins, raise the barrier for drug resistance. Some specific mutations, which exhibit resistance-inducing, as well as sensitizing effect towards different drugs, give an opportunity to efficiently combine ARV drugs (Larder *et al.* 1995; Ziermann *et al.* 2000; Wiesmann *et al.* 2011). Drug tolerability is one of the limiting factors in prescribing drug combinations; even treatments combining three different drugs are expected to cause drug resistance after several years (The UK Collaborative Group on HIV Drug Resistance 2005).

Resistance-associated mutations typically alter the affinity of inhibitor molecules to their protein targets. The change in the free energy of binding of inhibitors to the target protein, e.g. HIV protease, upon mutations, $\Delta\Delta G$, measures the difference of binding free energy between mutant protein and wildtype protein to inhibitor, ΔG_{MUT} and ΔG_{WT} respectively. This change in free energy of binding is directly related to the change in the inhibitor's affinity to the mutant protease (and thus to the resistance phenotypes), and hence accurately estimating the effect of a mutation on $\Delta\Delta G$ is highly desirable to predict its relation to drug resistance. MD simulations of protein and inhibitor allow to use different computational techniques for this purpose. Free Energy Perturbation (FEP) was employed in the 1990s during the development phase of the first PIs for predicting ΔG for various, mostly experimental inhibitors (Ferguson *et al.* 1991; M. R. Reddy *et al.* 1991; Rao *et al.* 1992; Tropsha and Hermans 1992; Rao and Murcko 1994; X. Chen and Tropsha 1995; Rao *et al.* 1996; Rao and Murcko 1996; M. R. Reddy and Erion 1998; McCarrick and Kollman 1999), but also in some more recent studies (E. C. B. Johnson *et al.* 2007; Y. Yu *et al.* 2015; S. T. Ngo *et al.* 2015). Thermodynamics Integration (TI) has also been applied for the same purpose (Cai and Schiffer 2010; Deng *et al.* 2015).

A conceptuatly different set of techniques developed in the late 1990s for estimating free energy of binding of small ligands and macromolecules are the so-called Molecular Mechanics/Poisson–Boltzmann Surface Area (MM/PBSA) and Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) methods (Srinivasan *et al.* 1998; Kollman *et al.* 2000). The names stem from their using molecular mechanics energy terms describing bonded, electrostatic, and van der Waals interactions and combining them with polar and non-polar solvation free energy terms, estimated by solving Poisson-Boltzmann equation or by using generalized Born model and solvent accessible surface area accordingly. Another term included in the analysis performed by these methods is entropy, estimated from normal-mode analysis of the vibrational frequencies, multiplied by temperature term. Unlike the alchemical methods, in these methods sampling is performed only in the end states (the complex and possibly free ligand and receptor), thus MM/PB(GB)SA are also referred to as end-point methods. Widely used due to their computational efficiency, MM/PB(GB)SA have also been applied extensively for studying the HIV protease–ligand complexes (Kalra *et al.* 2001; Hou and R. Yu 2007; Kožíšek *et al.* 2007; Altman *et al.* 2008; Hou *et al.* 2008; Hou *et al.* 2009; Stoica *et al.* 2008; Wittayanarakul *et al.* 2008; Alcaro *et al.* 2009; J. Chen *et al.* 2009; Hu *et al.* 2010; Shi *et al.* 2009; Cai and Schiffer 2010; J. Chen *et al.* 2010; Sadiq *et al.* 2010; Kar and

Knecht 2012a; Kar and Knecht 2012b; D. Li *et al.* 2012; Meher and Y. Wang 2012; Srivastava and Sastry 2012; Tzoupis *et al.* 2012; D. Li *et al.* 2014; Wright *et al.* 2014; R. Duan *et al.* 2015). However, studies using these methods which include a comparison to the experimental ligand binding affinities (Hou and R. Yu 2007; Kožíšek *et al.* 2007; Hou *et al.* 2009; Stoica *et al.* 2008; Wittayanarakul *et al.* 2008; Hu *et al.* 2010; Cai and Schiffer 2010; J. Chen *et al.* 2010; Sadiq *et al.* 2010; Kar and Knecht 2012a; Kar and Knecht 2012b; Meher and Y. Wang 2012; Tzoupis *et al.* 2012; Wright *et al.* 2014; R. Duan *et al.* 2015) are limited in their scope, i.e. in the number of mutations/inhibitors analysed, or often have poor correlation to the experimental estimates. It has been shown that the performance of these methods in terms of reproducing the ΔG of a ligand binding to a protein varies depending on the system (Kuhn *et al.* 2005; D. A. Pearlman 2005; Hou *et al.* 2011; Genheden and Ryde 2015).

In an experimental setting, the resistance of mutant proteins towards inhibitors, such as in the studies mentioned above, is typically measured in terms of the Inhibitor Concentration required to reduce enzymatic reaction activity by 50% (IC_{50}) or the Effective Concentration corresponding to the half-maximal response (EC_{50}). Thus, in the HIV research the ratio between IC_{50} or EC_{50} in mutant and the same measurement for the wildtype protease (typically with the consensus sequence from the HXB2), also called RF, is a useful descriptor for resistance of different mutated proteins. RF is directly related to the free energy of inhibitor binding, ΔG, and the protein enzymatic activity, K_m (Cheng and Prusoff 1973).

Studies of inhibitor affinity to the protease based on MD simulations also give a unique information into their structures, interactions, and changes thereof in time. Comparing wildtype and mutant complexes can therefore provide information on the underlying physical reasons of the effect that the mutation has on inhibitor binding. This knowledge can help in designing inhibitors with minimal vulnerability towards the resistance mutations of HIV protease, as well as potentially other viruses proteases, which are typically present in human-infecting viruses (Kräusslich and Wimmer 1988; Tong 2002) (for review on strategies on improving HIV and Hepathitis C Virus (HCV) protease inhibitors against resistance see (Yilmaz *et al.* 2016)). While some studies analyse the mechanistic effect of the select major RAMs on binding of different inhibitors (Hou and R. Yu 2007; Muzammil *et al.* 2007; Alcaro *et al.* 2009; Shen *et al.* 2010; Mittal *et al.* 2013; R. Duan *et al.* 2015; Y. Yu *et al.* 2015), and other analyse the effect of different RAMs on binding of the same inhibitor (Liu *et al.* 2005; Kovalevsky *et al.* 2006; Kožíšek *et al.* 2007; Hu *et al.* 2010; Cai and Schiffer 2010; J. Chen *et al.* 2010; Kar and Knecht 2012a; Meher and Y. Wang 2012; Ragland *et al.* 2014; Ragland *et al.* 2017), most of the studies are focused on single mutation and inhibitor combinations, particularly for major RAMs outside of the binding pocket, offering only a limited perspective on molecular mechanisms of protease resistance. The mechanism of action of a mutation on binding of different inhibitors is particularly interesting for those cases, where the same mutation is known to cause resistance to some drugs, while making the protein sensitive towards others. Such mutations are for example L76V, which is associated with resistance towards APV, Indinavir (IDV), Darunavir (DRV), and Lopinavir (LPV), but decreased resistance towards Atazanavir (ATV) and Saquinavir (SQV) (T. P. Young *et al.* 2010; Wiesmann *et al.* 2011), or N88S, which is a RAM towards IDV and Nelfinavir (NFV), but increases susceptibility towards APV (Ziermann *et al.* 2000; Resch *et al.* 2002; Vermeiren *et al.* 2007) or its prodrug FPV (Rhee *et al.* 2010).

In this chapter, a diverse set of HIV protease RAMs, both inside and outside the binding pocket, is analysed. These RAMs are divided into three parts: dataset 1 for different mutations with experimental measurements of $\Delta\Delta G$, dataset 2 for different mutations with experimental measurements of RF, and dataset 3 for L76V mutation in different sequence backgrounds with experimental measurements of RF. We demonstrate using dataset 1 that alchemical free energy calculations using a non-equilibrium approach based on MD simulations allow for an accurate estimation of the impact of mutations on inhibitor binding for a set of different RAMs in combination with different inhibitors. We show an overall good correlation of the estimated and empirical $\Delta\Delta G$ values and emphasize the importance of the correct choice of the protonation state for the two aspartic acid residues in the active site. We also demonstrate using datasets 2 and 3 that we can in general faithfully reproduce the effects of RAMs on IC_{50}. It is then analysed how the effects of the mutation propagate in the protease structure to affect the inhibitor binding at a site that can be remote from the mutated amino acid residue. We show coupled effects among different protease residues affecting the binding of the inhibitors. Thus, novel and general insights into the development of HIV resistance to drugs are provided.

4.2 Methods

4.2.1 Drug Resistance Measuring for Protease Acquired from Patient Samples

The experimental data in dataset 3 was derived from samples of patients who underwent multiple therapy failures with different PIs. These viral variants displayed resistance and re-sensitising effects of the PIs LPV and SQV. Those variants were observed in the diagnostic procedure, sequenced, and subsequently tested in a phenotypic assay. The tests were carried out after the patient's variant was cloned into a recombinant derivate of the HIV NL4-3, called pNL4-3-Delta-PRT5. These variants were analysed in cell culture experiments where they were exposed to different PIs in different concentrations to estimate their RF values (Table 4.1). Based on these variants, the clones were specifically modified by site-directed mutagenesis so that different variants of L76V could be tested in different genetic backgrounds. For simplicity, from here on regardless of the residue at position 76 of protease as present in the original clinical samples, L76 will be referred to as wildtype residue and V76 as the mutant residue as per HXB2.

4.2.2 System Preparation

Crystal structures of protease–inhibitor complexes were obtained from the Protein Data Bank (Berman *et al.* 2000) (PDB IDs 1HPV (APV), 1HXB (SQV), 1K6C (IDV), 1MUI (LPV), 1SDT (IDV), 2BPX (IDV), 2O4K (ATV), 2O4P (TPV), 2O4S (LPV), 3EKV (APV), 3EL1 (ATV), 3NU3 (APV) and 3PWR (SQV)). Modeller (Sali and Blundell 1993) version 9.12 was used to introduce mutations targeted in this study as well as the background mutations. Mutation I50V was introduced in structures 1SDT, 2O4K, 2O4P, 2O4S, and 3NU3, as well as mutations G48V, L90M, and G48V/L90M in structure 1HXB. For other RAMs analysed in this study, additional background mutations were required, described in the following, where first the background mutations are listed followed by the structure name and the RAM analysed (plus the genotype identifier in cases where L76V was studied) in the parenthesis:

- Q7K in 1HPV (I50V)
- Q7K in 2BPX (I50V)

Genotype		**ATV**	**SQV**	**IDV**	**LPV**
FB15	**L76**	63	74	–	–
	V76	0.9, 2.3, 4.5	3.6, 4.6, 5.8	–	–
GH9	**L76**	90	18.6	–	–
	V76	1.2, 2.4, 3.2, 3.6, 1.9	1.2, 1.3, 1.5	–	–
RU1	**V76**	2.7	–	–	157
	L76	9, 10, 8.4	–	–	27, 47, 46
iZ2	**V76**	4.1	–	59	71
	L76	12, 7.8, 30	–	7.9, 10, 2.2	11, 12, 5.5

Table 4.1: Protease RF values for **dataset 3**, with multiple measurements for the same protein separated by comma. For each genotype the first row represents the wildtype position as in the original sample and the second row represents the mutation introduced at position 76.

- K7Q, R14K, R57G, T82V, and V84I in 1K6C (N88S)
- K7Q, R14K, K41R, and V77I in 3EKV (N88S)
- L10F and S37N in 1MUI (I84V)
- L10F and S37N in 1HPV (I84V)
- S37N and A71V in 1HPV (M46I)
- L10F and S37N in 2BPX (I84V)
- S37N and A71V in 2BPX (M46I)
- K7Q, I13V, G16E, K20I, I33F, M36L, S37N, I62V, I63H, A67C, A71V, G73S, I84V, L90M, and 95C in 3PWR (V76L in the genotype FB15)
- K7Q, I13V, G16E, K20I, I33F, M36L, K41R, I62V, P63H, V64I, A71V, G73S, I84V, and L90M in 3EL1 (L76V in the genotype FB15)
- K7Q, L10V, I13V, G16E, K20R, I33L, E35D, M36I, S37N, M46I, I54V, Q58E, I62V, I63H, I64V, A67C, V82F, I84V, and A95C in 3PWR (V76L in the genotype GH9)
- K7Q, L10V, I13V, R14K, G16E, K20R, E35D, M36I, K41R, M46I, I54V, Q58E, I62V, P63H, V82F, I84V in 3EL1 (L76V in the genotype GH9)
- K7Q, L10V, I13V, R14K, K20M, E35D, M36I, M46I, I54V, Q58E, P63L, V64I, H69K, V82M, and L89I in 3EL1 (L76V in the genotype RU1)
- L10I, I13V, K20M, E35D, M36I, S37N, R41K, M46I, I54V, Q58E, H69K, V82M, and L89I in 1MUI (L76V in the genotype RU1)
- K7Q, L10I, I13V, R14K, L24I, L33F, K46L, I62V, A71V, T82A, V84I, and Q92K in 1K6C (L76V in the genotype iZ2)
- L10I, I13V, L24I, L33F, S37N, M46L, I62V, L63P, A71V, and V82A in 1MUI (L76V in the genotype iZ2)
- K7Q, L10I, I13V, R14K, L24I, L33F, K41R, M46L, I62V, V64I, A71V, V82A, and Q92K in 3EL1 (L76V in the genotype iZ2)

In two cases where the RAM I50V was analysed, we have investigated viral strains containing additional mutations L33I, L63I, C67A, and C95A, for which we used structures 1SDT and 3NU3. Denoted I50V* in this study, these mutations are introduced to protect the protease against autocatalysis as well as cysteine thiol oxidation; however this practice is not universally applied (Muzammil *et al.* 2007). All of the aforementioned mutations were introduced in both protease monomers in consistency with the data from the experimental studies of HIV protease and HIV infection *in vivo* (Josefsson *et al.* 2011; Josefsson *et al.* 2013).

To validate our approach of choosing the active site protonation state, the following structures resolved by X-ray/neutron crystallography were considered: 2ZYE (with a not used in ART inhibitor KNI-272), 5E5J (DRV), 5E5K (DRV), and 5T8H (APV). Hydrogen atoms were assigned to aspartic acid, glutamic acid, and histidine residues as they were resolved in the corresponding structures, with the exception of 5E5J, where D30 was assigned a deprotonated state as the deuterium atom in that structure had 50% occupancy on D30 and K45 residues (Gerlits *et al.* 2016), and 5E5K, where the E34 was assigned a protonated state as this state was inferred in the original study (Gerlits *et al.* 2016).

In the following, preparation for simulation of all the structures mentioned above is described in both holo and apo states. Exception to that are the wildtype and mutant apo structures for which multiple complexes had the same protein sequence (with mutation indicated in the parenthesis): 1HPV (I50V), 1SDT (I50V), 2O4P (I50V), 2O4S (I50V), 2BPX (I50V, M46I, and I84V), 1MUI (I84V), and 1HXB (I84V), whose simulation was not required (see Section 4.2.4). The Gromacs simulation software package was used to set up (versions 4.6.2 and 4.6.5), carry out, and analyse the MD simulations (versions 5.0.2 and 5.1.2) (Hess *et al.* 2008; Abraham *et al.* 2015). All crystallographic water and ion molecules were retained. The pK_a of residues was predicted using Propka (Søndergaard *et al.* 2011) and the protease was assigned a monoprotonated state on either D25/D25′, where the prime refers to the second subunit of the protein. Exception to that were structures 3NU3 and 1SDT, where previously suggested protonation was used (R. Duan *et al.* 2015), and 1HXB, where similar pK_a values (6.42 and 6.47) were predicted for both active site residues and protonation of D25 was chosen. The mutant reference protein protonation state was considered to be the same as that of the wildtype protein. The summary of protonated aspartic acids for all the complexes is reported in Tables B.1, B.4, and B.5. The Amber99SB*-ILDN force field was used for the protease parametrisation. Chemaxon Calculator (*Chemaxon Calculator 5.3.8* n.d.) was used to determine ligand protonation. Ligands ATV, LPV, APV were assigned neutral charge, IDV and SQV were assigned a +1 charge, and TPV was assigned a −2 charge. Gaussian09 (Frisch *et al.* 2010) was used to optimize ligand geometry and calculate the electrostatic potential at the HF/6-31G* level of theory. Partial charges were assigned by performing RESP fit (Bayly *et al.* 1993). Bonded parameters and atom types were obtained from the Generalized Amber Force Field (GAFF) (J. Wang *et al.* 2004). ACPYPE (Silva and Vranken 2012) was used for file conversion into the Gromacs format. The complex was solvated in TIP3P water molecules with 1.4 nm buffer in each dimension with 0.15 mol/L concentration of Cl and Na ions (Joung and Cheatham 2008) to neutralize the system.

4.2.3 Equilibrium MD Simulations

Each system was subjected to energy minimization using the steepest descent algorithm. A maximum of 5000 steps was performed until a maximum force of 1000.0 kJ mol^{-1} nm^{-1} was achieved. Ten replica 200 ns simulations for each complex were performed at 300 K with a time constant of 0.1 ps using the velocity-rescaling thermostat (Bussi *et al.* 2007), at a constant pressure of 1 atm with a time constant of 4 ps using the Parrinello–Rahman barostat (Parinello and Rahman 1981). The electrostatic interactions were calculated at every step with the PME method (Essmann *et al.* 1995), and the short-range repulsive and attractive dispersion interactions were simultaneously described by a Lennard-Jones potential with a cutoff of 1.2 nm. All bonds were constrained using the LINCS algorithm (Hess *et al.* 1997) and a time step of 2 fs was adopted. Simulated annealing in length of 1 ns was performed to avoid a too close contact between atoms before equilibrium simulation for 2BPX (50V variant) in both protonation states, 5E5J in D25′ protonation state, 1MUI (76V for genotype iZ2) in D25′ protonated state, 1MUI (76L for genotype iZ2) in D25 protonated state, 2BPX (46V and 84I variants) in D25 protonated state, 3EKV (88S variant) in D25 protonated state, and 3EL1 (76L and 76V for genotype iZ2) in D25′ protonated state. For all of the analyses that followed, the first 20 ns of the simulations were considered to be a part of the system equilibration process and thus discarded, with the exception of free energy calculations, where first 10 ns were discarded.

4.2.4 Free Energy Calculations

The protocol for free energy calculations was adjusted from the non-equilibrium simulation approach used in assessing changes in protein thermal stabilities and protein–protein interactions upon amino acid mutation (Gapsys *et al.* 2016). For calculating the free energy change upon mutation of apo structures, ΔG_1, estimates for the following structures were used in multiple $\Delta\Delta G$ estimates where the same mutation (indicated in parenthesis) in the same background sequence (as indicated in Section 4.2.2) was analysed: 2O4K (I50V), 3NU3 (I50V), and 1HPV (M46I and I84V). From each of the equilibrium simulations described in the previous section, trajectory frames were extracted equidistantly in time every 10 ns. For every snapshot the hybrid structures and topologies were generated using the pmx framework (Gapsys *et al.* 2014) for all of the residues to be mutated. Subsequently, short 20 ps simulations were performed to equilibrate the velocities. Finally, alchemical transitions were carried out in 50 ps. Alchemical transitions were extended to a total of 400 ps in complexes 1HPV (46I, 46M, 84I, and 84V variants) in D25 protonated state and 1MUI (84I and 84V variants) in D25 protonated state when estimating $\Delta\Delta G_{calc}$, as well as 1K6C (88N and 88S variants) in D25 protonated state and 3EKV (88N and 88S variants) in D25 protonated state when estimating $\Delta\Delta G_{WT}^{prot}$ and $\Delta\Delta G_{MUT}^{prot}$ used in the dataset 2, because of high variance of the $\Delta\Delta G$ estimates from those transitions. During these transitions mutations in both of the monomers were introduced simultaneously, resulting in four mutations for the double mutant G48V/L90M and two mutations for the rest of the complexes. The simulation parameters for the 20 ps equilibration and alchemical transitions were identical to those used in the 200 ns equilibrium simulations. During the transitions, non-bonded interactions were soft-cored (Gapsys *et al.* 2012). The CFT (Crooks 1999) was used to relate the obtained work distributions to the free energy values by employing the maximum likelihood estimator (Shirts *et al.* 2003). The error estimates were obtained by the bootstrap approach.

4.2.5 Combining Free Energy Estimates from Alternative Protonation States

To include the contributions of the alternative protonation states in the binding free energy change upon mutation, we need to add the free energies of those states for wildtype and mutant proteins. Assume that in our calculation of $\Delta\Delta G_{calc}$ in both cases D25 was protonated and D25′ was not and refer to it as state A. Then we want to estimate for both proteins $G(B)$, i.e. the free energy of protein where D25′ is protonated, which per Equations 3.4 and 3.7 corresponds to:

$$G(B) = -\frac{1}{\beta} \ln p(B). \tag{4.1}$$

Assuming there are only two possible protonation states for each protein, using $p(A) + p(B) = 1$ and Equation 3.2 we get:

$$p(B) = \frac{1}{1 + e^{-\beta \Delta G}}. \tag{4.2}$$

From Equations 4.1 and 4.2 we arrive at the final estimate of binding free energy change upon mutation, $\Delta\Delta G_{total}$:

$$\Delta\Delta G_{total} = \Delta\Delta G_{calc} + \frac{1}{\beta} \ln(1 + e^{-\beta \Delta G^{prot}_{WT}}) - \frac{1}{\beta} \ln(1 + e^{-\beta \Delta G^{prot}_{MUT}}). \tag{4.3}$$

4.2.6 Assessing the Resistance Factor Ratio from the Inhibitor Binding Free Energy Change

Cheng-Prusoff equation (Cheng and Prusoff 1973) relates inhibitor's IC_{50} and binding affinity K_i (Equation 2.9), which in turn can be estimated from the inhibitor binding free energy ΔG (Equation 2.10). Thus, given two RF values for two proteases with sequences A and B, RF_A and RF_B, their ratio can be related to $\Delta\Delta G$:

$$\mathrm{RF}_R = \frac{\mathrm{RF}_A}{\mathrm{RF}_B} = \frac{\frac{\mathrm{IC}^A_{50}}{\mathrm{IC}^{WT}_{50}}}{\frac{\mathrm{IC}^B_{50}}{\mathrm{IC}^{WT}_{50}}} = \frac{e^{\frac{\Delta G_A}{k_\beta T}}}{e^{\frac{\Delta G_B}{k_\beta T}}} \left(\frac{1 + \frac{[S]}{K^A_m}}{1 + \frac{[S]}{K^B_m}} \right) = e^{\frac{\Delta\Delta G}{k_\beta T}} \left(\frac{1 + \frac{[S]}{K^A_m}}{1 + \frac{[S]}{K^B_m}} \right) = e^{\frac{\Delta\Delta G}{k_\beta T}} C, \tag{4.4}$$

where IC^{WT}_{50} refers to the reference wildtype protein. We are interested in obtaining a distribution of the RF_R values after calculating the double free energy differences $\Delta\Delta G$:

$$p(\mathrm{RF}_R | \Delta\Delta G, C) \propto p(\Delta\Delta G, C | \mathrm{RF}_R) p(\mathrm{RF}_R). \tag{4.5}$$

When there are multiple RF_R measurements and $\Delta\Delta G$ calculations available, C can be expressed as a function of the available values $C = C_i(\Delta\Delta G_i, \mathrm{RF}^i_R)$, $i = 1, \ldots, n$. This gives the final posterior distribution:

$$p(\mathrm{RF}_R | \Delta\Delta G, \Delta\Delta G_i, \mathrm{RF}^i_R) \propto p(\Delta\Delta G, \Delta\Delta G_i, \mathrm{RF}^i_R | \mathrm{RF}_R) p(\mathrm{RF}_R), \tag{4.6}$$

where $C_i = \mathrm{RF}^i_R e^{\frac{-\Delta\Delta G_i}{k_\beta T}}$. The $\Delta\Delta G$ values are sampled from a Gaussian distribution with the mean and standard deviation corresponding to the calculated double free energy difference and estimated error, respectively.

4.2.7 Partial Least-Squares Regression

Partial-Least Squares (PLS) regression was performed using the functional mode analysis tool (Krivobokova *et al.* 2012). The following sets of input atoms (excluding hydrogen atoms) were used for the model: backbone, ligand, protein, and side chain. Constants 0 and 1 have been used as response variables for trajectories corresponding to mutant and wildtype protein simulations, respectively.

Cross validation for each mutation and inhibitor combination was done as follows: all trajectories for wildtype and mutant complexes were concatenated, superimposed to minimize the variance over the ensemble (Gapsys and Groot 2013), and divided into five equal parts. In each iteration, a model was trained on four parts of labelled input in equal parts from wildtype and mutant simulations, after which it was used to make predictions for the last part. The Pearson correlation between the actual signal and prediction was used to measure the prediction quality. The number of components in the final model has been selected using the so-called "elbow method", where the model complexity is increased by boosting the number of components until adding further components only marginally improves the quality of prediction (Tibshirani *et al.* 2001).

4.2.8 Mutual Information

Mutual Information (MI) I between pairs of backbone ϕ and ψ, as well as side chain χ dihedral angles of residues was estimated from their individual and joint entropies using the MutInf (McClendon *et al.* 2009) method. Twenty-four bins were used to get the discrete distributions of the dihedral angles. Dihedrals from 10 simulations for each case were kept separate for the later evaluation of I^{ind}, or the excess mutual information, for pairs of torsion angles from independent simulations. Monte Carlo sampling with adaptive partitioning of all torsion angle pairs was used to obtain a background distribution of I, and only those I entries which had a p-value of at most 0.01 according to this distribution were retained. False positives were removed based on $P(I < E[I_{ind}])$ criteria, namely that the true mutual information is lower than expected for independent torsion angles, I_{ind}. I^{ind} is then subtracted from the actual I so that incomplete sampling resulting from memory effects in the simulations can be corrected for. The mutual information for any two residues, I_{res}, is then estimated as the sum of the mutual information between pairs of those residues' torsion angles. For more details on the methods described we refer to the original paper (McClendon *et al.* 2009).

Bootstrap sets were created for each set, by randomly selecting with replacement n frames from a simulation of length n. This was done for each of the 10 simulation runs while preserving the same order across the different dihedral angles and simulation runs, repeating this procedure 10 times. For each resulting bootstrap set, I was estimated. The mean and standard deviation of I_{res} from the 10 bootstrap sets, $\mu_{I_{res}}$ and $sd_{I_{res}}$ respectively, were then calculated. When comparing the actual residue pairs, $I_{res_{WT}}$ and $I_{res_{MUT}}$, for wildtype and mutant simulations for some complex, only those cases were retained where

$$\left|\mu_{I_{res_{MUT}}} - \mu_{I_{res_{WT}}}\right| > \sqrt{\frac{sd^2_{I_{res_{MUT}}}}{10} + \frac{sd^2_{I_{res_{WT}}}}{10}} \tag{4.7}$$

from the bootstrapped sets. As a final filter, residues whose differences in Equation 4.7 were smaller than or equal to 0.3 kT were not considered for further analysis.

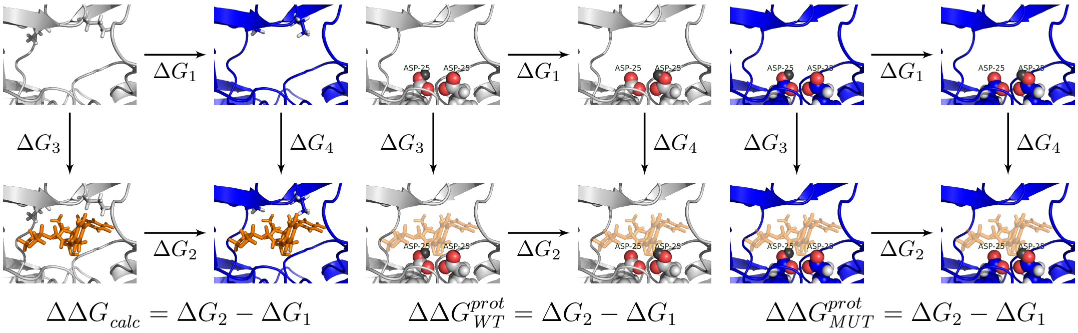

ΔG_1
ΔG_3
ΔG_4
ΔG_2
ASP-25
ASP-25
$\Delta\Delta G_{calc} = \Delta G_2 - \Delta G_1$
$\Delta\Delta G^{prot}_{WT} = \Delta G_2 - \Delta G_1$
$\Delta\Delta G^{prot}_{MUT} = \Delta G_2 - \Delta G_1$

Figure 4.1: Thermodynamic cycles for estimation of $\Delta\Delta G$ of protease-inhibitor complex upon mutation. The first cycle estimates the mutation effect on ligand binding, the second cycle estimates the effect of switching the proton between D25/D25′ for wildype protein, and the third cycle switches correspondingly for the mutant protein. Only the binding pocket is shown in the figures; the wildtype protein is shown in grey, the mutant is shown in blue, and the inhibitor is shown in orange. In the second and third cycles the inhibitor is set to semi-transparent for reasons of clarity, and the proton in question for D25/D25′ is shown as a black sphere.

4.3 Results

4.3.1 Estimation of the Change in Binding Free Energy with Explicit Probing of the Protonation State of the Active Site

As in all aspartyl proteases, the catalytic site of the HIV protease comprises two aspartic acid residues, of which only one is protonated in its active form (Hyland *et al.* 1991). In the HIV protease these residues are D25 and D25′ from the two subunits of the dimer that lie next to each other in the binding pocket (Figure 2.5). The three-dimensional structures of the HIV protease considered in this study have been obtained by X-ray crystallography, which typically does not provide information on the positions of hydrogen atoms. Hence it is unknown which of the two catalytic aspartate residues is protonated in each case. The apo form of the enzyme is symmetric, thus switching the protonation state of the two aspartic acid residues results in an identical dimer. On the other hand, protease inhibitors considered here are not symmetric molecules, and upon their binding to the protease, the resulting complex is no longer symmetric. Previous studies have suggested that the protonation state of the protease depends on the inhibitor bound (Baldwin *et al.* 1995; X. Chen and Tropsha 1995; Y. X. Wang *et al.* 1996; Tawa *et al.* 1998; Adachi *et al.* 2009; Kar and Knecht 2012a). For a complex with an experimental inhibitor including a diol group, a diprotonated HIV protease at its active site has been suggested (Yamazaki *et al.* 1994), but such a state seems to be rather an exception (McGee Jr *et al.* 2014).

In this study both monoprotonated active site states are considered in order to find the complex with the lowest free energy. In each case studied, MD simulations were conducted in both alternative protonation states (DH25/D25′ and D25/DH25′) for the wildtype and the mutant. The reference protonation state was identified with Propka (Søndergaard *et al.* 2011), with the exception of complexes analysed for I50V*, where the previously suggested protonation was used (R. Duan *et al.* 2015) (Table B.1). We then estimated the difference of the binding free energy in the wildtype and the mutant

Drug	Mutation	$\Delta\Delta G_{exp}$	$\Delta\Delta G_{calc}$	$\Delta\Delta G^{prot}_{WT}$	$\Delta\Delta G^{prot}_{MUT}$	$\Delta\Delta G_{total}$
ATV	I50V	2.7[a]	0.02 ± 0.19	-1.80 ± 0.16	-1.16 ± 0.2	0.34 ± 0.32
LPV	I50V	2.6[a]	1.39 ± 0.15	-0.81 ± 0.26	-0.1 ± 0.48	1.74 ± 0.57
TPV	I50V	2.1[a]	0.9 ± 0.18	1.04 ± 0.3	-0.56 ± 0.21	0.1 ± 0.41
APV	I50V	2.5[a]	1.16 ± 0.17	-1.24 ± 0.17	-0.81 ± 0.27	1.38 ± 0.37
IDV	I50V	1.9[a]	0.82 ± 0.12	0.91 ± 0.29	0.73 ± 0.27	0.73 ± 0.41
APV	I50V*	2.03[b]	2.11 ± 0.15	1.47 ± 0.35	-0.33 ± 0.45	1.21 ± 0.58
IDV	I50V*	2.33[c]	0 ± 0.1	-2.49 ± 0.45	-1.82 ± 0.76	0.32 ± 0.89
SQV	G48V	2.78[d]	3 ± 0.58	0.67 ± 0.15	0 ± 0.24	2.66 ± 0.65
SQV	L90M	1.60[d]	-0.09 ± 0.24	—”—	1.03 ± 0.18	0.09 ± 0.33
SQV	G48V/L90M	4.03[d]	5.32 ± 0.74	—”—	2.1 ± 0.21	6.03 ± 0.79

Table 4.2: Experimental and estimated $\Delta\Delta G$ values for **dataset 1** in kcal/mol. [a] $\Delta\Delta G_{exp}$ taken from (Muzammil *et al.* 2007). [bc] K_i values taken from (Shen *et al.* 2010) and (Liu *et al.* 2005) respectively and ΔG_{exp} calculated using $\Delta G_{exp} = -RT \ln K_i$. [d] K_i values taken from (Ermolieff *et al.* 1997) and (Maschera *et al.* 1996), ΔG_{exp_1} and ΔG_{exp_2} calculated using $\Delta G_{exp_j} = -RT_j \ln K_{i_j}$ and ΔG_{exp} at 300 K temperature then calculated by linear interpolation between ΔG_{exp_1} and ΔG_{exp_2}.

complexes between the two alternative protonation states, $\Delta\Delta G^{prot}_{WT}$ and $\Delta\Delta G^{prot}_{MUT}$ (Table 4.2, Figure 4.1). Large differences in the binding free energy between the two protonation state alternatives can be observed in several cases, e.g. -2.49 kcal/mol for IDV bound to the wildtype protease. Interestingly, the reference protonation state corresponds to the lowest free energy complex only in 44% of the cases (8 out of 18), indicating that Propka cannot be used as a reliable means for predicting protonation for the HIV protease. For some inhibitor/mutant combinations, the lowest free energy complex for the mutant has the opposite protonation state compared to the wildtype, i.e. the mutation affects the protonation probabilities upon inhibitor binding. Concomitantly, the interactions in the vicinity of the active site differ between the wildtype and the mutant enzymes, as demonstrated by the interaction energy analysis in Section 4.3.3. Analysis of an experimental structure of HIV protease in complex with APV resolved with X-ray/neutron crystallography (Weber *et al.* 2013), where the protein has the same sequence as the wildtype in our analysis of the I50V* mutation supports our assertion. In this structure, the hydrogen atoms on the aspartic dyad were resolved and the protonated aspartic acid residue matched the lowest free energy prediction for both complexes with APV from our study.

To further test the accuracy of our method in predicting the energetically favourable active site protonation state, we searched the PDB for structures of HIV protease with experimentally resolved hydrogen atoms. This resulted in a dataset of four additional protease complexes, including one with APV, two complexes with DRV resolved under different pH conditions, and one with an experimental inhibitor KNI-272. In all cases, the free energy change $\Delta\Delta G^{prot}$ of switching the proton from active site aspartic acid, which was resolved experimentally, to the active site aspartic acid on the opposite monomer was estimated to be energetically unfavourable (Table B.2). This included the two complexes with DRV, where the only manifestation of differences in experimental pH values in our simulations was in terms of different protonation states of several residues, including the

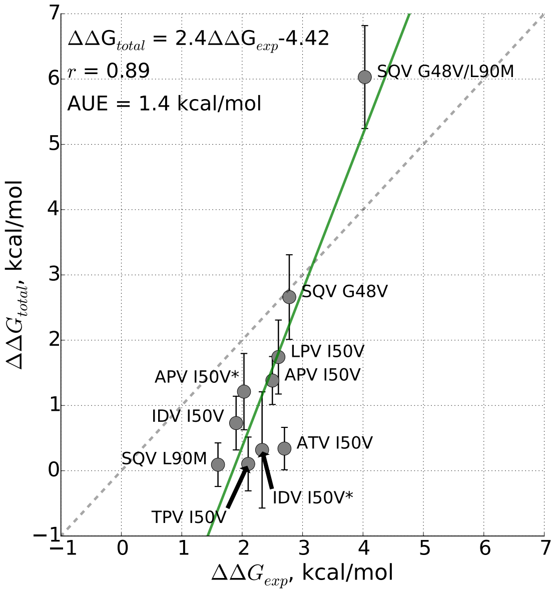
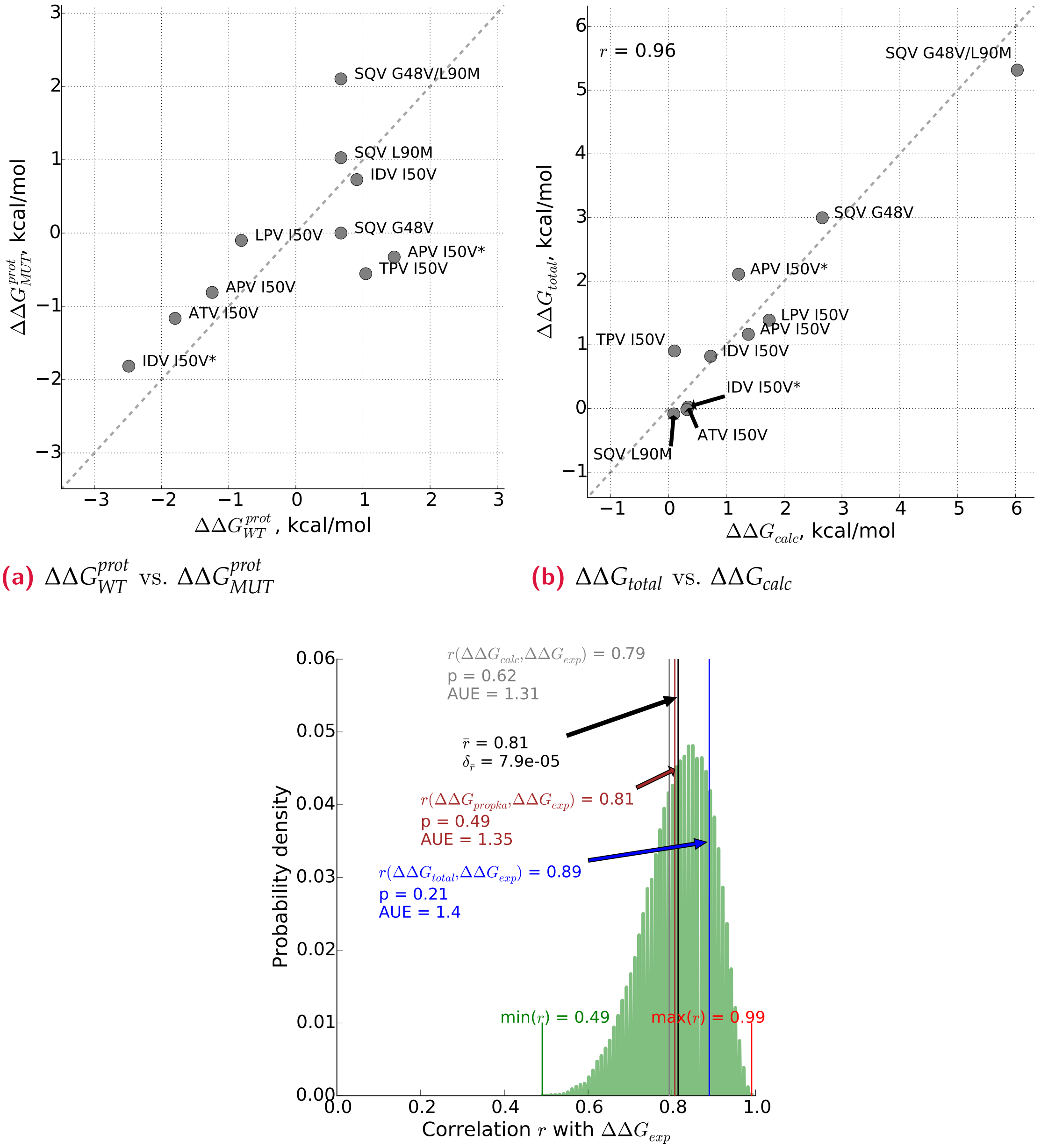


Figure 4.2: Calculated total estimate and experimental measurement of binding free energy change upon mutation for dataset 1.

active site. Here we correctly predicted the opposite active site protonation states of these complexes.

Taking into account the alternative protonation states of the active site, we measured the total difference of the binding free energy between the wildtype and the mutant complex, $\Delta\Delta G_{total}$, as defined in Equation 4.3. Considering the alternative protonation states results in a correlation between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{total}$ of 0.89 (average unsigned error 1.4 kcal/mol) (Figure 4.2).

It must be noted that for many of the complexes studied, there is a compensation in the changes of the binding free energy caused by switching the active site protonation state between the complexes. That is, if the simulation of a wildtype protein in the reference state was in an energetically unfavourable protonation state, the corresponding mutant protein also had a tendency to be in an energetically unfavourable protonation state which means they compensated each other to some degree in the overall estimate of $\Delta\Delta G$ (Figure 4.3a). This results in a high degree of correlation between $\Delta\Delta G_{total}$ and $\Delta\Delta G_{calc}$ (Figure 4.3b). Thus we evaluated how accurate can one expect to be in estimating $\Delta\Delta G$ if one were to choose randomly which of the active site residues is protonated in both wildtype and mutant protein. For that purpose all unique combinations of $\Delta\Delta G$ estimates for the set of all ten complexes were probed, where $\Delta\Delta G$ was a result of any possible linear combination of $\Delta\Delta G_{calc}$ with $\Delta\Delta G^{prot}_{WT}$ and $\Delta\Delta G^{prot}_{MUT}$. For each of these sets its correlation r with $\Delta\Delta G_{exp}$ was calculated, which resulted in 4^{10} data points (Figure 4.3c). Correlation between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{total}$ ($r = 0.89$) was higher than with the mean $\Delta\Delta G$ from all possible combinations ($r = 0.81$), with $\Delta\Delta G_{calc}$ ($r = 0.79$) or $\Delta\Delta G_{propka}$ ($r = 0.81$) (the latter corresponded to the evaluation of the most likely protonation state as predicted by Propka separately for wildtype and mutant protease complexes). This confirms the benefit of combining the simulations of HIV protease in both active site protonation states for obtaining the best accuracy when estimating the effect of mutation on ligand binding free energy. Thus we used this approach in all further calculations of this value and from here on $\Delta\Delta G$ refers to the complete estimate including simulations in both protonation states.



(a) $\Delta\Delta G_{WT}^{prot}$ vs. $\Delta\Delta G_{MUT}^{prot}$

(b) $\Delta\Delta G_{total}$ vs. $\Delta\Delta G_{calc}$

(c) Correlations of $\Delta\Delta G$ and $\Delta\Delta G_{exp}$

Figure 4.3: Effect of protease D25/D25′ protonation choice on $\Delta\Delta G$ for dataset 1 through comparison of (a) free energy of switching the protonation between active site residues in wildtype and mutant proteins, (b) calculated total estimate and calculated estimate of free energy binding change upon mutation where a single, reference protonation state was used, and (c) distribution of correlations between the experimentally measured binding free energy change upon mutation and calculated one, where all possible combinations of protonation states of wildtype and mutant protein have been considered. p corresponds to the probability to observe correlation r higher or equal than the one in question when choosing a random protonation state.

4.3.2 Estimation of Resistance Factor Ratio from the Inhibitor Binding Free Energy Change

Mutation		Background	IDV	SQV	LPV	FPV
M46I	wildtype	V3, N37, V71	0.6, 1.0, 1.2	–	–	0.3, 0.6, 0.7
	mutant		4.4	–	–	2.2
I84V	wildtype	V3, F10, N37	1.5	1.3	1.6	1.8
	mutant		2.1, 3.2	2.7, 3.7	6.2, 7.7	4.6, 8.4
N88S	wildtype	V3, N37, P63, G57/I77*	1.1	–	–	1.0
	mutant		2.6	–	–	0.1

Table 4.3: RF values from the HIVdb, further referred to as **dataset 2**. First column indicates the mutation analysed, while the second column indicates background mutations in both wildtype and mutant sequences compared to the reference HIV sequence HXB2. * indicates that the G57 background mutation was found in the sequences where RF for IDV was measured and I77 in the sequences where RF for FPV was measured. *Nota bene:* M46I wildtype and mutant measurements are reported in different studies, but performed using the same susceptibility test method.

Given the successful application of the method in reproducing the experimental estimates of the mutations' effect on inhibitor binding affinity $\Delta\Delta G$, we went a step further to analyse whether the effect of RAMs on IC_{50} could be predicted. More specifically, we aimed at evaluating whether the experimentally measured RF_R between two protease complexes differing in one mutation can be reproduced using the inhibitor binding free energy change upon mutation in the protein. For this purpose we selected dataset 2 comprising of sixteen complexes from the HIVdb database (*The Stanford HIV Drug Resistance Database* n.d.) for which resistance factors for inhibitors APV, IDV, LPV, and SQV were measured experimentally as reported in the literature (Tables 4.3 and B.3). These complexes could be paired amongst each other such that the RF has been measured for the same inhibitor and the same protease strain with and without the mutation, namely: IDV and APV with mutation M46I; APV, IDV, LPV, and SQV with mutation I84V; APV and IDV with mutation N88S. Moreover, for all of these mutations there was a correspondence of wildtype and mutant protein background sequences across the different inhibitors, with the exception of N88S, where complexes with APV had a background mutation L77I and complexes with IDV had a R57G background mutation present. Both R57G and L77I mutations are found next to each other on two parallel beta-sheets on the side of the protease. Although they are close to the resistance-associated positions 76 and 54, respectively, unlike for those residues, side chains of residues 77 and 57 are pointing away from the protease binding pocket. Nevertheless the mutation R57G has been suggested to be a protease-inactivating mutation (Ott *et al.* 2003). L77I on the other hand, has been reported to be a compensatory mutation for I84V restoring protein stability (M. W. Chang and Torbett 2011). M46I, I84V, and N88S are all considered to be major RAMs against the corresponding inhibitors as reported in the HIVdb, with the exception of N88S, where sensitivity towards APV/FPV was reported (Ziermann *et al.* 2000; Resch *et al.* 2002; Vermeiren *et al.* 2007; Rhee *et al.* 2010). Accordingly, for all

Inhibitor	Mutation	$\Delta\Delta G$	RF$_R$
APV	M46I	-0.35 ± 0.4	3.67–7.33*
IDV	M46I	-0.34 ± 0.72	3.14–7.33
APV	I84V	2.06 ± 0.47	2.56–4.67*
IDV	I84V	1.13 ± 0.57	1.4–2.13
LPV	I84V	1.25 ± 0.39	3.88–4.81
SQV	I84V	0.56 ± 0.41	2.08–2.85
APV	N88S	-0.97 ± 0.7	0.1*
IDV	N88S	1.41 ± 0.95	2.27

Table 4.4: Change of the binding free energy $\Delta\Delta G$ of inhibitors upon mutation for dataset 2, values in kcal/mol, and ± shows bootstrap error estimate. Last column indicates the RF$_R$ value range from previously reported experimental RF measurements, where * stands for measurements for FPV, the prodrug of APV.

of these pairs, protease with a RAM had a higher RF than the wildtype, while N88S reduced resistance towards APV in dataset 2.

We performed MD simulations of all the complexes and calculated the effect of mutations on the free energy of inhibitor binding. Overall resulting $\Delta\Delta G$ calculations indicated a good agreement between the effect of mutations on the free energy of inhibitor binding and their effect on RF, including the opposite effects of N88S towards IDV and APV (Table 4.4; for protonation switch estimates see Table B.4). An exception to that is M46I, where the mutation had a modest effect on ΔG which was within the error estimate.

A possibility of effect of mutations on the catalytic activity of the enzyme, K_m, precludes direct comparison of the $\Delta\Delta G$ estimates of mutation effects on the free energy of inhibitor binding and the RF$_R$ corresponding to that mutation. In previous studies on resistance mutations of another enzyme of HIV, reverse transcriptase, $\Delta\Delta G$ was considered to approximate changes in IC$_{50}$ (Rizzo *et al.* 2000; D.-P. Wang *et al.* 2001; Udier-Blagović *et al.* 2003). This is however a strong assumption, at least in case of the HIV protease, in which mutations were reported to affect its catalytic activity (X. Chen and Tropsha 1995; Pazhanisamy *et al.* 1996; Schock *et al.* 1996; Nijhuis *et al.* 1999). Similarly to the case of the reverse transcriptase studies, we had available only the experimentally measured effects of mutation on IC$_{50}$ for the same enzyme but different inhibitors. To account for the potential problems mentioned, we developed a Bayesian method which combines estimates of the effect of mutation on K_m from different complexes to calculate RF$_R$ (see Section 4.2.6). We then compared the estimated RF$_R$ values to the their experimental measurements (Figure 4.4). The increase of resistance towards inhibitors (RF$_R$> 1) was correctly predicted for all mutations, as was the sensitizing effect of N88S towards APV (RF$_R$< 1), and the experimental RF$_R$ values were within the corresponding calculated distributions based on the $\Delta\Delta G$ estimates (Figure B.1).

The third and final dataset which we analysed, dataset 3, included RF measurements of HIV protease acquired from clinical isolates in the presence and absence of hydrophobic core RAM L76V (Table 4.1). The RF measurements confirmed the resistance-inducing effect of this mutation against IDV and LPV, as well as the sensitizing effect towards ATV and SQV reported previously (T. P. Young *et al.* 2010; Wiesmann *et al.* 2011). Since,

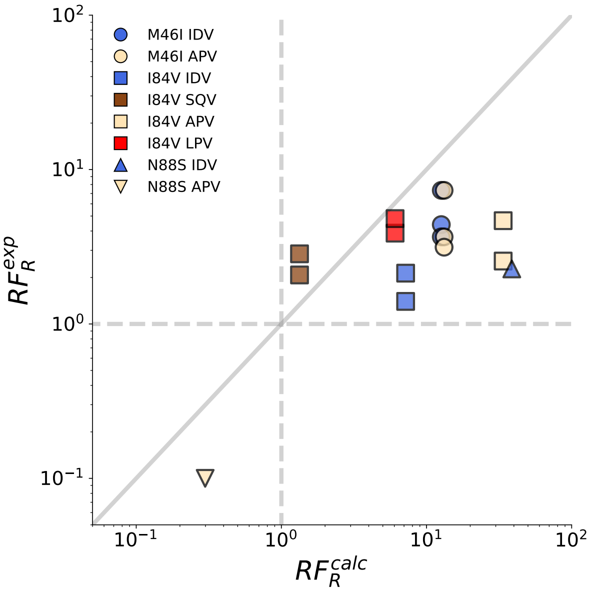


Figure 4.4: Predicted and experimental RF measurements for dataset 2. Each symbol corresponds to a unique sequence background and each colour corresponds to an inhibitor. *Nota bene:* in case of APV, RF_R^{exp} measurements are for its prodrug FPV.

similarly to the case of dataset 2, the effect of the mutation on inhibitor binding was measured in different background sequence contexts and towards different inhibitors, the same approach for estimating RF_R from $\Delta\Delta G$ could be applied for this dataset, too. Because the sequences of the protease complexes analysed had a large number of background mutations accumulated (compared to the reference sequence HXB2), it was difficult to find complexes in PDB with similar sequences. Thus in the protein modelling stage between 11 and 19 mutations had to be introduced to create protein models with sequences corresponding to those for which RF_R was measured (see Section 4.2.2). Including the target mutation L76V as well meant that up to 20% of protease residues had to be modelled *in silico*.

First we estimated the effect of the mutation L76V on inhibitor binding in terms of the change of the binding free energy $\Delta\Delta G$ (Table 4.5). The increase of the binding free energy, corresponding to the decrease in inhibitor affinity, was predicted for all complexes where mutations were observed to increase the protease RF. The decrease of RF, on the other hand, did not always correspond to a negative value of $\Delta\Delta G$: L76V was predicted to increase the affinity of inhibitor binding for inhibitors ATV and SQV only for the genotype GH9, but not for the genotype FB15, nor for inhibitor ATV in the context of the genotypes RU1 or iZ2. Of these genotypes, FB15 and iZ2 lack the background mutation M46I (the former being wildtype at that position and the latter having mutation M46L), which has been suggested to co-occur with L76V to compensate for its compromising effect on the replication capacity of HIV (Nijhuis *et al.* 2009; Louis *et al.* 2011).

We then used the $\Delta\Delta G$ estimates to calculate RF_R (Figure 4.5). For most of the complexes we correctly predicted whether the mutation made the protein more resistant or more sensitive towards the inhibitor. This included the prediction of sensitizing effect of mutation in the genotype FB15 for both ATV and SQV for which the inhibitor affinity

Inhibitor	Mutation	$\Delta\Delta G$	RF_R
ATV	FB15	0.78 ± 0.64	0.01–0.07
SQV	FB15	0.69 ± 0.62	0.04–0.08
ATV	GH9	-0.29 ± 0.49	0.01–0.04
SQV	GH9	-0.58 ± 0.47	0.06–0.08
ATV	RU1	0.74 ± 0.73	0.27–0.32
LPV	RU1	1.52 ± 0.6	3–5
ATV	iZ2	1.11 ± 0.65	0.14–0.53
IDV	iZ2	1.76 ± 0.66	5.9–26.82
LPV	iZ2	0.99 ± 0.88	5.91–12.9

Table 4.5: Change of the binding free energy $\Delta\Delta G$ of inhibitors upon mutation L76V in dataset 3, values in kcal/mol, and ± shows bootstrap error estimate. Last column indicates the RF_R value range from experimental RF measurements.

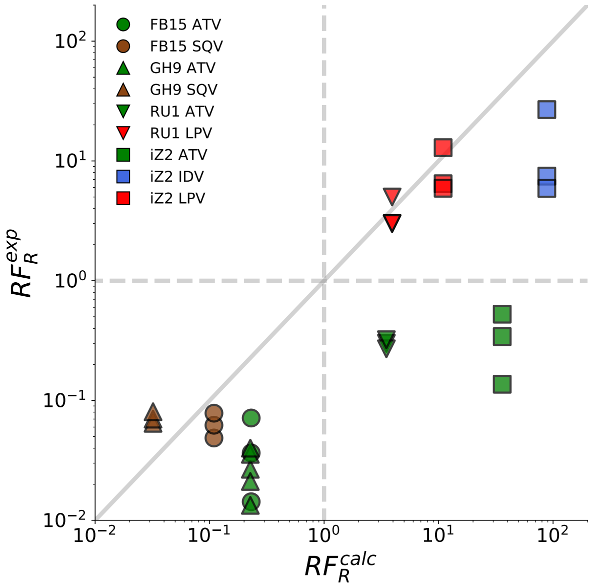


Figure 4.5: Predicted and experimental RF measurements for dataset 3. Each symbol corresponds to a unique sequence background and each colour corresponds to an inhibitor.

increased based on the $\Delta\Delta G$ estimates. Sensitization towards ATV was on the other hand not observed for genotypes RU1 and iZ2. The experimental RF_R values were however within the corresponding calculated distributions based on the $\Delta\Delta G$ estimates (Figure B.2).

4.3.3 Molecular Mechanisms of Resistance

Energetic Contributions of Individual Residues

To elucidate the energetic interactions of individual residues with the inhibitor, for each of the residues contributions of Lennard-Jones and short range electrostatic terms were

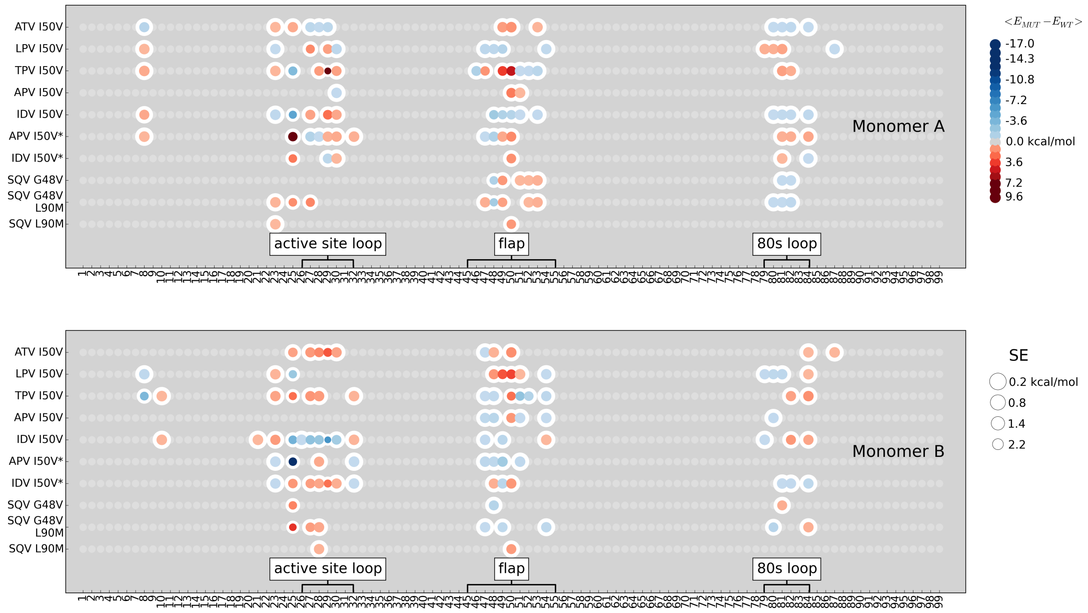


Figure 4.6: Energy differences of non-bonded interaction between protein and inhibitor in wildtype and mutant complexes. Residues, for which the difference ($E_{MUT} - E_{WT}$) between the wildtype and mutant complexes is higher than the propagated error (SE) and its absolute value is higher than 0.1 kcal/mol, are represented as a coloured circle, where the colour represents the relative interaction energy and the size of the circle relates inversely to the standard error of the estimate.

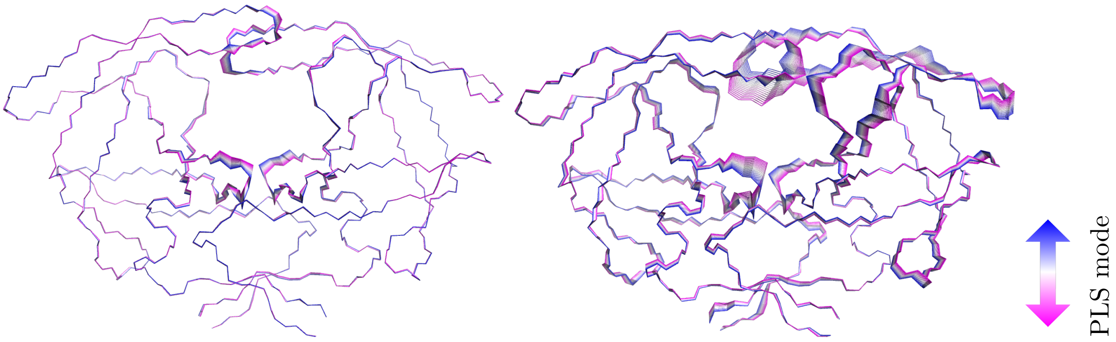
calculated and their comparison was made between the wildtype and the mutant complexes for RAMs G48V, I50V, and L90M, referred to as dataset 1 in this dissertation (Figure 4.6). In all cases, only the complexes with the protonation states corresponding to the lowest free energy were considered.

For the inhibitors ATV, LPV, TPV, and APV, the mutation I50V leads to an increase in the interaction energy between the inhibitor and the protein (Figures 4.6 and B.3), which may account for the observed resistance in this phenotype. The experimental measurements from ITC for ATV, LPV, and APV reported an enthalpic penalty $\Delta\Delta H$, i.e. losses in the binding enthalpy of inhibitor upon mutation, which also includes the direct protein–inhibitor interactions, caused by the I50V mutation (Muzammil *et al.* 2007). This was not the case for TPV, where a reduction in $\Delta\Delta H$ was reported for the same mutation (Muzammil *et al.* 2007). Mutation A71V, which tends to appear together with I50V (Mittal *et al.* 2013), is known to compensate for the loss of viral fitness due to primary RAM (Nijhuis *et al.* 1999). An ITC study of I50V+A71V double mutant (Mittal *et al.* 2013) suggested an increase in affinity toward ATV as a result of an increase in entropy which compensated the increase in enthalpy. The same compensatory effect in that study was observed for APV, but with the overall result of a binding penalty for the inhibitor. In the present study the mutation I50V* decreased the direct interaction energy between the protein and APV. This was mostly the result of a stronger interaction with the active site residues due to a difference in protonation state preference between the wildtype and mutant. To test whether the addition of mutation A71V results in a different thermodynamic profile, additional simulations of a 50V+71V mutant with the ATV and APV inhibitors were performed. The resulting $\Delta\Delta G$, 0.91 and 1.27 kcal/mol respectively, was close to the original estimates for a single mutation I50V for these inhibitors and it exhibited comparable direct protein–inhibitor interaction profiles (Figure B.3). Thus, the compensating effect of the double mutation in our simulation could not be reproduced. Similarly to APV, I50V and I50V* showed differing protein interaction profiles with IDV, with the former exhibiting stronger direct interactions and the latter having a small penalising effect on the interactions after mutation (Figures 4.6 and B.3). In combination with SQV, mutations G48V, L90M, and G48V/L90M had an unfavourable effect on the direct protein–inhibitor interaction energy, albeit to a smaller extent for the single-site mutations (Figure 4.6).

Direct protein–inhibitor interaction profiles were also calculated for the L76V mutation in the context of different background sequences (Figure B.4). We observed that the mutation has an unfavourable effect on direct protein–inhibitor interactions for all complexes with LPV (albeit to a small extent for the genotype iZ2) and with SQV; a favourable effect was observed for complexes with ATV (albeit to a small extent for the genotype RU1) and IDV. It should be noted that for mutations which change the preference of the protonation state of the active site residues D25/D25$'$ of the protease, namely FB15 and GH9 with SQV, as well as iZ2 with IDV (Table B.5), the largest energetic contributions overwhelming all others come precisely from the active site residues. Interestingly, measurable differences were observed for interactions of the residue at position 76 with the inhibitor for complexes of genotype FB15 with LPV and ATV and genotype iZ2 with IDV. But given that the side chain of residue 76 has minimal exposure to the binding pocket, those differences are small. Overall these effects of a mutation on the direct protein–inhibitor interactions are only partially in line with the effect of the mutation on binding free energy of the inhibitor.

Changes in Protease Structure and Residues Interactions

For all complexes with mutations G48V, L90M, and particularly I50V analysed (dataset 1), if the substitution had any notable effect on the direct interaction energy between protease and inhibitor, such effect extended beyond the mutated residue itself. The largest contributions were observed for residues in the flap region, i.e. near the mutation site. Other proximal affected residues were in the 80s loop, as well as around the active site (Figures 4.6 and B.3). A similar effect was observed for mutation L76V (dataset 3) (Figure B.4). Given such a wide distribution of mutation effects, we decided to investigate for dataset 1 whether this is related to structural changes in the protease or to a different pose of the inhibitor. Functional mode analysis based on PLS regression (Krivobokova *et al.* 2012), a supervised machine learning technique, allows to study the correlation of Cartesian input from MD trajectories to a desired functional property. In the present study we investigated whether the major collective motions discriminate the wildtype and mutant complexes: using Cartesian coordinates as an input, we attempted to predict whether a trajectory was generated by a wildtype or a mutant protein. This technique was applied with different input features: coordinates of heavy atoms of the backbone, inhibitor, protein, and side-chain, to create statistical models. Of these, for mutations of dataset 1 models with protein backbone input proved to be the most predictive and models with inhibitor input proved to be the least predictive in the cross-validation on average (Figure B.5). Models trained on protein backbone atoms proved to be capable of making a satisfactory distinction for all complexes except IDV in the context of I50V* and complexes with SQV in the context of the G48V or L90M mutations. Mapping the



(a) I50V TPV

(b) I50V* APV

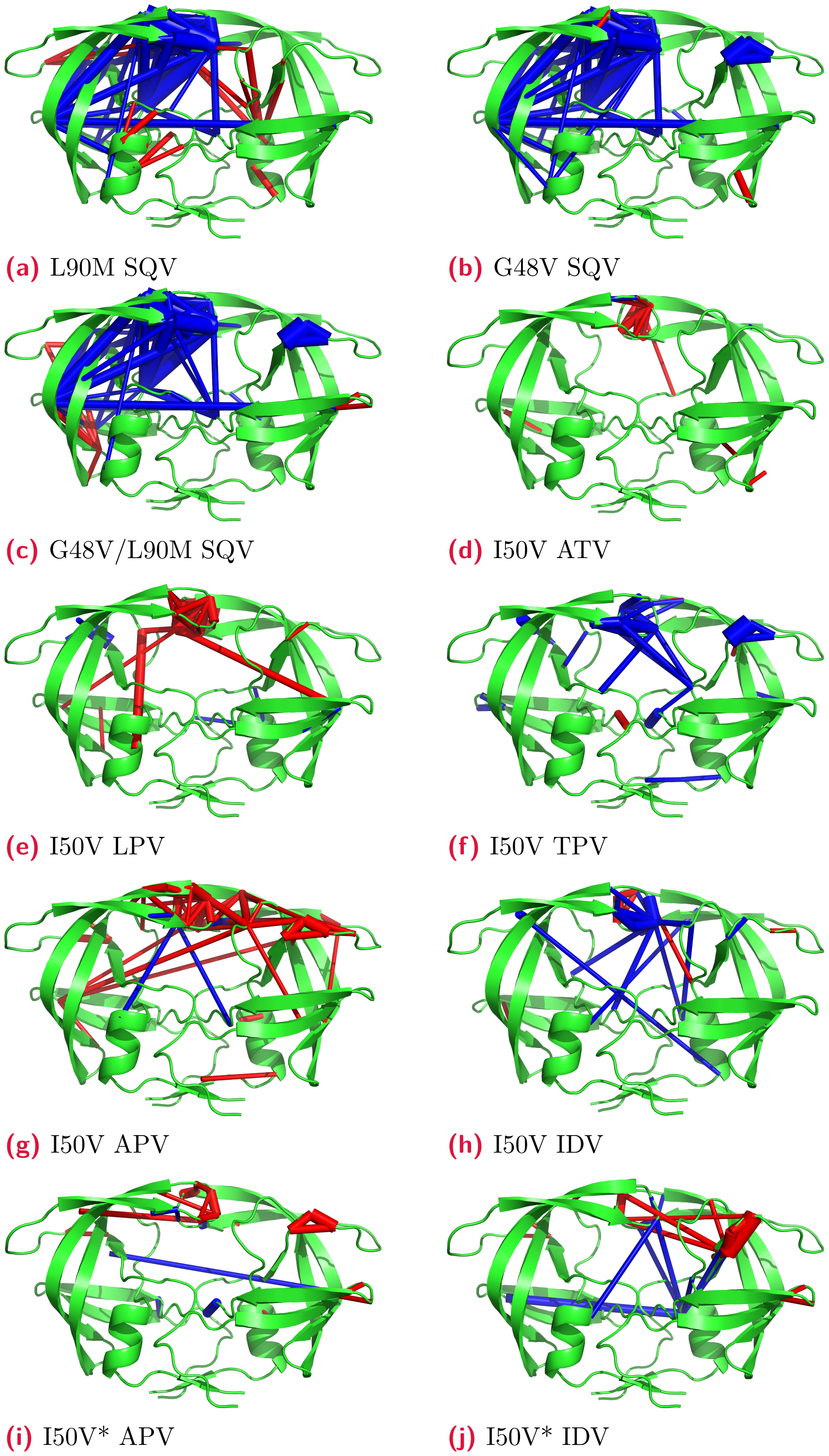
Figure 4.7: Interpolation between the extremes of the PLS models for the corresponding complexes. Blue-to-magenta bands correspond to the interpolation along the mode which relates the true label of simulation, wildtype or mutant, to the underlying differences in protein motions.

regions of the backbone contributing the most to the discriminative power of these PLS models confirmed the flap as the region that is the most important for distinguishing the trajectories of wildtype and mutant complexes (Figures 4.7 and B.6). The 80s loop is also prone to assume different conformations between wildtype and mutant complexes, as well as the loop with the active site residue, particularly so for cases in which the energetically favourable protonation state differs between wildtype and mutant proteins (Figures 4.7a and 4.7b).

The analysed mutations seem to have an effect on direct protein–inhibitor interaction as well as on protein backbone rearrangements outside of the immediate mutation site. Certainly, this must be the case for RAMs which are not in direct contact with the inhibitor, such as the mutation L90M. However no consistent effect of this mutation was observed in terms of changes in the protein backbone and just a minor effect on the protein–inhibitor interaction. To gain insight into other non-local effects of the mutations, differences in the correlated motion of all pairs of residues between the wildtype and the mutant protease-inhibitor complexes were inspected. Since these correlations are not necessarily linear, it was chosen to evaluate the pairwise mutual information based on the distributions of dihedral angles from molecular dynamics simulations (McClendon *et al.* 2009), a measure that depends both on individual Shannon entropies of the residues and their joint entropy.

First we inspected in which protein regions correlations are affected most by inhibitor binding. For this purpose residues' mutual information between apo and holo wildtype proteins were compared (Figure B.7). The flexible flap region for all of the complexes is affected to the largest extent. Correlations in the active site loop of the protease, as well as the 80s loop also, seem to be affected in many cases, particularly so upon binding of TPV and SQV. These results are expected, since the flap and 80s loop regions are suggested to be involved in inhibitor binding (Spinelli *et al.* 1991; Perryman *et al.* 2004; Perryman *et al.* 2006; Y. Yu *et al.* 2015), while the active site D25/D25′ typically makes hydrogen bonds with the inhibitor upon its binding, thus likely affecting its interactions with other residues in the protein.

Next, we estimated the mutual information network for holo protein complexes with mutations and compared it to the wildtype holo protein complexes in order to evaluate how the mutation affects residues' correlated motions in the protein (Figure 4.8). We



(a) L90M SQV

(b) G48V SQV

(c) G48V/L90M SQV

(d) I50V ATV

(e) I50V LPV

(f) I50V TPV

(g) I50V APV

(h) I50V IDV

(i) I50V* APV

(j) I50V* IDV

Figure 4.8: Mutual information mapped onto the protease structure. Cylinders connecting residues represent differences in mutual information between those residues in mutant and wildtype simulations, with the width of the cylinder proportional to the difference and red indicating higher correlation in mutant and blue indicating higher correlation in the wildtype protein. This corresponds to the degree to which residue pairs exhibit differences in the correlation of their motions.

measured what is the overlap between such pairs with those whose correlations were affected by inhibitor binding in the wildtype complexes. To evaluate the overlap between these sets, the Jaccard index was used, which is the ratio of the size of the intersection of the two sets and the size of their union. A Jaccard index close to 1 indicates that there is a high overlap between the sets of such pairs, whereas an index close to 0 would indicate that different residue pairs show correlated movements in the processes related to binding and propagation of mutation effects. An average Jaccard index value of 0.28 indicates that overall similar residue pairs are involved in these processes, but some differences exist (Table B.6). In contrast, when we choose random pairs to form sets of the same respective sizes as in the original data, an average Jaccard index value is 0.003. More generally, upon closer inspection it was observed that while the identity of the residues in these pairs may differ, they are located in the same protein regions: the flaps, the 80s loop, and the active site loop (Figures 4.8 and B.7).

Finally we analysed in detail the effect of mutations on the protein mutual information network. For the mutation L90M, the correlated motion of the mutation site with the active site residues D25/D25′ is changed. The active site residues directly interact with the inhibitor (Figure 4.8a), suggesting a path by which the mutation L90M, which is not located in the inhibitor binding pocket, affects its binding. The changed interaction pattern between L90M and the active site residues has been observed previously in crystal structures (Hong *et al.* 2000) and MD simulations (Ode *et al.* 2006). It has also been noticed that the side chain I84/I84′, which is typically in direct contact with D25/D25′, is oriented differently upon the mutation L90M (Ode *et al.* 2006), suggesting changes in the binding pocket of the protease. Indeed, in the present analysis, I84 demonstrates significant differences in the correlated motion with multiple residues in the binding pocket, as reflected by high values of the mutual information, for all three G48V, G48V/L90M, and L90M mutants in complex with SQV. These differences are possibly the result of different χ angle distributions of I84 between wildtype and mutant complexes (Figure B.8). Strong differences in mutual information involving the flap region of the complexes with SQV were also observed (Figures 4.8a-4.8c) and to a smaller extent in complexes with other inhibitors (Figures 4.8d-4.8j). Suspecting that such large differences are related to large structural rearrangements in the protease upon mutation, we performed Principal Component Analysis (PCA) on the backbone of the complexes involving SQV. The motion along the first principal component, which accounts for most of the structural variance, indeed showed the largest shifts in the 80s loop and the flap region (Figure B.9).

Next we focused on the structural changes of the protease in the genotypes FB15, GH9, RU1, and iZ2 in dataset 3 where the mutation L76V was introduced, for which purpose first we analysed the hydrogen bond network of the protein. It could be consistently seen across all of the different genotypes, that the mutation L76V increases the probability of observing hydrogen bond between residues D30 and K45 (Figure 4.9 and Table 4.6). These residues are located in the active site loop and flap respectively, both regions for which the direct interaction energies with inhibitor are affected by the mutation, including residues D30 and to a smaller degree K45 (Figure B.4). Seeking whether this was as a result of side chain rearrangement, we performed the PLS regression analysis on the heavy atoms of the protein (excluding mutation site). In analysing PLS models, we could see that the mutation L76V caused the tendency of the side chains of residues D30, K45, and Q/E58 to shift towards the binding pocket (Figure B.10). This shift is likely the result of fine rearrangement of residues in the region as a consequence of a larger side chain of leucine being replaced by a smaller valine. Specifically, the changes

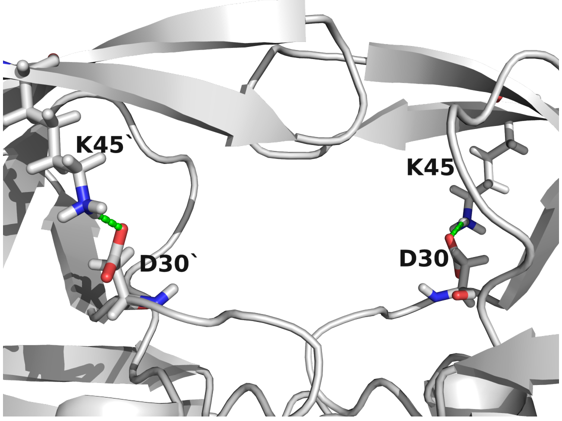


Figure 4.9: Hydrogen bonds between the protease residues D30 and K45 are shown with green dashed lines.

Drug	Genotype	D30–K45		D30′–K45′	
		L76	V76	L76	V76
ATV	FB15	0.068 ± 0.003	0.52 ± 0.04	0.07 ± 0.002	0.58 ± 0.02
SQV	FB15	0.12 ± 0.01	0.59 ± 0.02	0.11 ± 0.005	0.62 ± 0.004
ATV	GH9	0.07 ± 0.01	0.49 ± 0.12	0.06 ± 0.002	0.66 ± 0.12
SQV	GH9	0.07 ± 0.004	0.46 ± 0.04	$0.01 \pm 4 \times 10^{-5}$	0.19 ± 0.04
ATV	RU1	0.05 ± 0.001	0.43 ± 0.05	$0.04 \pm 1 \times 10^{-4}$	0.61 ± 0.09
LPV	RU1	$0.01 \pm 3 \times 10^{-5}$	0.16 ± 0.03	0.16 ± 0.006	0.56 ± 0.12
ATV	iZ2	0.16 ± 0.007	0.57 ± 0.07	0.08 ± 0.002	0.52 ± 0.04
IDV	iZ2	0.1 ± 0.01	0.39 ± 0.03	0.06 ± 0.002	0.51 ± 0.07
LPV	iZ2	$0.04 \pm 5 \times 10^{-4}$	0.19 ± 0.02	0.24 ± 0.01	0.63 ± 0.02

Table 4.6: Average hydrogen bonds number between residues D30 and K45 for protease wildtype and mutant complexes in dataset 3. ± indicates standard error of bond frequency across independent simulations.

of the D30/D30′ interactions with the inhibitor are in general in line with changes of RF_R, namely, negative (or favourable) interaction energy corresponding to $\mathrm{RF}_R < 1$ and positive (or unfavourable) interaction energy corresponding to $\mathrm{RF}_R > 1$. Exceptions to that are the protease of the iZ2 genotype when in complex with IDV, where a favourable effect on D30/D30′ direct interaction energy with inhibitor is observed (Figure B.4h), and the same enzyme in complex with LPV, where no notable effect on this interaction can be suggested from the simulations (Figure B.4i).

Given the observation that the mutation L76V affects the hydrogen bonding network of the protease, question arose on whether another mutation distant from the active site analysed in dataset 2, N88S, potentially also affects the protein's hydrogen bond network. In the analysis of N88S in complex with both APV and IDV the following effect was observed: S88 formed a hydrogen bond with D30 more frequently than N88, and N88 formed a hydrogen bond more frequently with T31 and T74 compared to S88 (Table B.7).

4.4 Discussion

In this study, we demonstrate the applicability of alchemical free energy calculations for an accurate estimation of the change of binding free energy, $\Delta\Delta G$, for different mutations in the HIV protease and different inhibitors. Taking into account alternative protonation states of the two aspartates in the protease active site improves in correlation with experimental $\Delta\Delta G$ values, compared to choosing one state based on empirical pK_a calculations with Propka (Søndergaard *et al.* 2011). A recent study of protonation of aspartic proteases, including HIV-1 protease, also suggested inaccuracy of prediction of protonation state using Propka (Huang *et al.* 2017). Although the importance of choosing the correct protonation state has been noted in multiple studies (Baldwin *et al.* 1995; X. Chen and Tropsha 1995; Y. X. Wang *et al.* 1996; Tawa *et al.* 1998; Wittayanarakul *et al.* 2008; Adachi *et al.* 2009; Kar and Knecht 2012a), predictions based on empirical pK_a estimates, as well as setting the protonation state based on a previous suggestion for protein–inhibitor complex, regardless of the specific sequence context, still remain a standard practice. The current study suggests that an explicit probing of both protonation states is needed to reproduce the correct ensemble. It was observed that choosing an alternative protonation state can contribute more than 2 kcal/mol to the change in free energy of the system. It was also shown that a point mutation can change the preferred protonation state of the protease, contrary to the often held assumption (Tawa *et al.* 1998).

For two of the complexes described here, the effect of the I50V* mutation on the binding free energy of APV and IDV has recently been addressed by Duan *et al.* (R. Duan *et al.* 2015) using the MM/PBSA approach. In their study, the authors could not reproduce the experimental values when sampling from 20 ns long MD simulations using the AMBER03 force field (Y. Duan *et al.* 2003) (Table B.8). In the present study, correct trends in free energy changes for these cases were obtained, and a hypothesis for the mechanism of resistance was provided. To compare our results for $\Delta\Delta G$ calculation from dataset 1 with the results of Duan *et al.*, snapshots from the first 20 ns of the trajectories for all complexes were used, totalling to 20 snapshots for each forward and backward transitions for ΔG calculations. This resulted in an overall correlation of -0.11 between $\Delta\Delta G$ and $\Delta\Delta G_{exp}$. Hence insufficient sampling might have been one of the issues contributing to the inaccuracies observed by Duan *et al.*

An unfavourable effect of mutation on inhibitor binding in dataset 1 has been correctly predicted for all of the cases analysed. Of these, predictions of $\Delta\Delta G$ were less accurate for complexes involving ATV and TPV for the mutation I50V, as well as to a lesser extent for IDV for the mutation I50V and for SQV in combination with the mutation L90M, where the complex destabilization of mutation was underestimated. The mutation I50V is however not considered to be a RAM against either ATV, TPV, or IDV inhibitors.

Change of binding free energy $\Delta\Delta G$ of inhibitor upon mutations M46I, I84V, N88S, and L76V was also computationally estimated in this study and, in absence of experimental $\Delta\Delta G$ values, it was compared to the effect of these mutations on experimentally measured RFs. In most cases a positive value of $\Delta\Delta G$ corresponded to an increase of RF and a negative $\Delta\Delta G$ to decrease of RF. Exceptions to this trend are the complexes of the protease with the mutation M46I and inhibitors APV and IDV, where a negligible favourable effect on ΔG was predicted despite $\mathrm{RF}_R > 1$ for both cases. While M46I has been associated with resistance towards different PIs, it typically appears in combination with other RAMs and it has been suggested to have a compensatory function for the protease's catalytic activity (Ho *et al.* 1994; Pazhanisamy *et al.* 1996; Schock *et al.* 1996;

Drug and mutation combination	ATV I50V	LPV I50V	TPV I50V	APV I50V	IDV I50V	APV I50V*	IDV I50V*	SQV G48V	SQV G48V/L90M	SQV L90M
p-value	0.12	0.05	0.02	2×10^{-3}	0.2	1×10^{-3}	0.07	0.06	0.05	3×10^{-5}

Table 4.7: p-values for Fisher's exact test for over-representation in holo protein of RAM sites amongst residues showing different correlations with other residues upon mutation or differences in direct interaction energies with inhibitor (absolute value larger than 0.1 kcal/mol).

Nijhuis *et al.* 2009; Louis *et al.* 2011; Henderson *et al.* 2012). Despite this inconsistency, our computational predictions of RF$_R$ were all in line with the experimental estimates.

Similarly, for dataset 3, there was also disagreement between $\Delta\Delta G$ and RF$_R$ for the mutation L76V in complexes with ATV and SQV (in FB15 genotype) and ATV (iZ2 genotype). Interestingly, both genotypes FB15 and iZ2 lack the background mutation M46I, which co-occurs with L76V to compensate the latter's diminishing effect on the replication capacity of HIV (Nijhuis *et al.* 2009; Louis *et al.* 2011). However for this mutation most of our computational estimates of RF$_R$ were in line with experimental measurements too. This suggests that in the absence of the compensatory mutation M46I, the dominant effect of the mutation L76V might be in terms of decreasing the protease catalytic activity K_m.

In the present study we show that the mutation I50V directly affects the interaction energy between protein and inhibitor, and the resulting impact on the enthalpy can be both unfavourable (APV, ATV, LPV, and TPV) and favourable (IDV).

Studying the structural changes in the protease, for the mutation I50V and even to a greater degree for the mutations G48V, L90M, as well as the combination of the latter two, we show that they perturb the correlated motions in the protein, both in the mutation site and beyond: this seems to be caused to a large extent by the movement of the protease flap region as well as the 80s loop. These regions, together with the loop proximal to the active site, harbour other sites in which mutations are associated with resistance toward PIs. In fact, major RAM sites are over-represented amongst residues whose interactions or dynamics is changed in MD simulations upon mutation, particularly in cases when the modelled mutation is a major RAM itself (Table 4.7). This hints at a possible relationship between different protease residues, mutations of which are associated with viral resistance toward inhibitors, and suggests their collective involvement in the process of inhibitor binding.

The mutation L76V on the other hand seems to affect the hydrogen bond between D30 and K45 residues, as their side-chains, as well as the side chain of the Q/E58 residue rearrange upon mutation. This has an effect on the direct interaction energies of D30 with the inhibitor, which is in general in line with the effect of this mutation on RF. These observations on energetic and structural consequences of this mutation agree with its previously reported effects on the ligand binding affinity for the inhibitor DRV through both changes in protein–inhibitor interactions and changes in the inter-residue distances in the binding pocket (Ragland *et al.* 2014). Another study suggested highly significant correlation between mutations of D30 and K45 (Margerison *et al.* 2008), indicating the importance of interaction between these two oppositely charged residues. Co-occurance of mutations L76V and Q58E has also been reported (Champenois *et al.* 2011) and both mutations were found in the patient samples RU1 in dataset 3.

Recently, a study was reported that analysed experimentally resolved wildtype and L76V mutant structures of the HIV protease with inhibitors DRV, LPV, TPV, as well as two experimental compounds, GRL-0519 and GRL-5010 (Wong-Sam *et al.* 2018). The study reported that mutation does not change the backbone structure of the protease, however residue 76 loses contacts with D30 and T74, and, for structures with LPV, a slight shift of K45 towards the binding pocket as a result of the mutation can also be observed. Overall similar interactions were reported between wildtype and mutant proteins with different inhibitors, with the exception of GRL-5010, which interacted with D30′ in an altered way. These results thus partially support the observations made in our study on the effects of the L76V mutation.

Substitution at another site of the protease distant from the binding pocket, N88S, has been suggested to create a hydrogen bond with residue D30, which in turn affects the interaction between D30 and the inhibitor NFV (Ode *et al.* 2007). Another mutation at this site that occurs in patients treated with NFV, N88D, has been reported to co-occur with mutation D30N, coinciding with losing interactions with residues T31 and T74 mediated by water molecules (Mahalingam *et al.* 2002). In the present study changes in direct hydrogen bonds upon N88S have been consistently observed between this site and D30, T31, and T74. Interestingly, mutation of N88, similarly to that of D30, has also been reported to be highly significantly correlated with substitution of K45 (Margerison *et al.* 2008). This points to the importance of an interaction pathway between RAM site 88 and sites 30 and 45 in a common undertone with the effect of L76V mutation on site 30 and 45 interaction.

5

Linear Binding Motifs Conferring Specificity of Protein–Protein Interactions in the MAPK-Related Pathways

In this chapter work on protein–protein interactions based on the linear motifs is presented. In the first part of this chapter, interaction of the docking D-motifs in the MAPKs is described. It was done in collaboration with András Zeke, Anita Alexa, Ágnes Garai, Bálint Mészáros, Klára Kirsch, Zsuzsanna Dosztányi, Olga V. Kalinina, and Attila Reményi and published in [Zeke et al. 2015]. The manuscript was written by András Zeke and Attila Reményi. Corresponding text and figures in Sections 5.1, 5.2, 5.3.1, 5.3.2, and 5.4 have been partially adapted from that publication. I, together with Bálint Mészáros and Olga V. Kalinina, performed the in silico analysis in this study. My main responsibility was to collect the candidate D-motifs from the human proteome and to build the predictive models for the D-motif interactome. Also presented in this chapter is the work on the clustering of phosphorylation sites. It was done in collaboration with András Zeke, Olga V. Kalinina, and Attila Reményi. I was responsible for the in silico part of this study, with the exception of tracing the conservation of the phosphorylation sites in vertebrates, which was implemented together with Olga V. Kalinina.

As outlined in Section 2.3, MAPK is a family of serine/threonine kinases well preserved in the eukaryotes, which plays a role in many important signalling pathways in the cell. Disruption of these pathways is associated with many of the major diseases affecting human health, such as cancer, neurological diseases including Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis (Kim and Choi 2015), as well as cardiovascular diseases among others (Muslin 2008). MAPK pathways have also been suggested to play a role in HIV replication and development of AIDS (Schrager *et al.* 2002; Toschi *et al.* 2006; Barraud *et al.* 2008; Medders and Kaul 2011). Crucial to the interaction between MAPKs and theirs targets, and thus for the specificity of the MAPK-related pathways, is the presence of a so-called D-motif in the interaction partner, a specific sequence motif in a protein disordered region, which binds to the surface of MAPK.

In this chapter, a study of MAPK interactome based on D-motif identification as well as on clustering of phosphorylation sites is presented. Section 5.1 briefly recapitulates the specifics of MAPK signalling that employ D-motifs. As background information of the study [Zeke *et al.* 2015], this section also introduces the idea of dividing D-motifs to separate classes based on their structural binding modes as well as results of dot-blot experiments testing different D-motifs' binding to MAPK. The main focus of Section 5.3.1 is presenting the different sequence-based computational predictive methods for identifying MAPK interaction partners from different classes of D-motifs. One of the methods, Position-Specific Score Matrix (PSSM), is then used to score the D-motifs from the candidate MAPK interaction partners, whereas their biological classification is discussed in Sections 5.3.2 and 5.4. In an additional study, an approach to find distinct clusters of phosphorylation sites targeted by proline-directed serine/threonine kinases is described in Section 5.3.3.

5.1 Introduction

Protein–protein interactions influence all aspects of cellular life and the most direct mechanism through which proteins can influence each other is by physical contact. This brings them into proximity so that they can exert control on each other's activity or to create opportunities for post-translational modification.

Protein–protein associations often involve so-called linear binding motifs which are short protein segments (5–20 amino acid long) lacking autonomous tertiary structure. These functional sites reside in the intrinsically disordered protein regions and adopt a stable conformation only upon binding. Currently, it can only be guessed how abundant linear motif-based interactions are; nevertheless, it was recently estimated that there are approximately 100 000 linear binding motifs targeting dedicated protein surfaces in the human proteome (Tompa *et al.* 2014).

As an example relevant to cellular signalling, MAPKs are prototypical enzymes that depend on short segments from partner proteins and on their dedicated protein–protein interaction hot spots. They mainly recognize their substrates, or target proteins, not with the catalytic site but with auxiliary docking surfaces on their kinase domains (Tanoue *et al.* 2000; Biondi and Nebreda 2003). As discussed in Section Target Recognition, the most important of these docking sites consists of a hydrophobic docking groove and the negatively charged CD groove regions (C.-I. Chang *et al.* 2002) (Figure 2.7). Together, they can bind the so-called D-motifs of the target proteins.

D-motifs are short linear motifs ranging from 7 to 18 amino acids in length and are typically found in the intrinsically disordered segments potentially far away from the target phosphorylation sites (Garai *et al.* 2012). Such docking elements are not only restricted to the substrates: they are also found in MAP2Ks, in MAPK phosphatases (MKPs), and in a variety of scaffold proteins. ERKs (ERK1 and ERK2), JNKs (JNK1, JNK2, and JNK3), and the p38s (p38α-δ) control diverse physiological processes, and they phosphorylate most of their substrates at serine-proline or threonine-proline ([ST]P) sequence motifs. This is a very promiscuous consensus to be the only element recognised by the catalytic site, and thus it is insufficient for selective target recognition, hence additional linear binding motifs provide specificity of binding (G. L. Johnson and Lapadat 2002; Bardwell 2006). Therefore, the MAPK D-motif protein–protein interaction system is an ideal test bed for linear binding motif discovery.

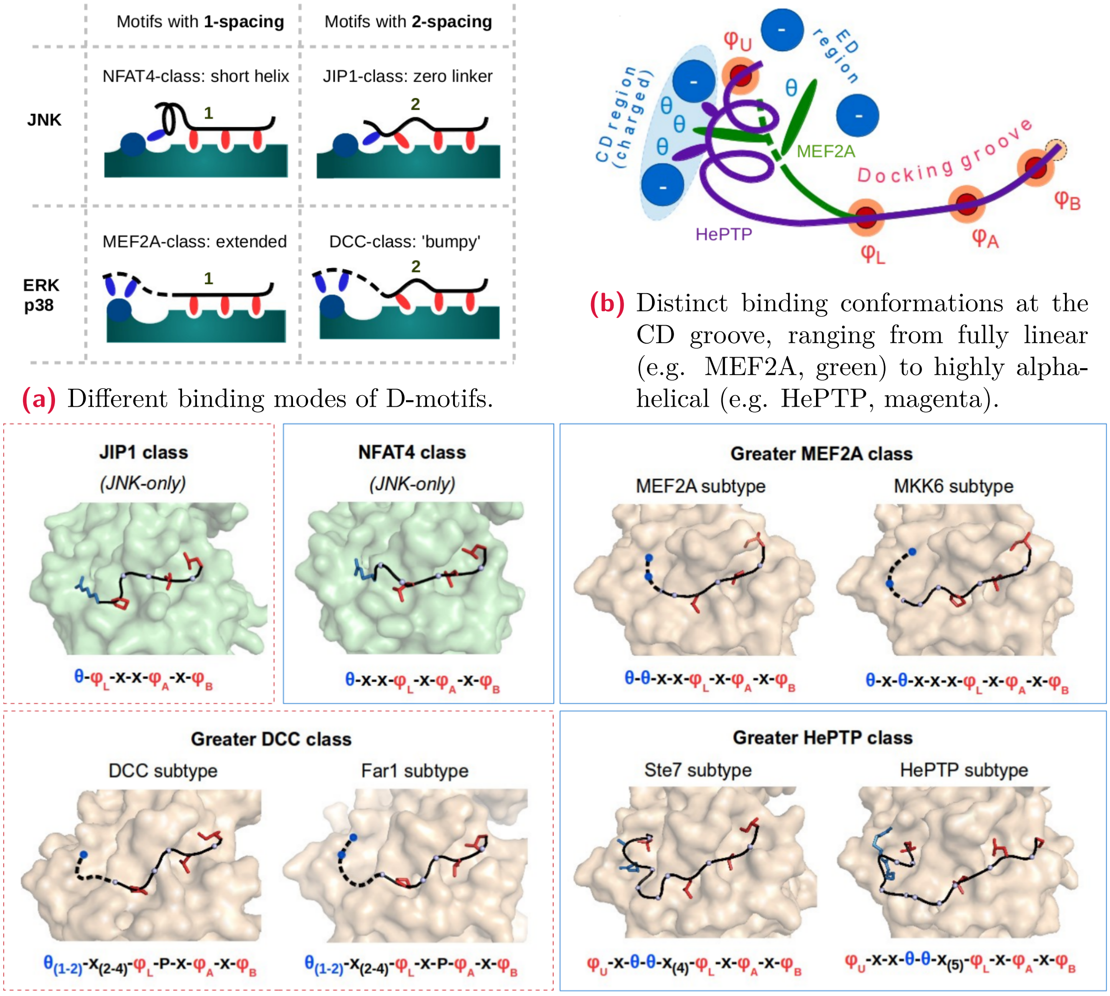
Several previous attempts were aimed at predicting MAPK-binding proteins from full proteomes by using a generic consensus of D-motifs as it had been established almost two decades ago (Sharrocks *et al.* 2000). This consensus was derived from an observation that D-motifs almost always include at least a single positively charged residue (termed the θ position: arginine or lysine) and a series of alternating hydrophobic residues (ϕ positions: frequently leucine), connected by a linker of a variable length (Dinkel *et al.*

2013). But despite the use of extensive multiple alignments and sophisticated algorithms, predictions had only low success rates and the large-scale assessment of predicted hits was not performed (Whisenant *et al.* 2010; Garai *et al.* 2012; Gordon *et al.* 2013).

In terms of experimental MAPK network discovery, ERK2 has been the most widely explored. For example, several different methods were utilized to identify ERK2 substrates by large-scale phosphoproteomics (Old *et al.* 2009; Pan *et al.* 2009; Kosako *et al.* 2009; Carlson *et al.* 2011; Courcelles *et al.* 2014). Unfortunately, pairwise overlaps between the lists of substrates are low across studies (e.g. ≈10%), with not a single overlap between five different studies that aimed to find ERK2-phosphorylated substrates (Courcelles *et al.* 2014), suggesting a great dependence on the experimental conditions used. It was noted that D-motif-like sequences are enriched in experimentally detected ERK2 substrates (Carlson *et al.* 2011), yet the detection or verification of direct physical association were not performed. In addition, studies that used a high-throughput approach to identify the partners of JNK1 (W.-K. Chen *et al.* 2014) or p38α (Bandyopadhyay *et al.* 2010) based on direct physical interaction resulted in low number of hits. Thus, it is likely that a protein–protein interaction-based MAPK network discovery could greatly benefit from a target-tailored approach, which takes into account — and possibly capitalizes on — specific biochemical and biophysical knowledge already available on known MAPK-partner protein interactions.

In recent years, the number of experimentally determined MAPK-partner protein complex structures increased considerably (Garai *et al.* 2012). This development made it possible to amend the definition of the underlying sequence motifs, and it became clear that D-motifs encompass multiple classes of similarly built, but structurally distinct linear motifs (similarly to the SH3 or PDZ-binding sequences) (Lim *et al.* 1994; Tonikian *et al.* 2008). In the current study, it is shown that by building a strategy that can handle this conformational diversity in binding, and using structural compatibility with specific interaction surface topography as an additional criterion for prediction, the identification of novel D-motifs can be dramatically improved. This analysis in combination with tailored experimental techniques for the validation of low-affinity (1-20 μM) protein–protein interactions produced unique, molecular-level insights into the physiological roles of MAPK-based protein networks.

Our analysis of MAPK–D-peptide complex structures revealed distinct D-motif binding modes in the MAPK-docking groove (Figure 5.1). The hydrophobic docking groove binds three hydrophobic amino acids in a row, while admitting two different spacing schemes. θ to ϕ linker length determines the MAPK specificity of a given motif. These two features can combine freely with each other, resulting in the four basic arrangements (Figure 5.1a). For example, D-motifs from the JNK-binding scaffold protein JIP1 and from the JNK-regulated transcription factor NFAT4 bind to the same docking surface differently (Heo *et al.* 2004; Garai *et al.* 2012; Laughlin *et al.* 2012). Similarly, ERK- and p38-binding D-motifs may also be structurally distinct. Nonetheless, D-motifs could be described with a common loosely defined consensus: $\theta_{1-2}x_{0-5}\phi_L x_{1-2}\phi_A x\phi_B$. However, the rules are much stricter for sequences that are compatible with a given MAPK-docking surface in a given binding mode. Because the CD groove region of ERK and p38 is wider compared to that of JNK, the N-termini of motifs binding to the two former kinases have larger conformational freedom (Figure 5.1b) (Garai *et al.* 2012). These can be classified as MEF2A- and DCC-type motifs named after the proteins in which they were first identified. Some motifs with longer intervening regions also exist (HePTP) (Zhou *et al.* 2006). The typical helical conformation at the N-terminus of HePTP-type docking motif is also characteristic to some MAPK interactors from yeast (Ste7) and peptides



(a) Different binding modes of D-motifs.

(b) Distinct binding conformations at the CD groove, ranging from fully linear (e.g. MEF2A, green) to highly alpha-helical (e.g. HePTP, magenta).

(c) Structural heterogeneity of D-motifs. The combinations of the three variable features yield structurally well-defined, distinct classes of D-motifs. JNK kinases only admit two major types of motifs, the NFAT4 class (1-spacing, short linker) and the JIP1 class (2-spacing, short linker). Known ERK1/ERK2 and p38 binder peptides may belong to the greater MEF2A class (1-spacing, longer linker, linear end), the greater HePTP class (1-spacing, longer linker, helical end), or the greater DCC class (2-spacing, longer linker, linear end). Subtypes and other variants within a given greater class based on structures of MAPK-docking motif complexes are also featured wherever applicable. Consensus motif of each subtype is shown below the corresponding three-dimensional structures representing the bound complexes.

Figure 5.1: Structural classification of MAPK-docking motifs. Dashed lines indicate N-terminal peptide regions that are usually not visible in the crystal structures.

with such motifs are known to bind human ERK2 with high affinity (Fernandes *et al.* 2007). Therefore, we also set up a hypothetical subclass of Ste7-type motifs, hitherto unknown in humans (Figure 5.1c). Interestingly, D-motifs and their binding modes may be conserved from yeast to human as the docking surface is ancient and well conserved across all eukaryotes (Reményi *et al.* 2005; Grewal *et al.* 2006).

70 known, as well as newly discovered, constructs were tested in a dot-blot assay for phosphorylation enhancement (see Section 5.2), and a total of 52 of these constructs were found to interact with at least one of the three MAPKs (ERK2, JNK1, or p38α). This

included novel interactors based on the JIP1, NFAT4, MEF2A, MKK6, DCC, and Ste7 models. Classes JIP1, NFAT4, as well as merged MEF2A and MKK6 classes (further referred to as *greater MEF2A*) were considered to have enough members to use them to construct *in silico* prediction models for other interactors in these classes and were analysed in this study.

Another part of this study is on the role of the phosphorylation in human cells signalling networks. As described in Section 2.3.1, phosphorylation can have a wide range of diverse effects on the protein, including affecting protein–protein interactions through their presence on the binding interfaces of the proteins. Also, if multiple phosphorylation sites are found on a protein, they tend to be clustered in close proximity, thus such clusters are potentially particularly indicative of having direct role in affecting protein–protein binding. The goal was thus to analyse pairs of known [ST]P phosphorylation sites in human proteins in order to see whether there are distinct groups within them which would be potentially indicative of common downstream binding partners that bind selectively to the phosphorylation sites in those groups.

5.2 Methods

5.2.1 Experimental Screening of D-motifs

A dot-blot assay based on substrate phosphorylation enhancement on a solid-phase support was used to test the binding of D-motifs to MAPKs. An artificial substrate was constructed containing the D-motifs as well as the Thr71 phosphorylation site from activating transcription factor 2 (AFT2), which is a well-known MAPK target site (Livingstone *et al.* 1995). After applying activated MAPKs on the assay and later washing the assay, it was developed using western blot techniques using an anti-phospho-T71 AFT2 antibody. Phosphorylation signal was then read either by luminescence or fluorescence.

5.2.2 Selecting Ortholog Sequences and Weights Assignment

To increase the sequence space coverage in the computational prediction models of D-motifs, more than just (known or novel) human MAPK-docking motifs were included. A method was devised to use evolutionarily weighted sequences for each independently evolved (or sufficiently unique) motif and to collect all known vertebrate orthologs. For this purpose, alignments were built from vertebrate proteins obtained by BLAST (Altschul *et al.* 1990) searches. Based on the refined consensus, motifs were classified as either potentially functional or non-functional. The motifs deemed potentially functional were realigned (with no gaps allowed) to the original sequence. If multiple paralog instances were present, they were considered within the single ortholog group. Weights were assigned to all ortholog groups using the Gerstein-Sonnhammer-Chothia scheme (Gerstein *et al.* 1994) based on pairwise sequence identity. This scheme assigns lower weight to sequences with higher identity and distant sequences receive higher weights. The purpose of these weights is that groups of closely related sequences can make a final contribution comparable to a contribution of a single distant sequence, accounting for over-representation of certain too close strains in the dataset. Finally, weights of peptides in each ortholog group are divided by the total number of peptides in the group.

Position-Specific Scoring Matrix

A Position-Specific Score Matrix (PSSM) was built for each of the JIP1, NFAT4, greater MEF2A, and greater DCC classes using formerly known and newly found, validated human motif instances as well as all their identifiable vertebrate orthologs, including 10 residue long flanking positions from the corresponding protein. In a PSSM, each row represents one of 20 possible residues, and each column represents a position in a motif. Thus, the score for residue X at position i is defined in the following way:

$$X_i = \frac{\sum_s (w_s I(s_i = X)) + pX_b}{\sum_s w_s + p}, \tag{5.1}$$

where s is a peptide sequence, w_s is the weight of that sequence based on the species from which it stems, I is the indicator function, which evaluates as 1 when its argument is true and 0 otherwise, p is the pseudo-count defined as the square root of total number of training peptides from the class and is used to account for potential other residues that are not observed at position i, and X_b is the overall background frequency of that residue (based on UniProtKB/Swiss-Prot (The UniProt Consortium 2016) Release 2013.05). For computational efficiency and to account for background frequencies of residues, log-odds scores of X_i were used in the following form:

$$X_i{}' = \log_{10}\left(\frac{X_i}{X_b}\right). \tag{5.2}$$

The score of a peptide s is then calculated as the sum of the $X_i{}'$ scores of individual positions i. If a peptide is missing any of the flanking positions, the average score at that position of the whole dataset which is being scored is used instead.

PSSMs can be visualized by so-called sequence logos. In the sequence logos the height of residue X at position i is directly proportional to its PSSM value X_i (with $p = 0$) and the information content of that position R_i, which in turn depends on uncertainty at that position H_i:

$$\begin{aligned} R_i &= \log_2(20) - H_i, \\ H_i &= \sum X_i \log_2(X_i). \end{aligned} \tag{5.3}$$

One Class Support Vector Machine

Support Vector Machine (SVM) models using string kernels were constructed using the R package KERNLAB (Karatzoglou *et al.* 2004) with the SHOGUN (Sonnenburg 2017) package for computing the string kernels.During the training phase, for each string kernel with degrees $k \in \{1,2,3,4\}$, the number of support vectors was varied with the parameter $\nu \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$. SVM models using the Radial Basis Function (RBF) kernel were constructed using the R package E1071 (Meyer *et al.* 2018). The physical-chemical encoding (Venkatarajan and Braun 2001) with five parameters per residue was used for this kernel to numerically encode peptides. Kernel spread was controlled with the γ parameter set to $\gamma = \frac{1}{df}$, with d being the data dimension, which equals the length of the peptide in its numerical encoding, and $f \in \{0.05, 0.1, 0.2, 0.5, 1, 1.5, 2, 5, 10, 20\}$ (with fixed $\nu = 0.2$) or alternatively $\nu \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$ (with fixed $\gamma = \frac{1}{d}$). During the testing phase, new data points were classified as belonging or not to the one class on which the model was trained on, depending on which side of the hyperplane they fall.

Hidden Markov Model

A Hidden Markov Model (HMM) method trained with the Viterbi algorithm, called D-finder, was used for D-motifs scoring (Whisenant *et al.* 2010). The training parameters of the learning rate (0.01) and the number of epochs in training (300) were left unmodified from the original implementation. In the original version of the algorithm, the sequence scanning using a basic hydrophobic site motif defined as $\phi x \phi$, where $\phi \in \{V, I, L, M\}$ and $x \notin \{V, I, L, M\}$, was used to down-score sequences that did not contain such motifs. This part of the algorithm was not used in our scoring of D-motifs.

5.2.3 Motif Scan and Filtering

Candidate MAPK-binding D-motif instances were collected from the human proteome. Protein sequences were downloaded from the reviewed section of the UniProt database (The UniProt Consortium 2016). The resulting Human Proteome Database (HPD) contained 20248 sequences. The HPD was scanned for motif hits with basic pattern matching using the regular expressions of 7 different D-motif classes/types:

1. JIP1-class: [RK]P[^P][^P]Lx[LIMVF],
2. NFAT4-class: [RK][^P][^P][LIM]xLx[LIMVF],
3. MEF2A-class-MEF2A-type: [RK][RK]xx[LIVMP]x[LIV]x[LIVMFP],
4. MEF2A-class-MKK6-type: [RK]x[RK]xxx[LIVMP]x[LIV]x[LIVMFP],
5. DCC-class: [RK]x_{2-4}[LIVP]xx[LIV]x[LIVMPF],
6. HePTP-class-Ste7-type: [LIV]x[RK][RK]x_4[LIVMP]x[LIV]x[LIVMFP],
7. HePTP-class-HePTP-type: [LIV][^P][^P][RK][RK]Gx_4[LIVMP]x[LIV]x[LIVMFP].

This yielded 87857 hits. Aiming to select the biologically relevant instances, the resulting hits were then filtered as outlined below. Separately, the greater MEF2A class (uniting the two MEF2A classes) was considered with a merged motif [RK]x_{2-4}[LIVMP]x[LIV]x[LIVMFP].

Structural Disorder

The estimation of the interaction potential of the selected protein regions was done with the ANCHOR algorithm, a method trained to recognize binding regions in disordered protein segments (Mészáros *et al.* 2009). In linear motif selection, a more permissive version of ANCHOR can be used; therefore, the default 0.5 cutoff value was lowered in an adaptive way so that at least 90% of the 47 formerly known D-motifs are retained. Motif hits were kept only if they overlapped with either region predicted by ANCHOR with a cutoff of 0.4, or region predicted by ANCHOR using a cutoff of 0.3, but at least one of the 20 residue flanking regions of the motif hit had to have a sufficiently high average disorder value (> 0.45) predicted with IUPred (Dosztanyi *et al.* 2005). As a result, the number of hits was reduced to 21201.

Intracellular Accessibility

Motif hits were discarded if they resided in proteins that were predicted to have a signal peptide by SignalP (Petersen *et al.* 2011) (score > 0.5), and if they were also predicted not to have a transmembrane region predicted by Phobius (Käll *et al.* 2007). These motif hits were predicted to be localized outside of the cell, which is incompatible with MAPK binding. Alternatively, Phobius was also used to predict signal peptides presence if SignalP score was sufficiently high (> 0.3). If a motif instance resided in a protein that was predicted to have a signal peptide but it also had at least one transmembrane region, the localization of the motif region was further checked. If it was entirely intracellular, it was kept, otherwise discarded. This filtering step reduced the number of motif hits to 18952.

Cellular Localization

All hits that were predicted by WoLF PSORT (Horton *et al.* 2007) to be extracellular (with score $\geq$ 25), membrane protein ($\geq$ 25), localized to the endoplasmic reticulum ($\geq$ 15), or the Golgi ($\geq$ 9) were filtered out, unless they harboured transmembrane regions, and the region containing the motif was predicted to be localized in the intracellular space. There were 18637 hits remaining after this step.

Structural Accessibility

Motif hits that were determined to reside in Pfam domain (Finn *et al.* 2014) regions were discarded. Some hits were also discarded in a manual curation process if they were located in Pfam Family/Repeat/Motif regions likely to have a stable structure in isolation. Furthermore, motif occurrences that overlapped with coiled-coil regions predicted by COILS (Lupas *et al.* 1991) were removed as well. Finally, there were 14062 motifs remaining for further analysis, including more than 90% of the known positives. Of these, 298, 382, and 4875 candidate motifs belonged to the classes JIP1, NFAT4, and greater MEF2A, respectively.

5.2.4 Evaluation of Methods Performance

To evaluate the performance of the methods for classifying the peptides from the different D-motifs classes as MAPK binders, an artificial negative dataset was created for each of the JIP1, NFAT4, and greater MEF2A classes. For this purpose, the human proteome was scanned for motif hits with basic pattern matching for each of the motifs classes and the hits which were lying in the structured region as defined by Pfam section A domains were retained, as these motifs were expected not to bind to MAPK. To be able to perform a comparison using the same set of negatives, the motifs which had incomplete flank regions due to being close to protein terminus were removed, resulting in a total of 587, 1995, and 1901 motifs for classes JIP1, NFAT4, and greater MEF2A, respectively. This last criterion was also applied for the positives dataset, resulting in 947, 381, and 446 motifs for the same classes.

For each of the motifs classes a five-fold cross-validation setting was applied. During this procedure, each set was split in five approximately equally-sized parts in terms of number of evolutionarily independent D-motif instances. Four parts were then used for training and the one remaining part, together with all the motifs from the simulated negatives for that class, was used for testing, while not allowing motifs from the same vertebrate ortholog set to be used for both training and testing. This procedure was repeated 100 times, with the exception of greater MEF2A class, where only 91 combinations of training and test dataset were possible for the five-fold cross validation procedure. AUC, the area under the ROC curve, was calculated using the R package ROCR (Sing *et al.* 2005). The ROC curves representing False Positive Rate (FPR) and True Positive Rate (TPR) values were constructed for thresholds corresponding to the following three values: i) distance of each motif to the separating hyperplane for the

SVM models, ii) each motif's score for the PSSM, and iii) the Viterbi score for each motif for the HMM models.

5.2.5 Selection and Clustering of Phosphorylation Sites

Candidate human phosphorylation sites with the [ST]P consensus have been collected from the PhosphoSitePlus database (Hornbeck *et al.* 2015). Sites where two phosphorylation residues follow each other with a maximum distance of five residues between them (counting from the [ST] site) have been grouped based on that distance into five separate groups. Each group thus consisted of peptides with two phosphorylation sites with distances $d = 1, \ldots, 5$ between them and their flanks on both sides, which were of length $min(7, 3 + d)$. Alignments of proteins embedding these peptides together with their vertebrate homologs, as reported in the eggNOG database (Jensen *et al.* 2008), were then analysed. Only evolutionarily conserved peptides, namely those whose double phosphorylation sites were conserved in at least one of different species of fish (*Latimeria chalumnae, Tetraodon nigroviridis, Ictalurus punctatus, Cyprinus carpio, Salmo salar, Oryzias latipes, Takifugu rubripes, Gadus morhua, Osmerus mordax, Oncorhynchus mykiss, Dicentrarchus labrax, Anoplopoma fimbria, Oreochromis niloticus, Danio rerio*) were kept for further analysis.

Each group was then analysed for similar peptides within it. Specifically, peptides from the same gene as well as peptides with similar sequences $\Bigg(\frac{D_{edit_{i,j}}}{|i|} < 0.4$, where $D_{edit_{i,j}}$ is the edit distance between peptides i and $j\Bigg)$ which came from the same Pfam A domain or the same Uniprot entry, have been removed and represented instead by a consensus sequence of those peptides for the clustering. For all $d = 1, \ldots, 5$ between the two phosphorylation sites this resulted in a total of 197, 248, 361, 182, and 175 peptides, respectively.

For the clustering procedure, the distance between peptides i and j was defined as follows:

$$D_{i,j} = \sum_{k=1}^{|i|} d_{i_k,j_k} w_k, \tag{5.4}$$

where d is the Euclidean distance between the corresponding entries in the physical-chemical properties matrix (Venkatarajan and Braun 2001) and w_k is the position k dependent weight. w_k is defined as $max(1 - \frac{1}{10}dist_{[ST]_1}, 1 - \frac{1}{10}dist_{[ST]_2})$, where $dist_{[ST]_{\{1,2\}}}$ is distance to the first and second phosphorylation site, respectively. This weighting is used to linearly down-weight the importance of the position as one moves away from the phosphorylation sites.

The clustering itself was performed using the tight clustering algorithm (Tseng and Wong 2005) using the R package TightClust (Tseng and Wong 2012). Briefly, tight clustering is an iterative procedure for performing k-means clustering with a pre-set k with resampling, with the goal of finding clusters that are most tight (data points within it tend to cluster together in separate clustering iterations) and stable (tight cluster candidate which is chosen repetitively for increasing k starting with k_0). The core idea of the method is resampling of data points: at each iteration a pre-set portion of all data points are randomly left out of the clustering. When the data space tends to be clustered in a similar way in different such iterations, it is expected to have some structure on top of random noise. Note that in this approach not all data points get assigned to a specific cluster in the end, but only those for which this structure is observed.

Some of the important parameters of the procedure are thus $\alpha \in [0, 1]$ (where values closer to 0 select tighter clusters), $\beta \in [0, 1]$ (where values closer to 1 select more stable clusters), the resampling iteration number B, and the resampling rate p. Authors of the tight clustering algorithm identify $k_0 \in [1, \infty]$ (which is related to expected true number of clusters, with large k_0 values tending to select small tight clusters) as the most important parameter of the procedure and suggest setting $k_0 > l + 5$, where l is the expected number of clusters (Tseng and Wong 2012). For the remaining parameters, authors suggest default values of $\alpha = 0.1$, $\beta = 0.6$, $B = 10$, and $p = 0.7$. When testing performance of clustering for a specific parameter, the rest of the parameters were set to default values, with the exception of $B = 20$. Tight clustering was performed for each group of peptides for cluster sizes $l \in [2, 15]$.

The quality of clustering was measured using average of silhouette value $s(i)$ of all clustered data points i using R package CLUSTER (Maechler *et al.* 2017):

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{5.5}$$

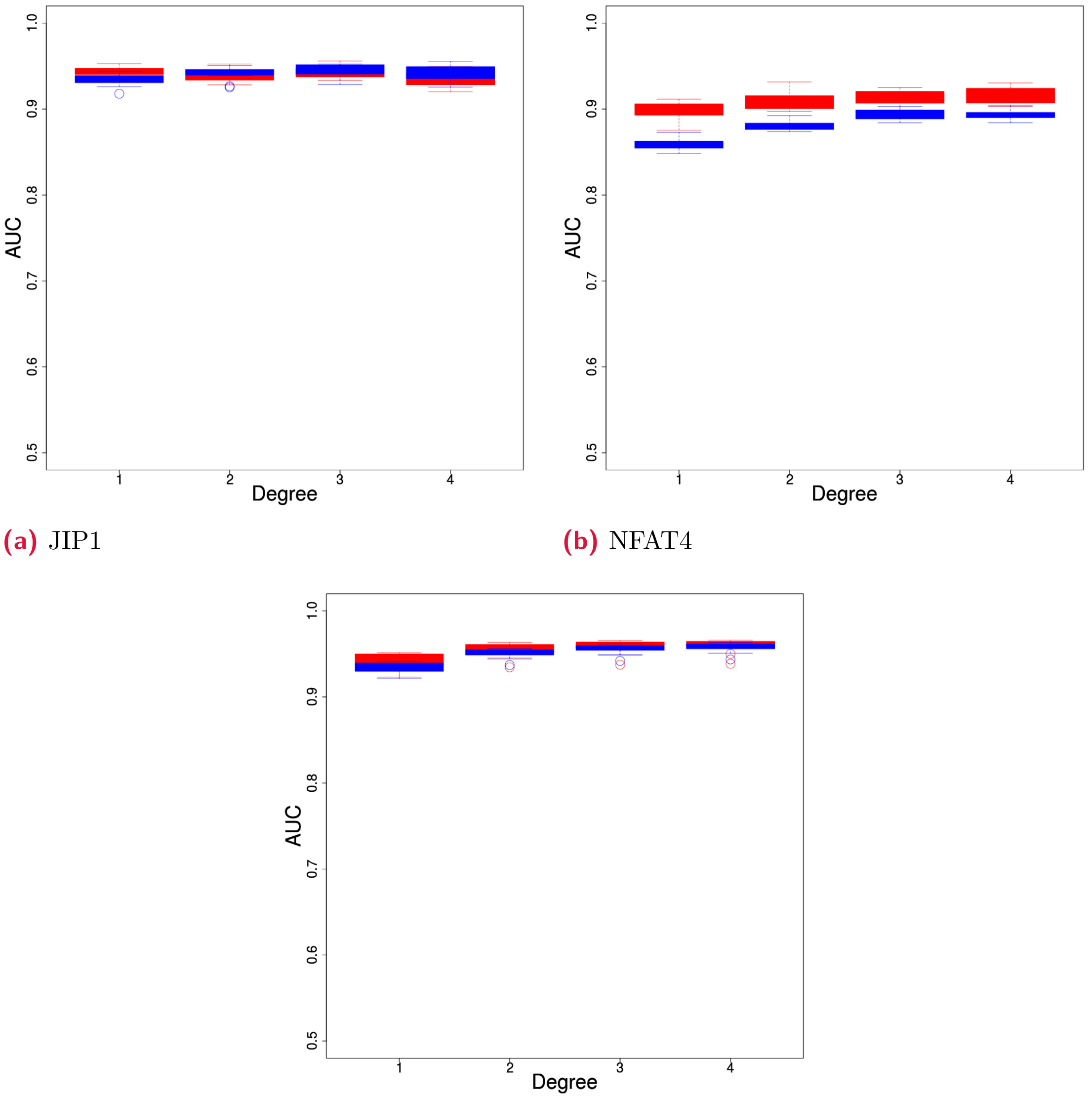
where $a(i)$ is the average dissimilarity between i and all the other data points in the same cluster, and $b(i)$ is the smallest average dissimilarity of i to the points in other clusters. A silhouette value close to 1 thus indicates presence of a well-ordered structure of the clusters and -1 indicates a lack thereof. When deciding on which number of clusters to choose, one typically chooses a number y which gives a local peak of silhouette value, such that $y + 1$ has a relatively much lower average silhouette value.

5.3 Results

5.3.1 Prediction Models for D-motif Classes

We employed two techniques to learn from each experimentally verified class of MAPK binders and their homologs: SVM and PSSM. The same procedure was applied for testing the performance of those techniques in predicting D-motifs. Namely, 5-fold cross-validation was performed, selecting 4 folds of D-motifs from that class for training and 1 remaining fold, together with the peptides from Pfam A domains, fitting the D-motif definition for that class, for testing (see Section 5.2.4).

One-class (also called novelty-detection) SVM (Schölkopf *et al.* 2000) is a kind of SVM classifier which uses data from only a single class in its training procedure, during which it aims to separate that data from the origin with a maximum margin. We built classifier models for each of peptide classes JIP1, NFAT4, and greater MEF2A. The purpose of these models is to predict whether an observed test instance of a D-motif peptide is a member of the class of D-motifs used to train it or not. In the testing phase, peptides from the founding class are classified together with simulated negatives. We used two kinds of kernels for comparison of D-motifs: string kernel, which can take D-motifs directly as input, and RBF kernel, which requires numerical input. Weighted Degree (WD) (Rätsch and Sonnenburg 2004) and Weighted Degree with Shifts (WDS) (Rätsch *et al.* 2005) string kernels were employed, which estimate string similarity based on the number of matching k-mers between them at fixed (WD) or variable (WDS) starting positions. The expectation was that the latter, which allows for shifts in positions when comparing k-mers, would show an improved performance when compared to the former in the case of greater MEF2A class of D-motifs which has a variable length linker region between the charged and hydrophobic sites. The results of performance of the models in terms of AUC for both kinds of string kernels using k-mers of up to length 4 are presented in Figure 5.2.



(a) JIP1

(b) NFAT4

(c) Greater MEF2A

Figure 5.2: Performance of SVM classification of D-motifs of JIP1, NFAT4, and greater MEF2A class in terms of AUC using WD (red) and WDS (blue) kernels of different degrees.

Comparing WD and WDS kernels, JIP1 and greater MEF2A classes had comparable performance, while the NFAT4 class models with the WD kernel performed better. This indicated WD kernel as the best overall choice. With respect to the kernel degree, increasing it improved the performance for classes NFAT4 and greater MEF2A, while for JIP1 class there was no consistent effect on the performance. Overall all models for classes JIP1 and greater MEF2A performed well in terms of AUC irrespective of

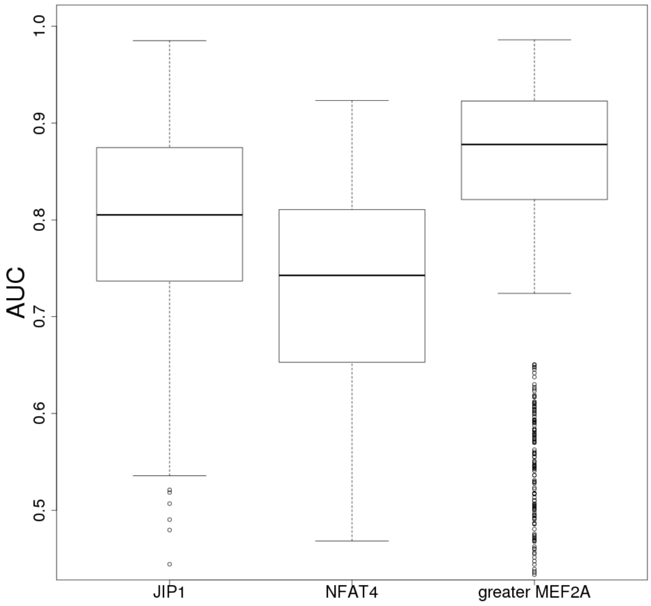


Figure 5.3: Performance of SVM models using RBF kernel with physical-chemical residues encoding in terms of AUC with $\gamma = \frac{1}{d}$.

the choice of the type of the string kernel and their degree, and NFAT4 models performed slightly worse, with the mean AUC values for WD kernels 0.94, 0.95, and 0.91 respectively.

For the one-class SVM models with RBF kernel, we converted peptide sequences to quantitative descriptors as described by (Venkatarajan and Braun 2001). In this five-dimensional property space, residues that are similar in physical-chemical properties such as hydrophobicity, size, tendency to occur in α-helices and others, will be more closely located, whereas residues with large discrepancies in these properties will tend to have larger distance between them. Usage of such kernel can be seen as a logical extension of the definitions of D-motifs classes, which are composed of hydrophobic and charged parts. Figure 5.3 shows the performance of the models built with parameter γ, which controls the spread of the kernel, set at $1/d$, where d is the data dimension. The average performance for classes JIP1, NFAT4, and greater MEF2A at 0.8, 0.73, and 0.84, respectively, was lower than for SVM models with string kernels and also displayed higher variance. This variance could be attributed neither to the parameter ν, which sets the upper bound on the fraction of class outliers and the lower bound on the fraction of support vectors (Schölkopf *et al.* 2000), (Figure B.11), nor to the parameter γ (Figure B.12).

A completely different kind of approach to create a scoring function based on the peptide classes was to create a PSSM for each class. In this approach, in the training procedure the homologs could be weighted based on their sequence similarity (see Section 5.2.2). In the testing phase, we calculated AUC for different PSSM score thresholds for the peptides in the test set (Figure 5.4). With the mean AUC values for classes JIP1, NFAT4, and greater MEF2A at 0.98, 0.91, and 0.95, respectively, the performance was very similar to that of SVM with WD kernel, with a slightly better performance for the JIP1 class.

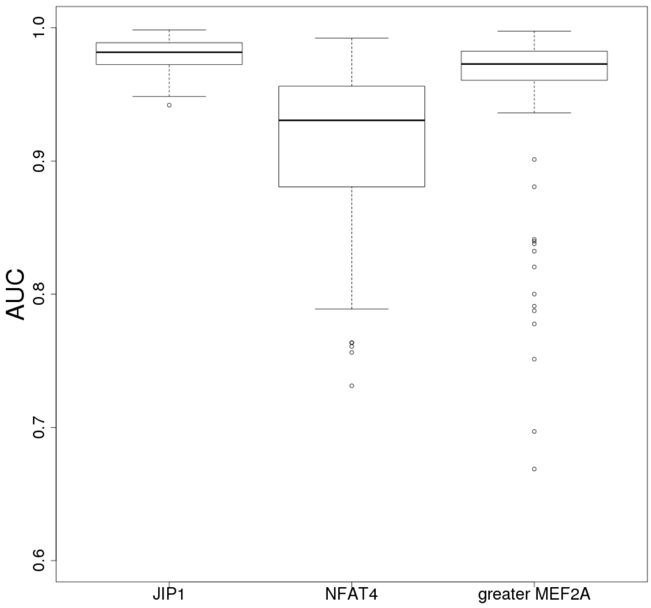


Figure 5.4: Performance of PSSM scoring in terms of AUC.

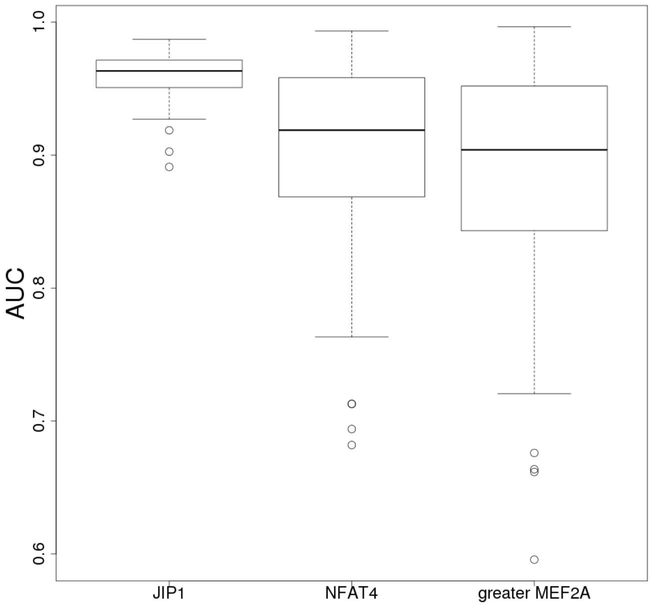


Figure 5.5: Performance of D-Finder scoring in terms of AUC.

An HMM-based predictor of D-motifs crafted for finding JNK-type binding motifs, called D-finder, has been previously reported (Whisenant *et al.* 2010). We decided to evaluate the performance of D-finder with the dataset from our study for comparison purposes. In their approach, the authors implemented a combined search for D-motif with the generic consensus, together with a scoring function based on the HMMs trained using the Viterbi algorithm. Since we already preselected the motifs corresponding to different D-motif classes, we excluded the part of the algorithm that used consensus-based motif search. In the testing phase, we calculated AUC for different score thresholds for the peptides in the test set (Figure 5.5). With the mean AUC values for classes JIP1, NFAT4, and greater MEF2A at 0.96, 0.91, and 0.89, respectively, this method displayed performance comparable to PSSM-based scoring for the NFAT4 class and slightly lower performance for JIP1 and greater MEF2A classes, while also showing greater variance for the latter.

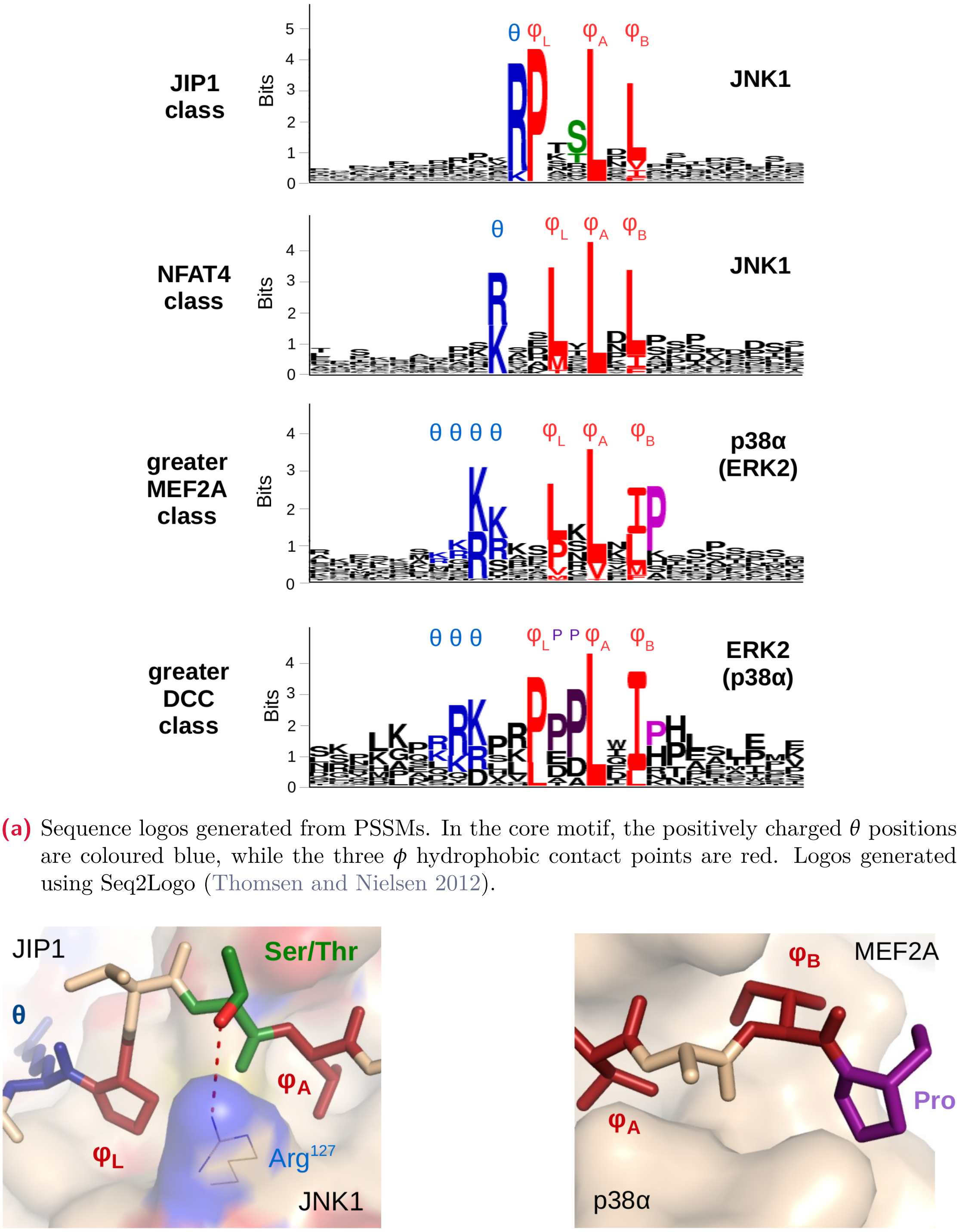
Based on these results and because of its performance, simplicity, independence of parameters, ability to include sequences of different length, and interpretability, we chose PSSM for scoring of candidate D-motifs in the human proteome.

5.3.2 D-motif Based MAPK Interactome

We now used all motifs from the three classes employed in the performance evaluation phase, as well as all D-motifs with short flanks due to being close to the N-terminus of the protein, to create a PSSM for each of the classes. We also used peptides from another D-motifs class, DCC, to create a separate PSSM. This matrix was trained using only 6 independent D-motif instances. Thus this class was not used in the testing phase, and was built only for comparison purposes with the other PSSMs. We built sequence logos based on these matrices for all classes (Figure 5.6). Due to small number of independent instances of D-motifs used to calculate its PSSM, there is less uncertainty in terms of the information content in sequence logo of DCC class compared to the others (Figure 5.6a). Overall there are few cases of conservation outside of the positions predefined by the regular expressions, which is also in agreement with the fact that the flanking regions have a small beneficial effect for the performance of the motifs classification (Figure B.13). One of the exceptions to that is serine or threonine in the position immediately preceding ϕ_A in the JIP1 class. An experimentally resolved structure of JIP1–JNK1 complex (Heo *et al.* 2004) indicates that this amino acid has the ability to form a hydrogen bond with the underlying arginine side chain of JNK1 (Figure 5.6b). Other examples are from the greater MEF2A and DCC classes, the first of which includes p38α-binding, as well as a small fraction of ERK-binding peptides, and the second containing primarily ERK2 and some of its members also associating with p38α. In these classes a proline is prefered after ϕ_B, which, based on the structure of MEF2A peptide with p38α (C.-I. Chang *et al.* 2002), can form hydrophobic interactions with the surface of p38α upon binding (Figure 5.6c).

Having the PSSMs trained on the peptide sequence preferences for all of the classes, we scanned the human proteome for the presence of motifs of same classes by using regular expressions. We then filtered the hits based on different structural and localisation criteria for those motifs as well as their host proteins (see Section Motif Scan and Filtering). The remaining 298, 382, and 4875 hits for the JIP1, NFAT4, and greater MEF2A classes respectively, were then scored with the corresponding PSSMs. For the purpose of getting an overview of the kind of biological processes motifs from these classes mediate, we annotated the top 100 hits from each class based on their UniProt labels, domain composition, and literature sources.

Out of the three classes examined, the JIP1 type had by far the highest number of validated hits. Thus, the predictions for this class were deemed most reliable, shedding some light on the interactome of JNK1 (Figure 5.7). Among the less surprising categories discovered were the MAPK pathway components themselves, especially at the MAP3K



(a) Sequence logos generated from PSSMs. In the core motif, the positively charged θ positions are coloured blue, while the three ϕ hydrophobic contact points are red. Logos generated using Seq2Logo (Thomsen and Nielsen 2012).

(b) JIP1 peptide with JNK1 complex structure.

(c) MEF2A peptide with p38α complex structure.

Figure 5.6: Sequence logos for different D-motifs classes and the interactions of JIP1 and MEF2A peptides with MAPKs.

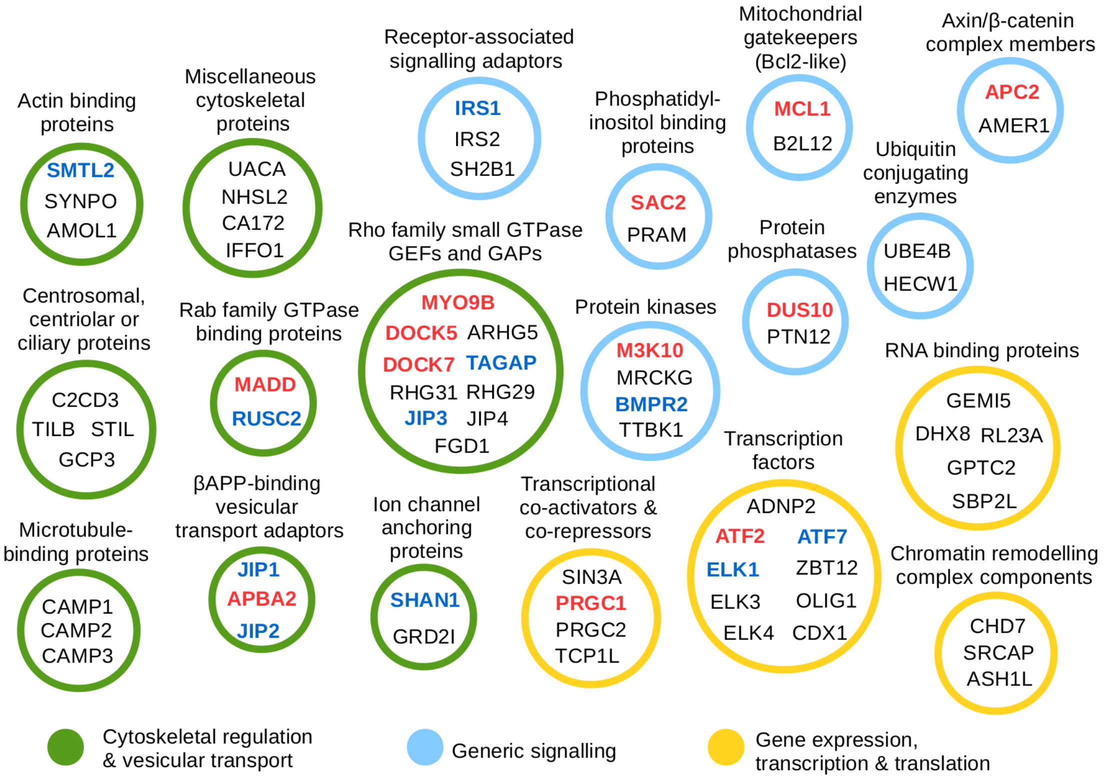
level serving as potential feedback elements. Also expected hits were several transcription factors, and other gene expression regulatory systems, or various ubiquitin ligases. A considerable number of experimentally-tested or predicted JNK-interacting proteins have preferentially or exclusively neuronal functions (Figure 5.7b). Interestingly, the majority of JNK-associating proteins (both experimentally validated and predicted) seem to be involved in cytoskeletal regulation, as numerous actin-binding or microtubule-binding proteins, molecular motors as well as small G-protein partners were encountered. Docking motifs were even found in proteins localized to centrosomes, basal bodies, or those involved in the formation of primary cilia. Several other high-scoring hits suggest that JNK is intimately involved in the regulation of endo- and exocytosis. The presence of insulin signalling pathway components in the lists may also explain many previous observations on the causative role of JNK in type II diabetes. This kinase is involved in pathways overactivated by cytokines derived from adipose tissue. JNK1 knockout mice are also known not to develop type II diabetes induced by obesity (Hirosumi *et al.* 2002; Sabio *et al.* 2009). Proteins bearing JIP1-type docking motifs (e.g. MADD, IRS1, PGC1A) are located in critical points of networks responsible for insulin signalling, and these are the same pathways that are also targeted by most antidiabetic pharmaceuticals (Y. H. Lee *et al.* 2003; Finck and Kelly 2006; Olson *et al.* 2008; D. Li *et al.* 2014).

The analysis of the best 100 hits for the NFAT4 class yielded results similar to the JIP1-type motifs, with some differences (Figure B.14a). In contrast, members of the greater MEF2A class were markedly dissimilar from those of the JIP1 class. Here, the proportion of cytoskeletal proteins was minimal, while the fraction of transcription factors was considerably higher. Proteins involved in other functions related to gene expression, such as chromatin remodelling or histone methylation, were also present in high numbers.

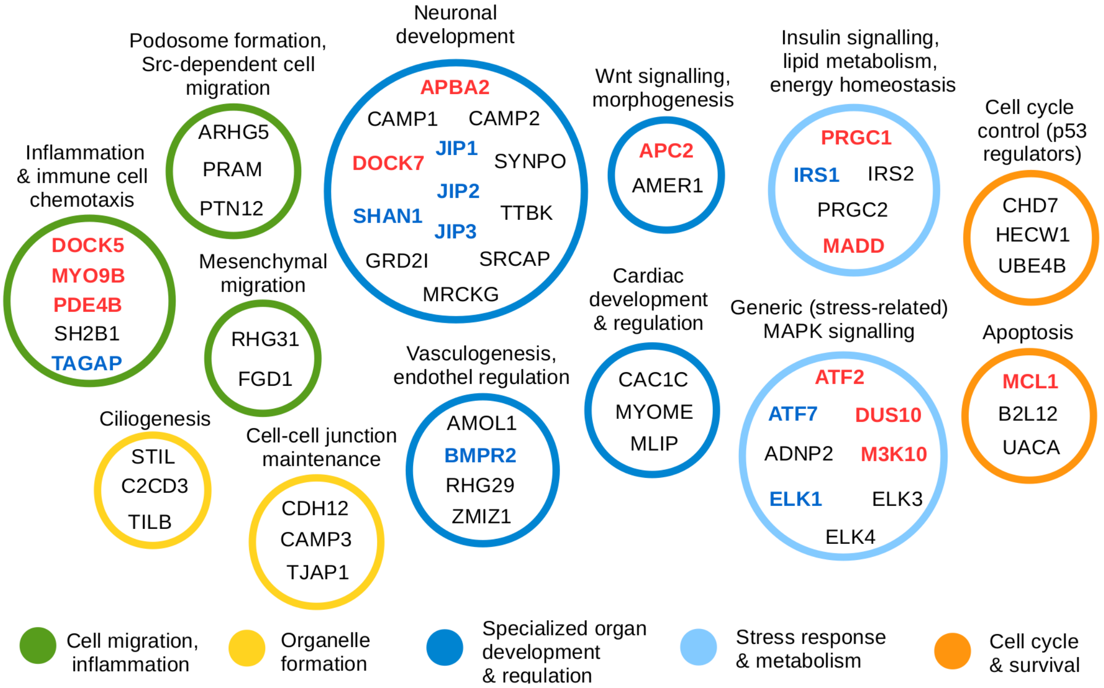
When comparing distributions of protein functions, the NFAT4 class appeared to lie between the two extremes represented by the JIP1 types (mostly cytoplasmatic targets) and greater MEF2A types (mostly acting in nucleus) (Figure B.14a). The similarity of NFAT4-type motif-containing proteins to JIP1-type bearing ones is easy to understand: both primarily interact with JNK1. In certain protein families, one can discover closely related pairs in which one protein contains a JIP1-type docking motif and the other contains a likely independently evolved NFAT4-type docking motif (Figure B.14b). On the other hand, the NFAT4-type motifs are structurally compatible with MEF2A types (unlike JIP1 types); thus, some of the predicted best binders are shared between the latter two lists. The dot-blot experiments indeed corroborated that the overlapping motif definitions result in a naturally overlapping set of interactors for JNK1 and p38α. (Figure B.14c).

5.3.3 Clustering of Phosphorylation Sites

As a separate part of the study on cellular signalling networks, we analysed known phosphorylation sites, such as those targeted by MAPK, to find potential clusters among them. For this purpose candidate phosphorylation sites have been collected from the PhosphoSitePlus database (Hornbeck *et al.* 2015). This database includes sites of human proteins which had been phosphorylated based on low- and high-throughput experimental evidence. Among these, phosphorylation sites pairs found in close proximity to each other (up to 5 residues) were selected together with their flanking regions. For this set of peptides clustering using the tight clustering approach was performed. Since one can expect to have a high number of false positives among all the candidate peptides in the dataset, attempting to cluster all peptides can potentially result in low quality clusters in terms of them being distorted and difficult to interpret, as well as making the estimation of the true number of clusters problematic. One would instead prefer to find

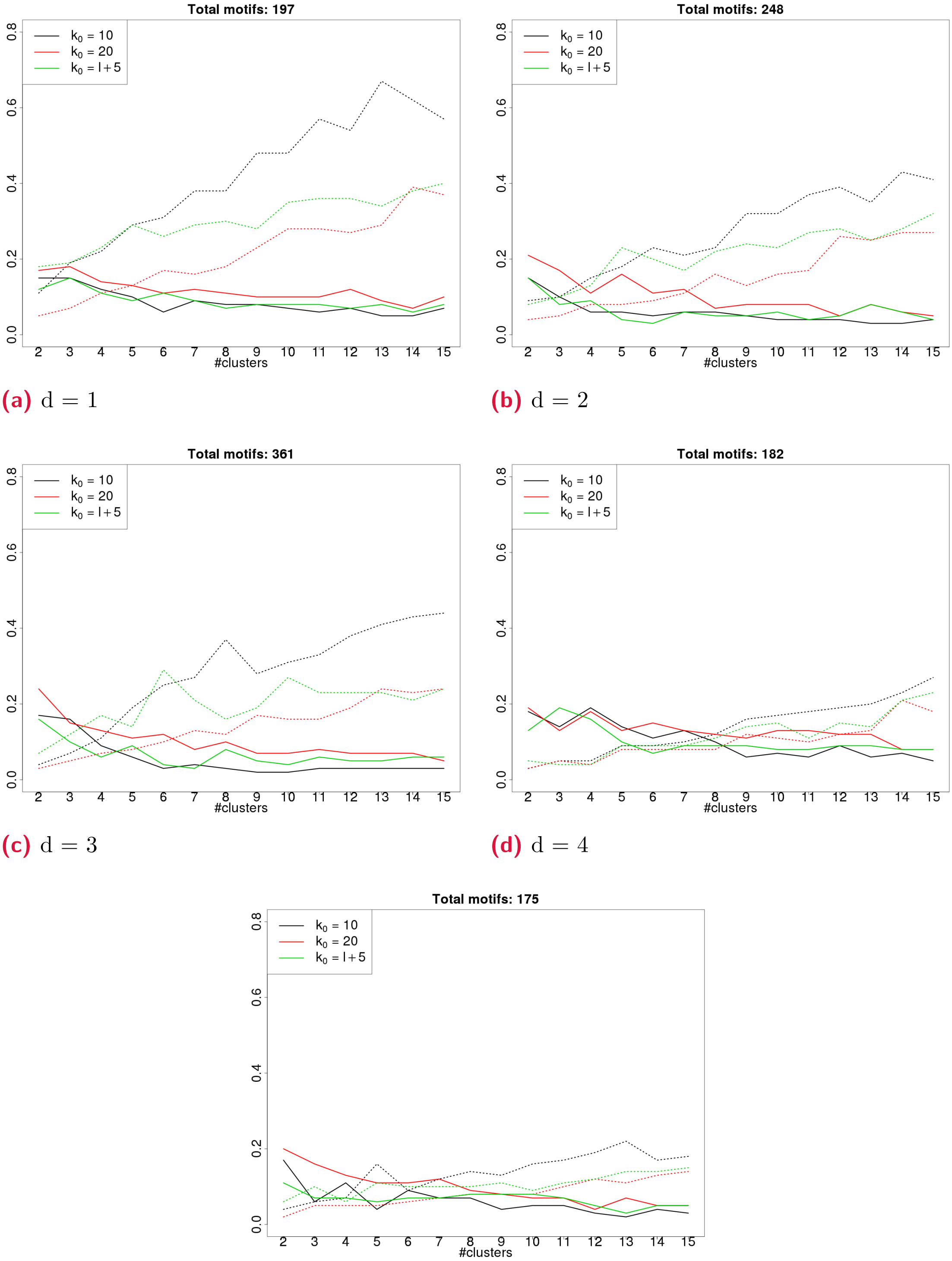

Actin binding proteins
SMTL2
SYNPO
AMOL1
Miscellaneous cytoskeletal proteins
UACA
NHSL2
CA172
IFFO1
Receptor-associated signalling adaptors
IRS1
IRS2
SH2B1
Phosphatidyl-inositol binding proteins
SAC2
PRAM
Mitochondrial gatekeepers (Bcl2-like)
MCL1
B2L12
Axin/β-catenin complex members
APC2
AMER1
Ubiquitin conjugating enzymes
UBE4B
HECW1
Rho family small GTPase GEFs and GAPs
MYO9B
DOCK5 ARHG5
DOCK7 TAGAP
RHG31 RHG29
JIP3 JIP4
FGD1
Protein phosphatases
DUS10
PTN12
Centrosomal, centriolar or ciliary proteins
C2CD3
TILB STIL
GCP3
Rab family GTPase binding proteins
MADD
RUSC2
Protein kinases
M3K10
MRCKG
BMPR2
TTBK1
RNA binding proteins
GEMI5
DHX8 RL23A
GPTC2
SBP2L
Transcription factors
ADNP2
ATF2 ATF7
ELK1 ZBT12
ELK3 OLIG1
ELK4 CDX1
βAPP-binding vesicular transport adaptors
JIP1
APBA2
JIP2
Ion channel anchoring proteins
SHAN1
GRD2I
Transcriptional co-activators & co-repressors
SIN3A
PRGC1
PRGC2
TCP1L
Microtubule-binding proteins
CAMP1
CAMP2
CAMP3
Chromatin remodelling complex components
CHD7
SRCAP
ASH1L
● Cytoskeletal regulation & vesicular transport
● Generic signalling
● Gene expression, transcription & translation

(a) Low-level functional classification of JIP1-type motif-bearing proteins.


Neuronal development
APBA2
CAMP1 CAMP2
DOCK7 JIP1 SYNPO
JIP2
SHAN1 TTBK
JIP3
GRD2I SRCAP
MRCKG
Podosome formation, Src-dependent cell migration
ARHG5
PRAM
PTN12
Wnt signalling, morphogenesis
APC2
AMER1
Insulin signalling, lipid metabolism, energy homeostasis
PRGC1
IRS1 IRS2
PRGC2
MADD
Cell cycle control (p53 regulators)
CHD7
HECW1
UBE4B
Inflammation & immune cell chemotaxis
DOCK5
MYO9B
PDE4B
SH2B1
TAGAP
Mesenchymal migration
RHG31
FGD1
Cardiac development & regulation
CAC1C
MYOME
MLIP
Generic (stress-related) MAPK signalling
ATF2
ATF7 DUS10
ADNP2 M3K10
ELK1 ELK3
ELK4
Apoptosis
MCL1
B2L12
UACA
Vasculogenesis, endothel regulation
AMOL1
BMPR2
RHG29
ZMIZ1
Ciliogenesis
STIL
C2CD3
TILB
Cell-cell junction maintenance
CDH12
CAMP3
TJAP1
● Cell migration, inflammation
● Organelle formation
● Specialized organ development & regulation
● Stress response & metabolism
● Cell cycle & survival

(b) High-level functional classification of JIP1-type motif-bearing proteins.

Figure 5.7: JNK interactome analysis using 100 highest scoring JIP1-type motifs. Only categories that contain more than a single protein are shown. Many proteins are present whose docking motif is already known (blue letters), or was validated in the dot-blot essays of the present study (red letters).



(a) d = 1

(b) d = 2

(c) d = 3

(d) d = 4

(e) d = 5

Figure 5.8: The evaluation of clustering performance of phosphorylation site groups of different distances between the [ST] sites (d) in terms of silhouette value (solid lines) and proportion of peptides assigned to clusters (dashed lines) while varying k_0 parameter.

a subset of peptides which form tight clusters and are thus potential candidate sites of being recognised by a specific downstream protein in the signalling cascade. The tight clustering approach was designed to solve this problem (Tseng and Wong 2005). This clustering approach is based on k-means clustering using resampling with the purpose of finding tight and stable clusters. There are a number of tuning parameters in this procedure (see Section 5.2.5). Although the Tseng and Wong claim the choice of the number of clusters k is secondary, k_0, which is related to the expected number of true clusters, is one of the most important parameters. Thus, we performed tight clustering on the different groups of peptides, separated based on the distance between the two phosphorylation sites while varying the k_0 parameter and the number of target clusters (Figure 5.8). The low silhouette values, which start at ≈ 0.2 for most of the classes, suggest a lack of clear groups in the data. As the proportion of peptides assigned to clusters grows, the silhouette values drop further close to zero, indicating that on average peptides are as close to other peptides in the cluster as they are to the closest neighbouring cluster. For groups of peptides with distances of 4 or 5 between the phosphorylation sites, increasing the number of target clusters increases the proportion of data points assigned only marginally.

We then performed clustering with varying other parameters of the clustering procedure, namely α, β, and p (Figure B.15). Varying β did not have much effect on performance neither in terms of silhouette value nor on the proportion of peptides assigned to clusters. On the other hand, increasing α and p values had the expected effect of increasing the fraction of peptides clustered. However, in no tested parameter combination did the silhouette value raise more than marginally above the 0.2 value. This suggests the underlying data does not lend itself to structuring using this approach and/or the adopted distance measure.

5.4 Discussion

Protein kinases often use dedicated domains for substrate recognition. Known examples include the Src-family tyrosine kinases (SH2 and SH3 domains) (Alexandropoulos and Baltimore 1996; Pellicena *et al.* 1998), SPAK/OSR kinases (Vitari *et al.* 2006) and Polo-like kinases (Polo-box domains) (K. S. Lee *et al.* 2014). In other cases, recruitment of the target proteins is provided by the catalytic domain itself, but by a distinct surface which is noncontiguous with the catalytic site. This appears to be common among the so-called cyclin-dependent kinase, MAPK, glycogen synthase kinase, CDC-like kinase group (CMGC). However, each kinase family uses a different surface with strikingly different recognition modes. Thus, motifs recognized by GSK3 or SRPKs (Dajani *et al.* 2003; J. C. K. Ngo *et al.* 2005) are unrelated to D-motifs or FxFP-motifs of MAPKs, or to CDK-docking motifs recognized by the cyclin subunit (Lowe *et al.* 2002). Based on our results on MAPK-binding D-motifs, it may be anticipated that insights into other recruitment motif-based systems will greatly contribute to a system-level understanding of protein kinase-based intracellular signalling networks.

In the current study, we demonstrate that canonical, D-motif-dependent partners of MAPKs are in fact quite common. However, a number of partners with atypical or "naturally defective" docking motifs do exist (e.g. MKK3, MKK7, TAB1), and these are difficult to predict (C.-I. Chang *et al.* 2002; De Nicola *et al.* 2013). Often such defective motifs act in a non-autonomous way: these weak elements may be complemented by additional protein stretches, motifs, or domains (Glatz *et al.* 2013). Besides, not all MAPK-binding elements are linear motifs. Folded domains such as the rhodanese domain of dual-specificity phosphatases may bind to the same site as intrinsically disordered docking motifs (Y.-Y. Zhang *et al.* 2011). MAPKs interactions can also be mediated by other linear motifs, i.e. the so-called FxFP-motifs (Jacobs *et al.* 1999; Fantz *et al.* 2001; J. Zhang *et al.* 2003). A considerable number of interactions might also be indirect, mediated by a third partner.

Nevertheless, directly interacting with a MAPK solely through short linear motifs appears to be a major and widespread phenomenon in mammals. Although experimental testing of all putative MAPK D-motifs is challenging to perform, a suggestion can be made that the fraction of the human proteome that harbours high-scoring D-motifs may be representative of the full interactomes for three distinct MAPKs. This may be best captured for JNK1 by the procedure presented in this study. Some of the newly-identified partners directly fit into the core of MAPK pathways. These include specific phosphatases as well as MAP3Ks. While there can be little doubt that docking motifs of phosphatases would be required for MAPK dephosphorylation, the presence of docking motifs in MAP3Ks is a more intriguing observation. It is probable that phosphorylation of proteins acting on the MAP3K level (like on MEKK1, MLK1/MLK2, or KSR2) would allow direct feedback control of MAPK pathways (Flotho *et al.* 2004).

However, the majority of novel hits appear to lie outside the core MAPK pathway module, and these are probably simple downstream elements (i.e. substrates). Most of the novel proteins are expected to be either direct MAPK substrates or scaffold proteins (i.e. enabling phosphorylation of indirect MAPK substrates through protein complexes).

The wide distribution of D-motifs in a functionally diverse set of proteins explains how MAPKs can regulate such a broad spectrum of physiological processes. Previously, their specific regulatory roles were often attributed to single target proteins. For example, the role of JNK in axonal growth was attributed to the JNK–JIP1 interaction, and the association of JNK with diabetes was attempted to be explained by the JNK1–IRS1 interaction alone (Y. H. Lee *et al.* 2003; Dajas-Bailador *et al.* 2008). In contrast, results from the current study imply that these interactions may only be two examples of a substantially more complex protein network, and JNK (as all MAPKs) connects to its targeted physiological systems by a large number of direct interactions. While individual connections might not be stable (especially in the evolutionary sense), multiple specific linkages could provide the key mechanism for a robust and adaptable physiological regulation.

Surprisingly, many of the newly implied MAPK partners have a restricted expression pattern enabling fine-tuned regulation in specialized tissues. Because of the latter phenomenon, a great deal of these interactions are unlikely to be discovered by large-scale protein–protein interaction screens. Easy-to-handle cell lines and mass-spectrometry-based analyses provide a powerful tool, but not for proteins that are only expressed in special, differentiated tissues (e.g. AAKG2, which is only abundant in cardiomyocytes) or restricted to certain embryonic developmental stages (e.g. DCX is almost exclusively expressed in developing neuroblasts) (Lang *et al.* 2000; Brown *et al.* 2003). Here, a modelling-driven interactome search is the most suitable tool to fill in the gaps in our knowledge. In addition, a reliable sequence-based prediction procedure sets the stage for an easy examination on how MAPK signalling partners changed over time during evolution.

In the study of the human proteome phosphorylated sites, we analysed peptides with pairs of [ST]P sites in close proximity to each other, which are targeted by proline-directed kinases (such as MAPKs). The purpose was to find groups which would be recognised by specific downstream proteins in the signalling cascade. Despite the fact that the tight clustering approach that we employed is supposed to select the part of the data that forms tight and stable clusters, we could not observe much of such structure in our set of peptides. This can potentially indicate the need to modify the hypothesis with regards to specifics of the motifs recognised by the downstream protein or the recognition mechanism. Alternatively one could narrow down the hypothesis by assuming that the recognition mechanism is used only by select proteins and to search for these groups within clusters formed. The performance of tight clustering approach itself is dependent on a number of different tuning parameters, with the number of expected clusters being one of them. Tuning these parameters is of course problematic in the context of unsupervised learning where there is no target variable. Finally, one can use a different measure for distance between the motifs, or a different density-based clustering algorithm.

6 Perspective

In this work, sequence- and structure-based approaches were applied to analyse different protein–protein and protein–inhibitor interaction systems. Despite the differing contexts of those systems, the replication of a virus and the eukaryotic signalling network, being disparate in numerous facets, both direct (Greenway *et al.* 1996; Yang and Gabuzda 1998; Yang and Gabuzda 1999; Gupta *et al.* 2011; Dochi *et al.* 2014) and indirect (Schrager *et al.* 2002; Toschi *et al.* 2006) interaction has been suggested between some of the main actors of these systems, HIV proteins and MAPKs. In fact, MAPK was the first protein kinase originating from cells which was detected within HIV particles (Cartier *et al.* 1997; Jacqué *et al.* 1998; Giroud *et al.* 2011). Similarly, despite the use of sequence- and structure-based methods being restricted in the context of this work to MAPK signalling network and the HIV protease drug resistance, respectively, studying these individual problems also benefited from both of these approaches. For example, the analysis of the MAPK interaction partners through D-motifs is largely sequence-based, but it has doubtlessly benefited from the premise of the project to use structural considerations to divide the D-motifs based on their different binding modes to MAPKs. We see potential future directions of expanding these projects both on the basis of principles of methodological approaches predominantly used in those projects so far and those from the other category. This chapter concludes the dissertation by summarizing its key developments and results and provides a perspective for future work in the area.

6.1 Conclusions

In the study of HIV protease, the major focus was a diverse set of RAMs in combination with different PIs. The goal was to see whether the experimental measurements of the resistance, both in terms of free energy of binding of the inhibitor, ΔG, and phenotypical resistance assays, RF, can be reproduced by the computational techniques based on MD simulations as well as to understand the underlying resistance mechanisms. It is first shown on a dataset of ten complexes that the change of the free energy of the inhibitor binding upon mutation, $\Delta\Delta G$, calculated using alchemical calculations correlates well with the experimental measurements. It was observed that explicit probing of alternative protonation states of the active site of the protease contributes to this accuracy, in addition to enabling us to select the most likely protonation state of the complex. It is then shown for two datasets with RF measurements where multiple proteases share the same sequence but have different inhibitor bound, totalling seventeen different complexes, that those RF measurements can be in most cases reproduced from the calculated $\Delta\Delta G$ values for the same complex. Comparison of RF and $\Delta\Delta G$ values for the same mutation also provided insights into whether the change in resistance is predominantly a result in drug binding affinity or of a change in the catalytic activity of the protease. We correlate these observations with the presence or lack of specific secondary mutations. Finally, the analysis of the MD simulations underlying the $\Delta\Delta G$

estimations reveals various effects of mutations on the protease structure and its interactions with the inhibitor. For RAMs found in the active site, the effect of mutations is mostly through changes in direct protein–inhibitor interactions; for RAMs in distant sites, protein–inhibitor interactions are also affected by mutation, but the underlying mechanisms are more complex and include displacements of specific side chains, changes in hydrogen bond network of the protease, differences in correlated movements of the residues and/or structural rearrangements of the protein.

In the study of the specificity of D-motifs in the MAPK signalling we made several important observations regarding the MAPK interactome. After observing that D-motifs acquire distinct binding modes when bound to a specific MAPK, we split known and experimentally validated human D-motifs into separate classes. Together with the related vertebrate D-motifs, peptides from each class could then be used to train different kinds of predictive models, which could successfully distinguish them from a dataset of simulated negatives. Models based on PSSMs trained on different motif classes performed best in our comparison and were used to score D-motifs in the human proteome with the following annotation of the top-scoring hits. This gave us insight into the different physiological processes, in which proteins that embed these D-motifs participate, as well as into the differences between proteins with D-motifs from different classes in terms of their cellular localization. In this study, we also analysed different kind of linear binding motifs which are phosphorylated, among other proteins, by MAPKs. We performed clustering on pairs of phosphorylated [ST]P sites in short distance to each other, aiming at finding tight clusters of motifs potentially indicative of being recognition sites of downstream proteins in the signalling cascades. Despite using a clustering approach which selectively clustered only those data points that tended to form clusters with each other, the goal of finding a cohesive structure among motifs with pairs of phosphorylation sites proved difficult to achieve.

6.2 Outlook

The current work presents a successful application of alchemical methods based on MD simulations for an accurate estimation of the effects of different HIV protease major RAMs, both inside and outside of the active site, on the enzyme resistance towards various drugs. This allowed to gain valuable insights into the resistance mechanisms of HIV protease at the molecular level. However, there is still a big gap in understanding the molecular details of the resistance of a number of major RAMs towards different inhibitors, including, e.g. I84V, which was part of the current study in predicting RF based on free energy estimates, let alone numerous minor RAMs. Understanding the underlying molecular phenomena could potentially provide important insights into the HIV drug resistance, as well as a possibility to transfer this knowledge to treatment of other viruses, and for drug design.

The current study also provided an insight into the balance between the most important effects of primary and secondary mutations, namely free energy of the inhibitor binding and catalytic activity, two phenomena that both contribute to the observed RFs. Statistical analysis of co-occurring mutations in the HIV protease, which already is being explored in connection to the analysis of resistance of this protein (Hoffman *et al.* 2003), combined with free energy and virological assay resistance studies, might reveal further cases of this interplay. On the other hand, finding mutations which, while selected on their own through, e.g. drug pressure, do not co-occur, or mutations which are resistance-inducing and sensitizing towards different drugs, such as some of those discussed in this study, might offer an advantage directly applicable for patient treatment. Overall, unlike statistical learning approaches in optimizing the drug therapy based on clinical data, which have found a way in application in clinical practice (Lengauer *et al.* 2014), statistical mechanics-based approaches on evaluating drug resistance are of course a few steps away from such immediate application. As demonstrated in the present study, the problem setting can also be very specific to each molecular complex under consideration. There is however continuous progress in both hardware, in terms of computational power, and software, which, e.g. now allows to automate the tedious and technical knowledge-demanding task of preparing the hybrid structures for alchemical simulations (Gapsys *et al.* 2014). With such advancements, one can envisage a wider application of molecular mechanics-based approaches for free energy estimations, which provides the advantage of having a solid physical basis and offering a high accuracy, even for large scale mutational studies.

There are many prospective avenues for further work on MAPK signalling network. We take part in an ongoing project to identify peptides with multiple phosphorylation sites indicative of specific biological function, with experiments being conducted by our collaborators to verify peptides from individual clusters as potentially binding to F-box/WD repeat-containing protein 7 (FBW7) as a part of MAPK signalling. The basic motif required for binding, which includes the phosphorylation sites, potentially needs refining to find the underlying pattern essential for signal transduction in the network. Alchemical methods for estimating free binding energy could be applied to supplement the experiments verifying candidate peptide binders. Specifically, closed thermodynamic cycles could be constructed to test how changes of amino acids in specific positions affect the free binding energy of that peptide to the target protein. The same approach could of course be used for the D-motifs binding to MAPKs to further refine consensus D-motifs sequences as preferred by different MAPKs. Since MAPKs are involved in various widely spread human diseases, they are natural pharmaceutical targets. Alchemical approaches could also be employed here for evaluating different candidate drugs, targeting both the active site and its other interaction sites, such as the D-motifs binding site. The questions posed in the current dissertation on the minimal required motifs and interaction mechanisms in the MAPK signalling network are undoubtedly crucial for efficient targeting of specific pathways in this network.

A
Glossaries

List of Abbreviations

EC_{50}	Effective Concentration corresponding to the half-maximal response
IC_{50}	Inhibitor Concentration required to reduce enzymatic reaction activity by 50%
AIDS	Acquired Immunodeficiency Syndrome
APV	Amprenavir
ART	Antiretroviral Therapy
ARV	Antiretroviral
ATP	Adenosine Triphosphate
ATV	Atazanavir
BAR	Bennet Acceptance Ratio
CCD	Charged-Coupled Device
CFT	Crooks Fluctuation Theorem
CRF	Circulating Recombinant Form
cryo-ET	cryo Electron Cryotomography
cryo-TEM	cryo Transmission Electron Cryomicroscopy
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleoside Triphosphate
DRV	Darunavir
FDA	U.S. Food and Drug Administration
FEP	Free Energy Perturbation
FPR	False Positive Rate
FPV	Fosamprenavir
HIV-1	Human Immunodeficiency Virus type 1
HIV-2	Human Immunodeficiency Virus type 2
HIVDR	Human Immunodeficiency Virus Drug Resistance
HMM	Hidden Markov Model
IDV	Indinavir

INI	Integrase Strand-Transfer Inhibitor
ITC	Isothermal Titration Calorimetry
LPV	Lopinavir
MC	Monte Carlo
MD	Molecular Dynamics
MI	Mutual Information
MM	Molecular Mechanics
MM/GBSA	Molecular Mechanics/Generalized Born Surface Area
MM/PBSA	Molecular Mechanics/Poisson–Boltzmann Surface Area
mRNA	Messager RNA
NFV	Nelfinavir
NMR	Nuclear Magnetic Resonance
NNRTI	Non-Nucleoside Reverse-Transcriptase Inhibitor
NRTI	Nucleotide or Nucleoside Reverse-Transcriptase Inhibitor
PB	Poisson–Boltzmann
PBMC	Peripheral Blood Mononuclear Cell
PCA	Principal Component Analysis
PDB	Protein Data Bank
PEP	Post-Exposure Prophylaxis
PI	Protease Inhibitor
PLS	Partial-Least Squares
PME	Particle-Mesh-Ewald
PrEP	Pre-Exposure Prophylaxis
PSSM	Position-Specific Score Matrix
PTM	Post-Translational Modification
QM	Quantum Mechanics
RAM	Resistance-Associated Mutation
RBF	Radial Basis Function
RESP	Restrained Electrostatic Potential
RF	Resistance Factor
RNA	Ribonucleic Acid
RTV	Ritonavir
SIV	Simian Immunodeficiency Virus
SQV	Saquinavir
SVM	Support Vector Machine
TI	Thermodynamics Integration
TPR	True Positive Rate
TPV	Tiprananavir

WD	Weighted Degree
WDS	Weighted Degree with Shifts
WHO	World Health Organization

List of Genes, Transcripts, Proteins, and Complexes

AAKG2	5'-AMP-activated protein kinase subunit gamma-2
AFT2	activating transcription factor 2
APOBEC3G	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G
CA	capsid
CCR5	C-C chemokine receptor type 5
CD4	cluster of differentiation 4
CDK	cyclin-dependant kinase
CMGC	cyclin-dependent kinase, MAPK, glycogen synthase kinase, CDC-like kinase group
CXCR4	C-X-C chemokine receptor type 4
DCC	deleted in colorectal carcinoma
DCX	neuronal migration protein doublecortin
Elk-1	ETS domain-containing protein Elk-1
env	envelope
ERK	extracellular signal-regulated kinase
ERK1	mitogen-activated protein kinase 3
ERK2	mitogen-activated protein kinase 1
ERK3	mitogen-activated protein kinase 6
ERK4	mitogen-activated protein kinase 4
ERK5	mitogen-activated protein kinase 7
ERK7	mitogen-activated protein kinase 15
FBW7	F-box/WD repeat-containing protein 7
gag	group-specific antigen
gp120	glycoprotein 120
gp41	glycoprotein 41
GSK3	glycogen synthase kinase 3
HePTP	protein tyrosine phosphatase non-receptor type 7
IN	integrase
IRS1	insulin receptor substrate 1
JIP1	c-Jun-amino-terminal kinase-interacting protein 1, also called MAPK8IP1 or JNK1 interacting protein 1
JNK	c-Jun N-terminal kinase
JNK1	mitogen-activated protein kinase 8
JNK2	mitogen-activated protein kinase 9
JNK3	mitogen-activated protein kinase 10
KSR2	kinase suppressor of ras 2
LTR	long terminal repeat

MA	matrix
MADD	MAPK-activating death domain protein
MAP2K	mitogen-activated protein kinase kinase
MAP3K	mitogen-activated protein kinase kinase kinase
MAPK	mitogen-activated protein kinase
MEF2A	myocyte-specific enhancer factor 2A
MEKK1	MAP2K kinase 1
MKK3	dual specificity mitogen-activated protein kinase kinase 3
MKK4	dual specificity mitogen-activated protein kinase kinase 4
MKK7	dual specificity mitogen-activated protein kinase kinase 7
MKP	MAPK phosphatase
MLK1	MAP2K kinase 9
MLK2	MAP2K kinase 10
NC	nucleocapsid
Nef	negative regulatory factor
NFAT4	nuclear factor of activated T-cells 4
NLK	nemo-like kinase
OSR	oxidative stress-responsive kinase
p38	p38 mitogen-activated protein kinase
p38α	mitogen-activated protein kinase 14
p53	tumor protein p53
PDK1	phosphoinositide-dependent protein kinase 1
PDZ	post synaptic density protein (PSD95), Drosophila disc large tumor suppressor (Dlg1), and zonula occludens-1 protein (zo-1)-domain
PGC1A	peroxisome proliferator-activated receptor gamma coactivator 1-alpha
pol	polymerase
PR	protease
Rev	regulator of expression of virion proteins
RT	reverse transcriptase
SH2	Src homology 2
SH3	Src homology 3
SPAK	sterile20 related proline-alanine-rich kinase
SRPK	serine/threonine-protein kinase
TAB1	MKK7-interacting protein 1
Tat	trans-activator of transcription
Vif	virion infectivity factor
Vpr	viral protein R
Vpu	viral protein U

Glossary

apo

Refers to the unbound state of the protein.

AUC

Refers to the Area Under ROC curve.

CD groove

Common Docking groove. Refers to a negatively charged region on the the MAPK surface, which, together with hydrophobic docking groove, is where the D-motifs bind.

D-motif

Docking motif. Refers to a short linear peptide sequence which facilitates the interaction between the protein, in which it is found, and MAPK by binding to the docking groove of the later.

holo

Refers to the ligand-bound state of the protein.

homolog

A gene related to a another gene by descent from a common ancestor.

HXB2

Name of a nucleatide sequence of HIV-1 group M subtype B which is comonly used as a reference wildtype sequence of HIV-1. Numbering of nucleotide and amino acid positions in this sequence can be found at https://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html

hydrophobic docking groove

Refers to a hydrophobic region on the the MAPK surface, which, together with CD groove, is the binding site for D-motifs.

NPT

Refers to isothermal-isobaric ensemble, where number of particles N, pressure P, and temperature T are constant.

ortholog

A homolog that is the result of a speciation event.

paralog

A homolog that is the result of a duplication event.

RF$_R$

Describes RF ratio between two mutant sequences (as compared to wildtype HXB2).

ROC curve

Receiver Operating Characteristic curve describes performance of a binary classifyer at different decision thresholds in terms of TPR and FPR.

B
Supplementary Material

Structure	Inhibitor	Reference protonated aspartic acid
2O4K	ATV	D25
2O4S	LPV	D25
2O4P	TPV	D25′
1HPV	APV	D25′
2BPX	IDV	D25
3NU3	APV	D25′
1SDT	IDV	D25′
1HXB	SQV	D25

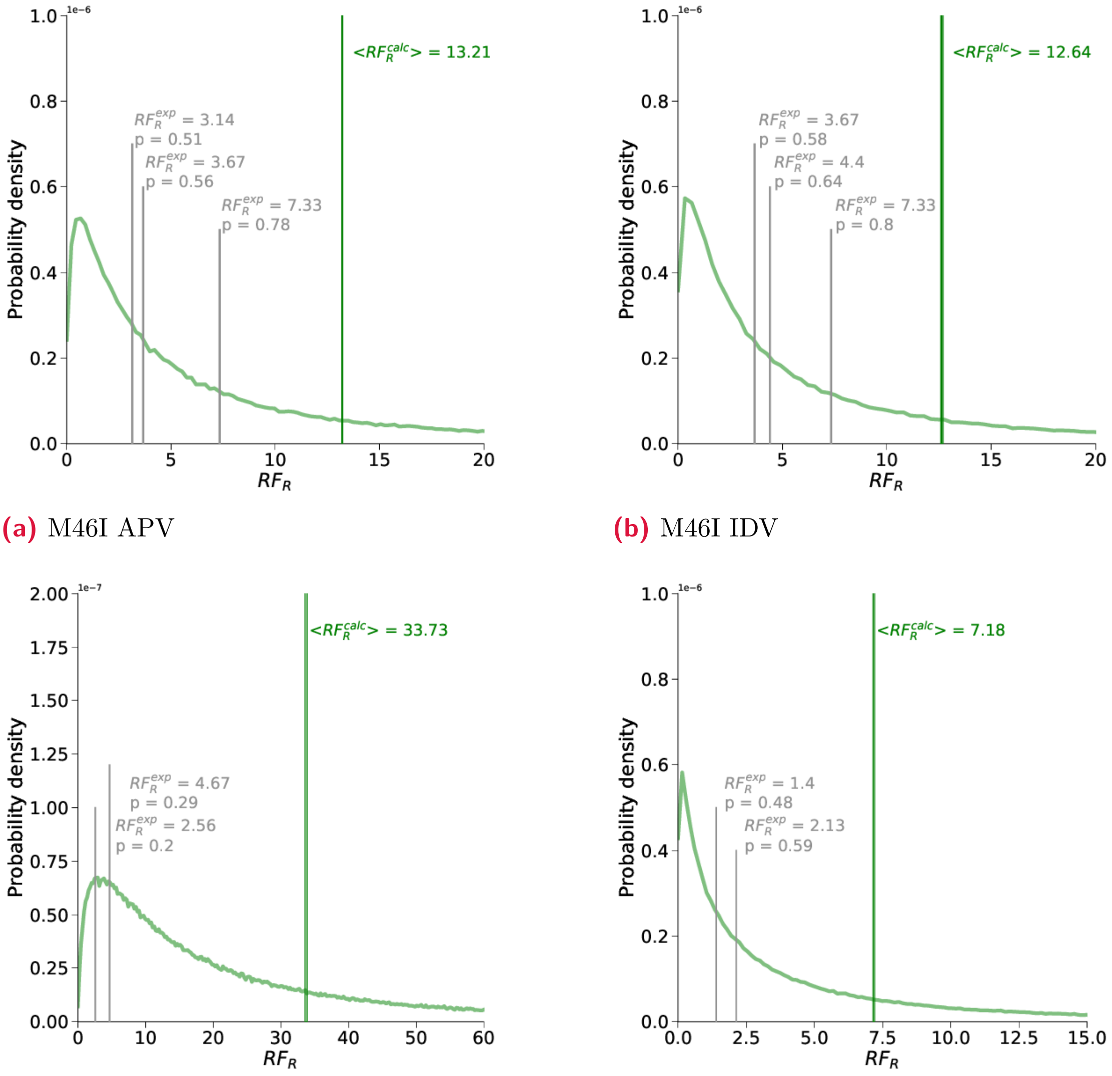
Table B.1: Selected reference protonation states for dataset 1. If alternative inhibitor orientation states A and B with the same occupancy were present in the original structure, state A was always selected, with the exception of 3NU3, where state B was selected, resulting in opposite orientations of the inhibitor APV with respect to the two monomers when comparing 1HPV and 3NU3.

Structure	Inhibitor	Reference protonated aspartic acid	$\Delta\Delta G^{prot}$
5T8H	APV	D25	1.46 ± 0.19
5E5J	DRV	D25′	2.36 ± 0.44
5E5K	DRV	D25	2.32 ± 0.36
2ZYE	KNI-272	D25	1.01 ± 0.30

Table B.2: Free energy change upon switching the proton from the reference active site aspartic acid as found in structure to the active site aspartic acid on the opposite protease monomer.

Mutation		Isolate	Reference
M46I	wildtype	71V-11, A71V-7, Bru-A71V-3	(Colonno *et al.* 2004)
	mutant	P372	(Petropoulos *et al.* 2000)
I84V	wildtype	JGP-M1C	(Prado *et al.* 2002)
	mutant	JGP-M2C, JGP-M2R	(Prado *et al.* 2002)
N88S	wildtype	RZ27 (IDV), RZ28 (FPV)	(Ziermann *et al.* 2000)
	mutant	RZ22 (IDV), RZ-L4 (FPV)	(Ziermann *et al.* 2000)

Table B.3: Isolates, for which measurements of RF were used in dataset 2 as reported in HIVdb, and references to the studies where RF measurements were performed.



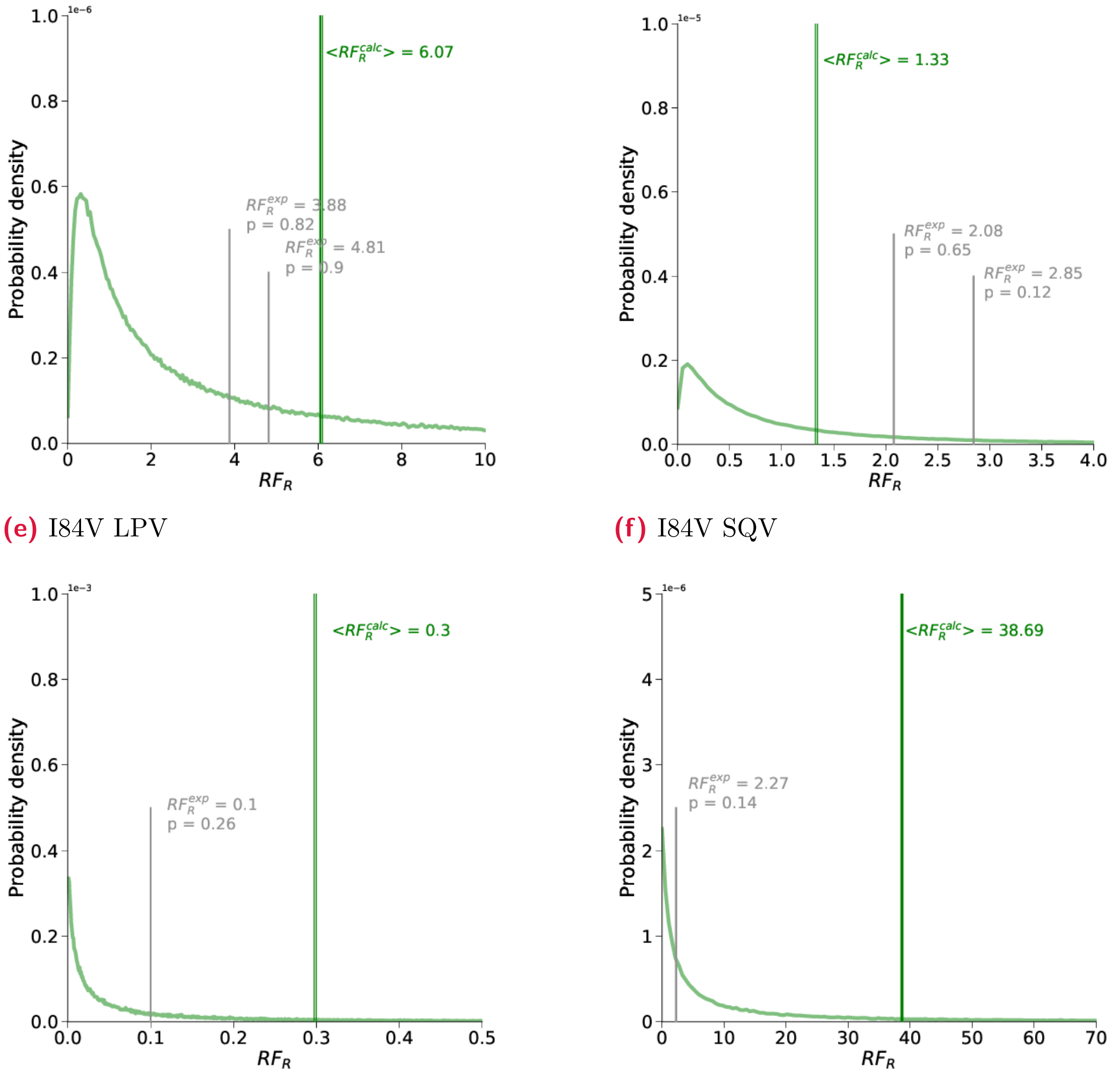
(a) M46I APV

(b) M46I IDV

(c) I84V APV

(d) M84I IDV

Figure B.1: Calculated RF$_R$ distributions and experimental estimates for dataset 2. p designates the proportion of RF_R^{calc} at least as extreme as RF_R^{exp} compared to $< RF_R^{calc} >$. *Nota bene:* in case of APV, RF_R^{exp} measurements are for its prodrug FPV.



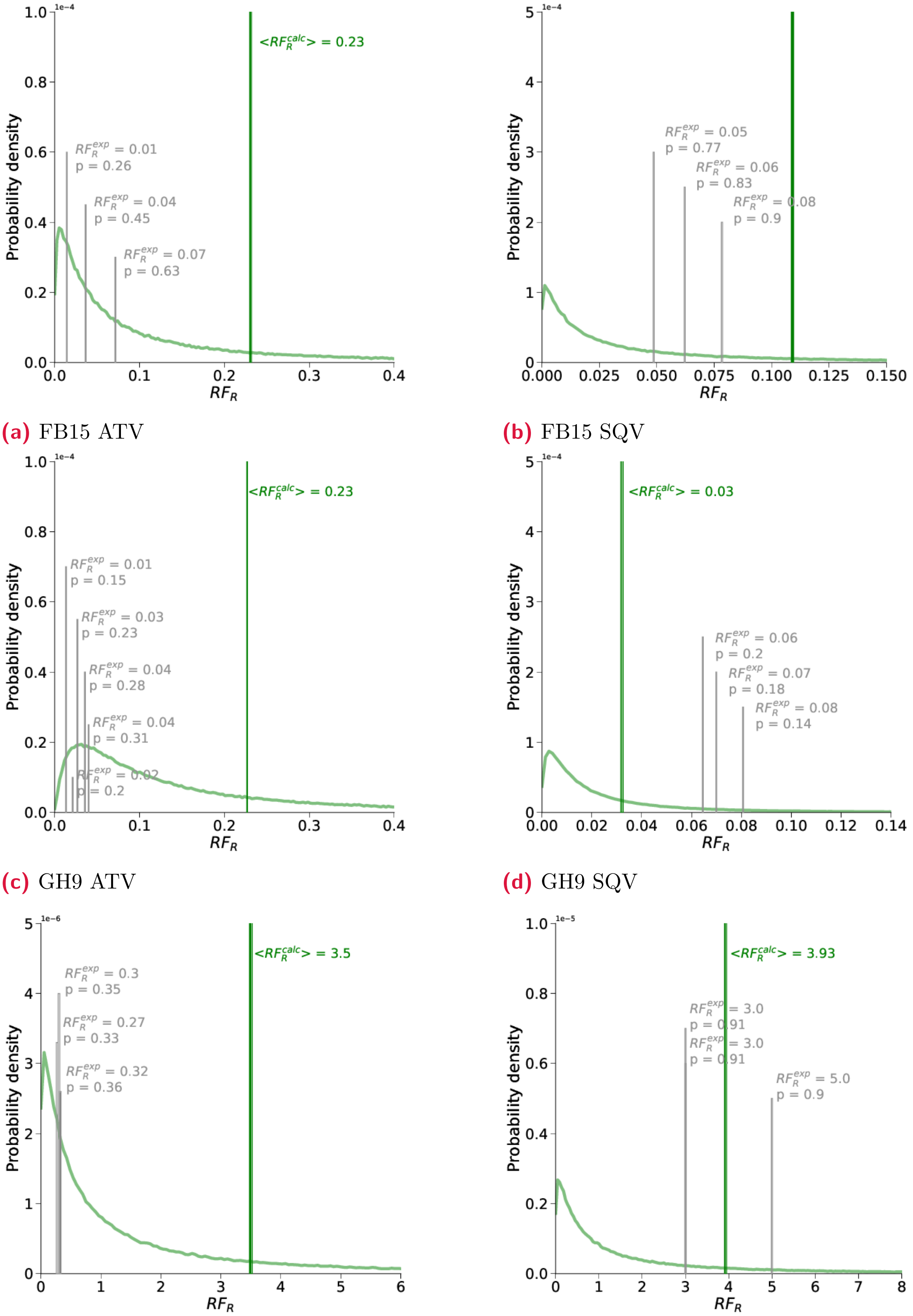
(e) I84V LPV

(f) I84V SQV

(g) N88S APV

(h) N88S IDV

Figure B.1: Calculated RF_R distributions and experimental estimates for dataset 2. p designates the proportion of RF_R^{calc} at least as extreme as RF_R^{exp} compared to $< RF_R^{calc} >$. *Nota bene:* in case of APV, RF_R^{exp} measurements are for its prodrug FPV (cont.)



(a) FB15 ATV

(b) FB15 SQV

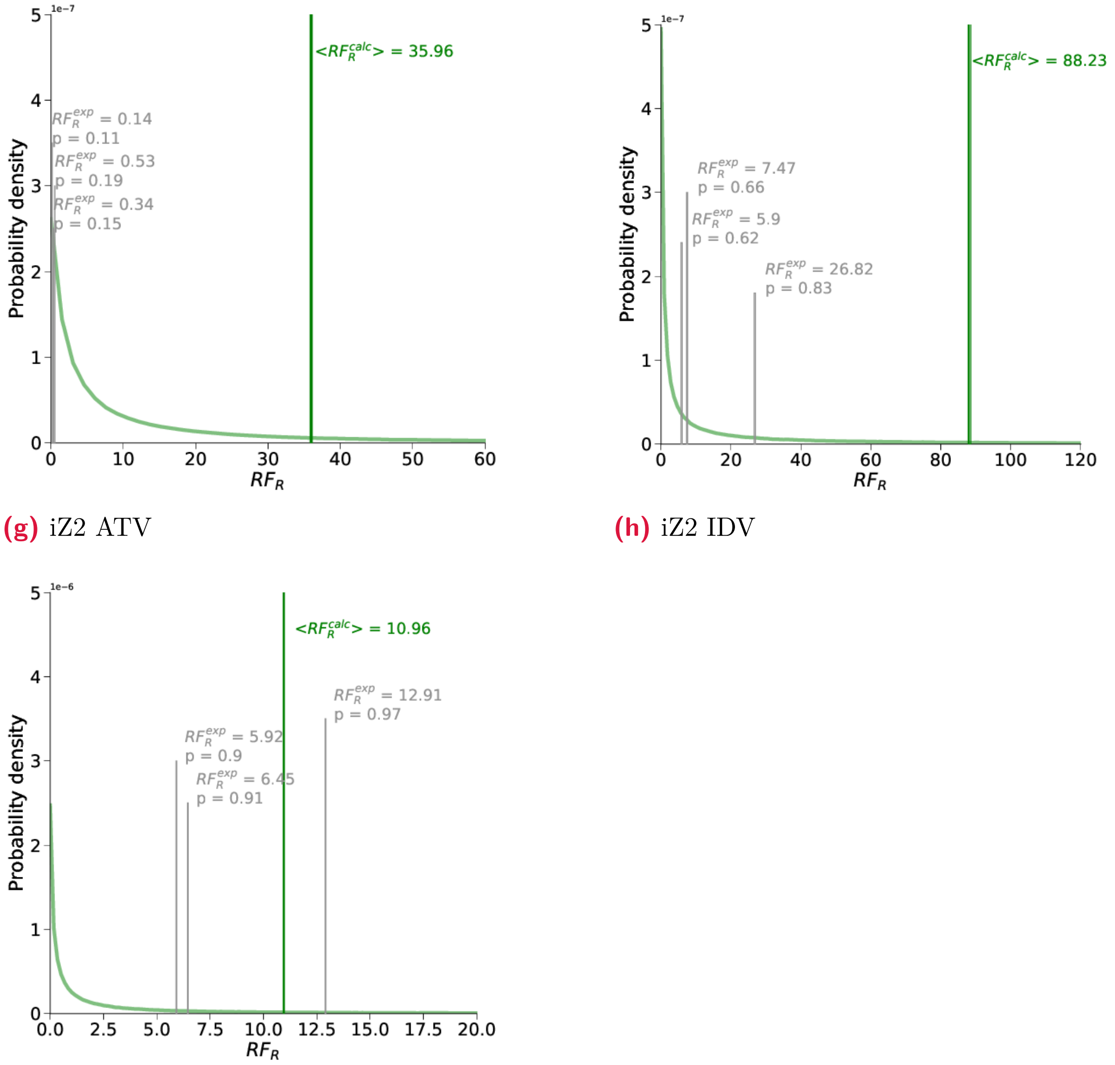
(c) GH9 ATV

(d) GH9 SQV

(e) RU1 ATV

(f) RU1 LPV

Figure B.2: Calculated RF$_R$ distributions and experimental estimates for dataset 3. p designates the proportion of RF_R^{calc} at least as extreme as RF_R^{exp} compared to $<RF_R^{calc}>$.

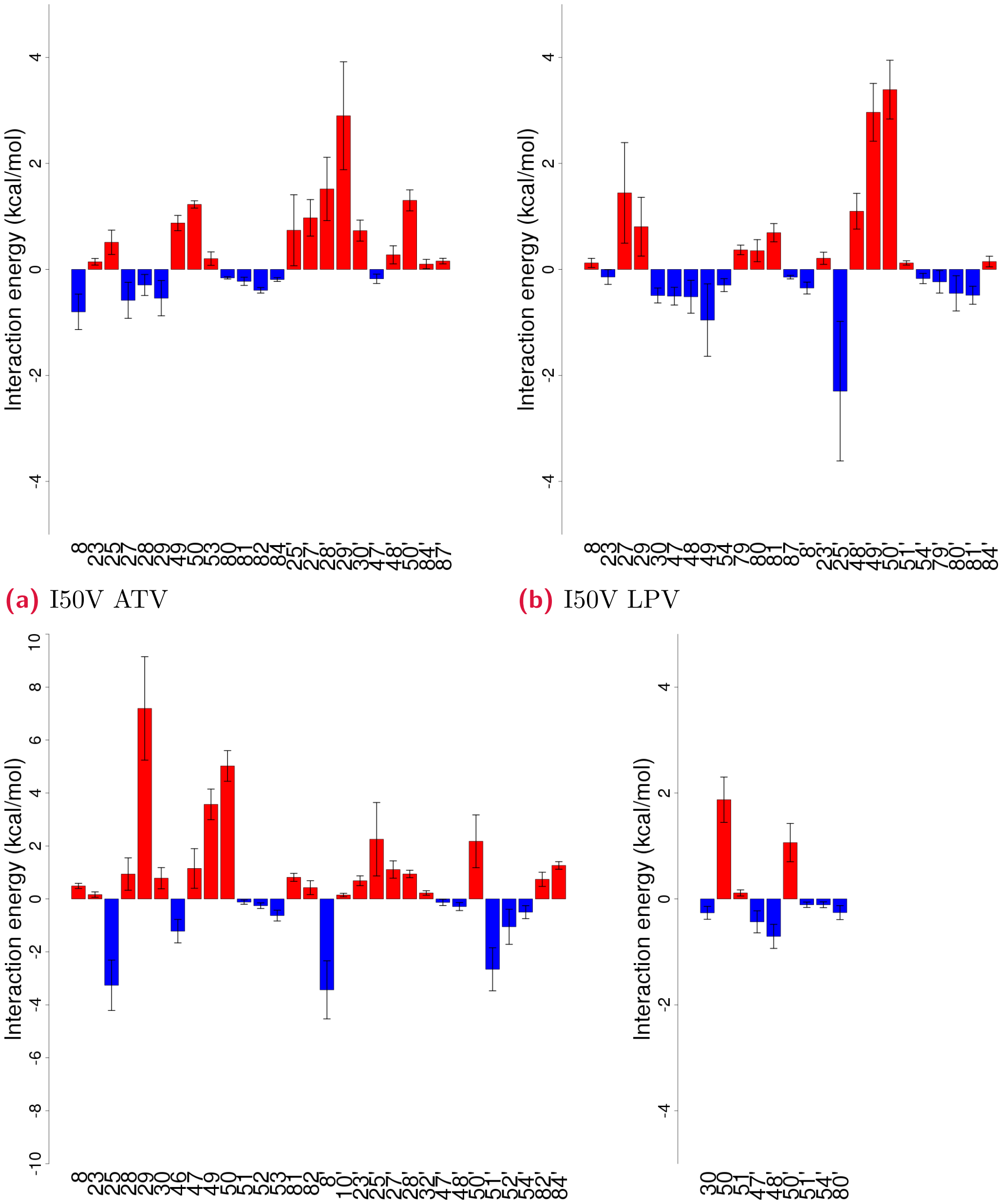


(g) iZ2 ATV

(h) iZ2 IDV

(i) iZ2 LPV

Figure B.2: Calculated RF_R distributions and experimental estimates for dataset 3. p designates the proportion of RF_R^{calc} at least as extreme as RF_R^{exp} compared to $<RF_R^{calc}>$ (cont.)



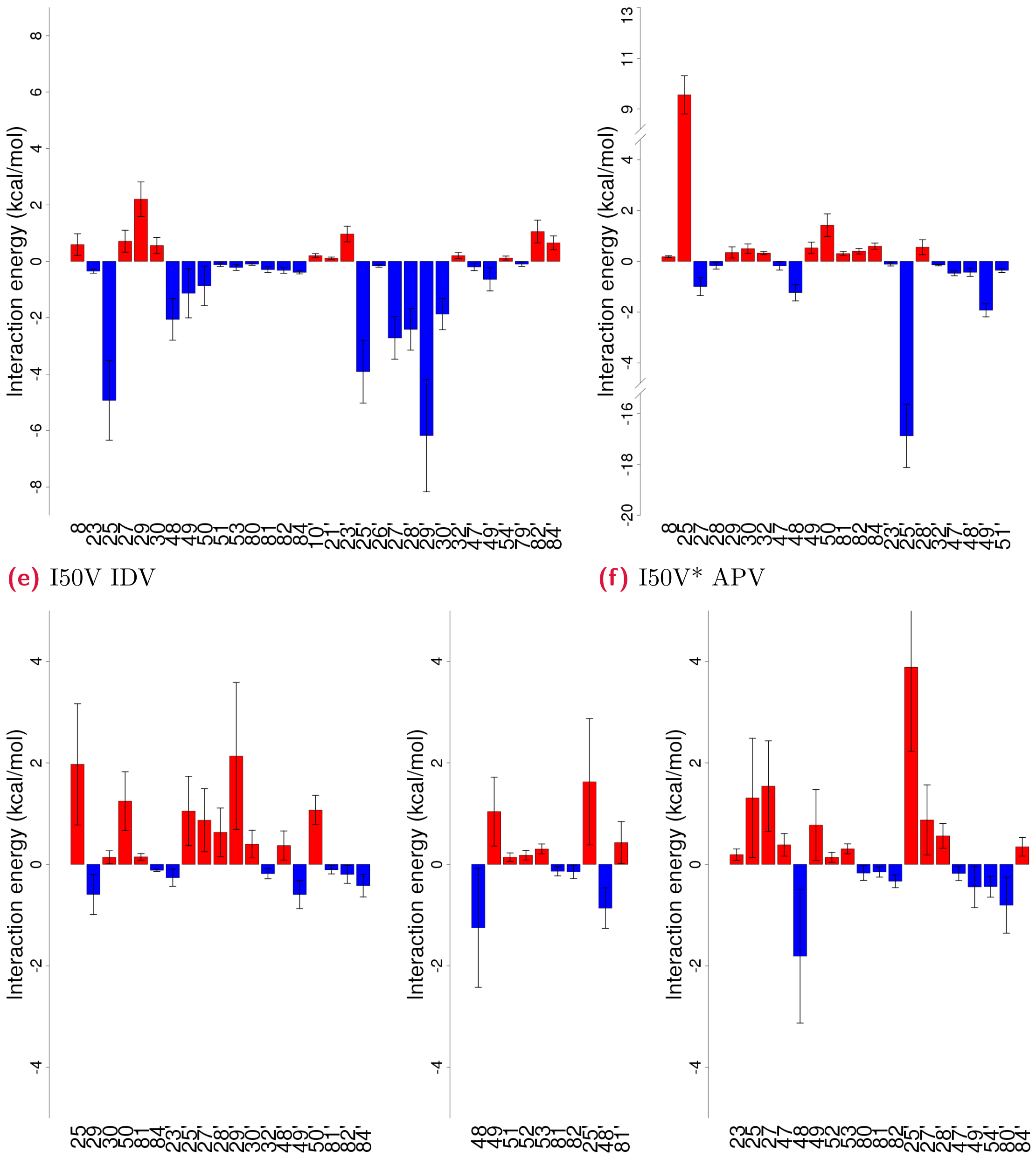
(a) I50V ATV

(b) I50V LPV

(c) I50V TPV

(d) I50V APV

Figure B.3: Energy differences of non-bonded interaction between protein and inhibitor in wildtype and mutant complexes for dataset 1. Only residues, for which the difference between the wildtype and the mutant complexes is higher than the propagated error and its absolute value higher than 0.1 kcal/mol are shown.



(e) I50V IDV

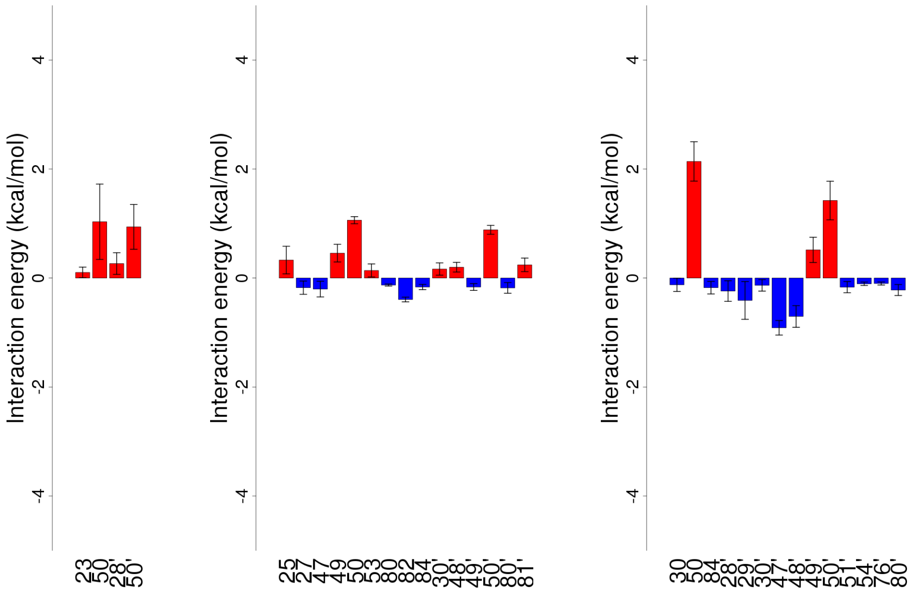
(f) I50V* APV

(g) I50V* IDV

(h) G48V SQV

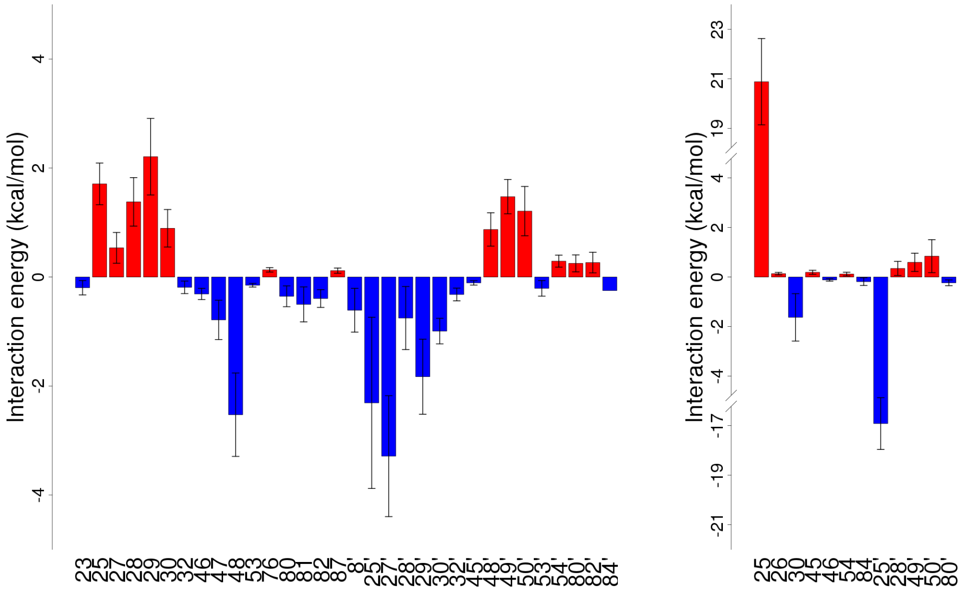
(i) G48V/L90M SQV

Figure B.3: Energy differences of non-bonded interaction between protein and inhibitor in wildtype and mutant complexes for dataset 1. Only residues, for which the difference between the wildtype and the mutant complexes is higher than the propagated error and its absolute value higher than 0.1 kcal/mol are shown (cont.)



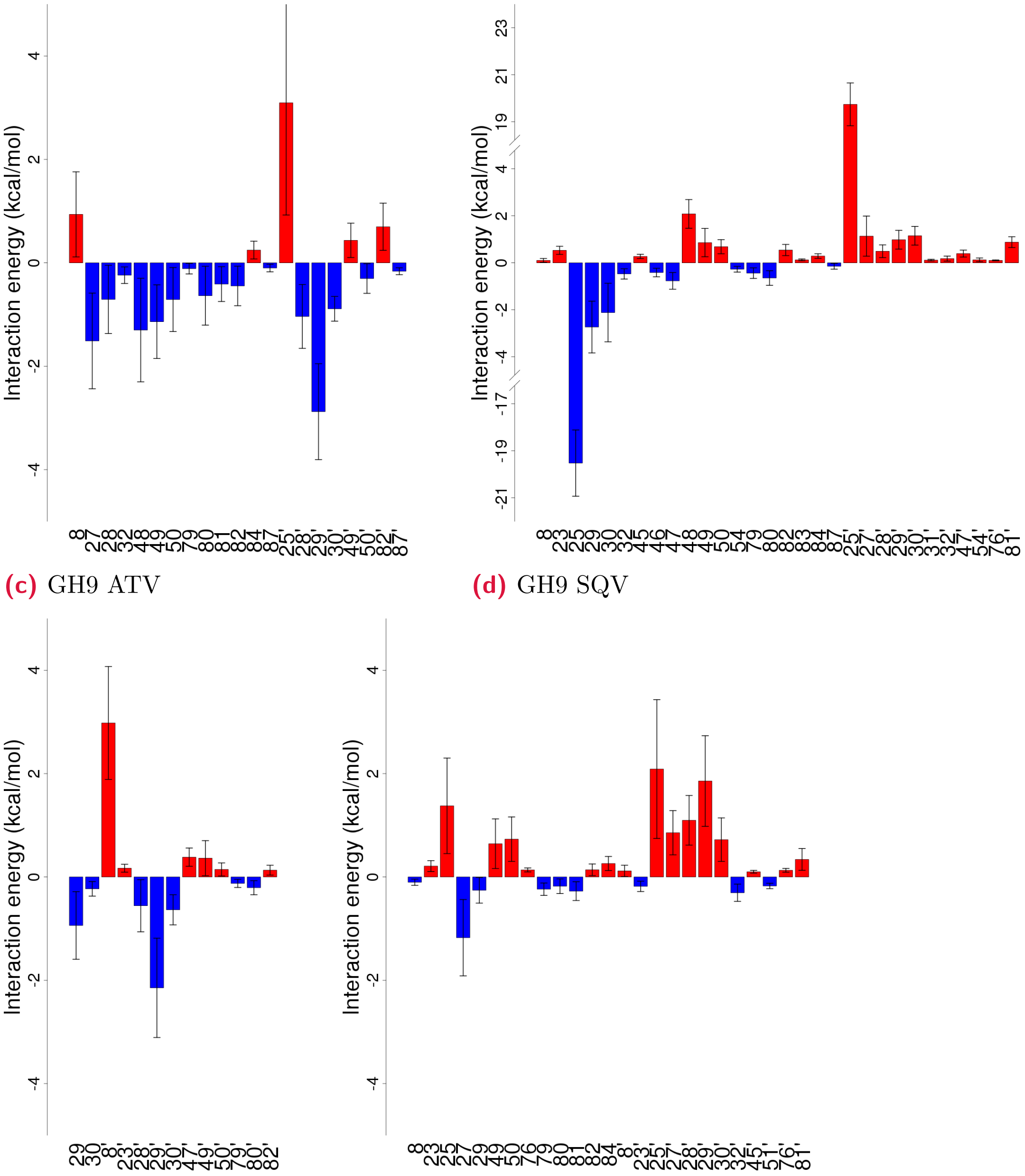
(j) L90M SQV **(k)** I50V/A71V ATV **(l)** I50V/A71V APV

Figure B.3: Energy differences of non-bonded interaction between protein and inhibitor in wildtype and mutant complexes for dataset 1. Only residues, for which the difference between the wildtype and the mutant complexes is higher than the propagated error and its absolute value higher than 0.1 kcal/mol are shown (cont.)



(a) FB15 ATV **(b)** FB15 SQV

Figure B.4: Energy differences of non-bonded interaction between protein and inhibitor in wildtype and mutant complexes for dataset 3. Only residues, for which the difference between the wildtype and the mutant complexes is higher than the propagated error and its absolute value higher than 0.1 kcal/mol are shown.



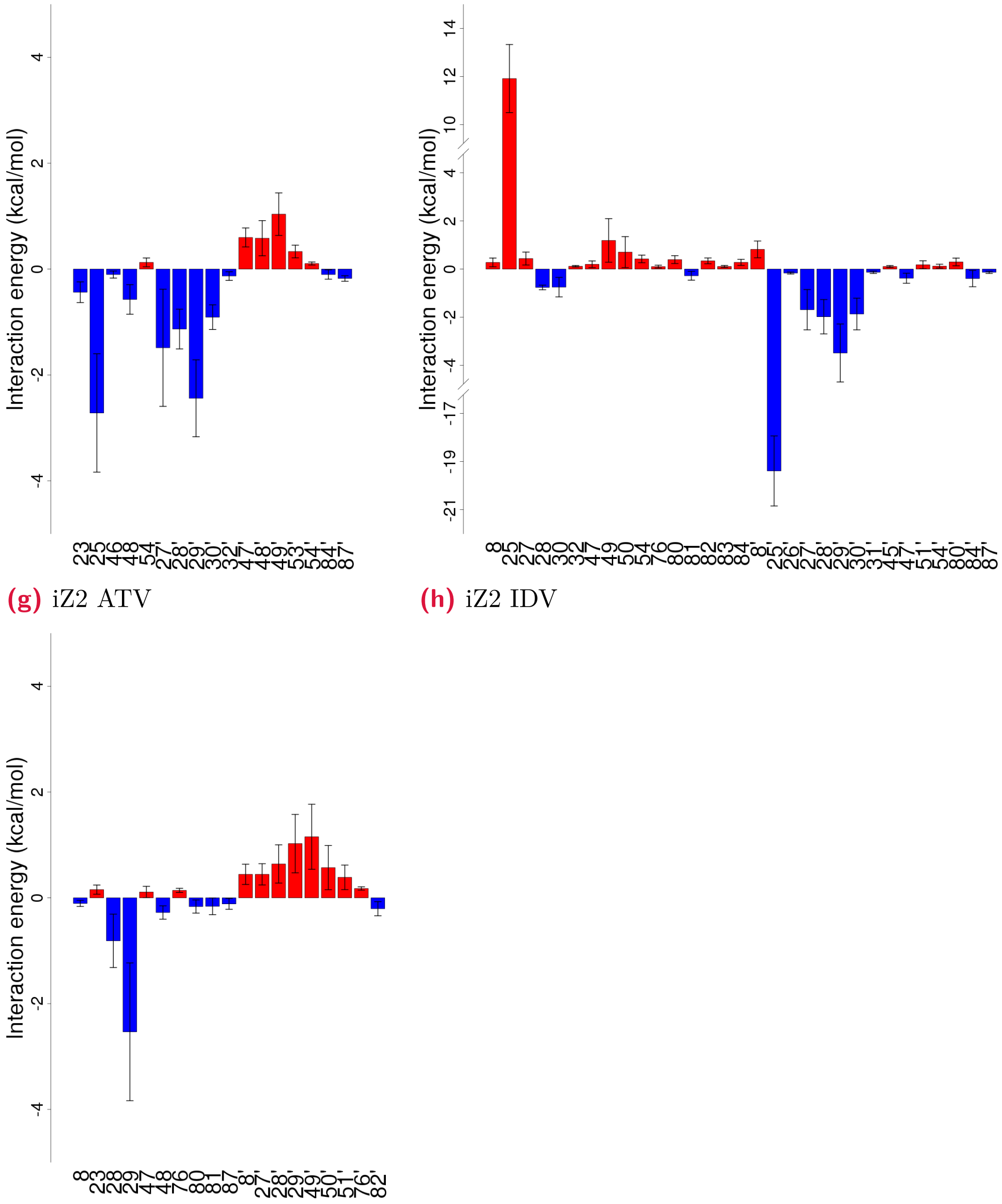
(c) GH9 ATV

(d) GH9 SQV

(e) RU1 ATV

(f) RU1 LPV

Figure B.4: Energy differences of non-bonded interaction between protein and inhibitor in wildtype and mutant complexes for dataset 3. Only residues, for which the difference between the wildtype and the mutant complexes is higher than the propagated error and its absolute value higher than 0.1 kcal/mol are shown.

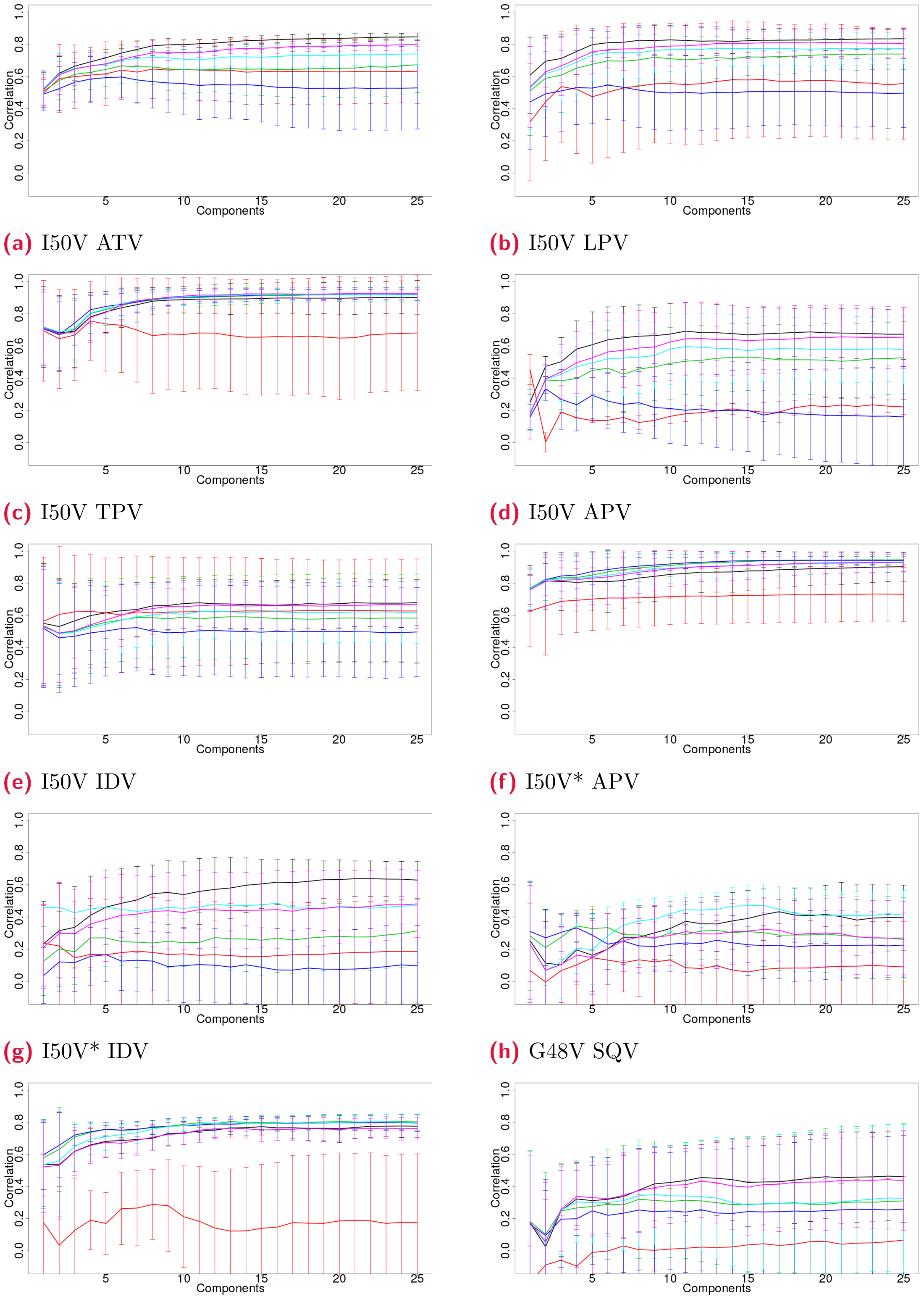


(g) iZ2 ATV

(h) iZ2 IDV

(i) iZ2 LPV

Figure B.4: Energy differences of non-bonded interaction between protein and inhibitor in wildtype and mutant complexes for dataset 3. Only residues, for which the difference between the wildtype and the mutant complexes is higher than the propagated error and its absolute value higher than 0.1 kcal/mol are shown (cont.)



(a) I50V ATV

(b) I50V LPV

(c) I50V TPV

(d) I50V APV

(e) I50V IDV

(f) I50V* APV

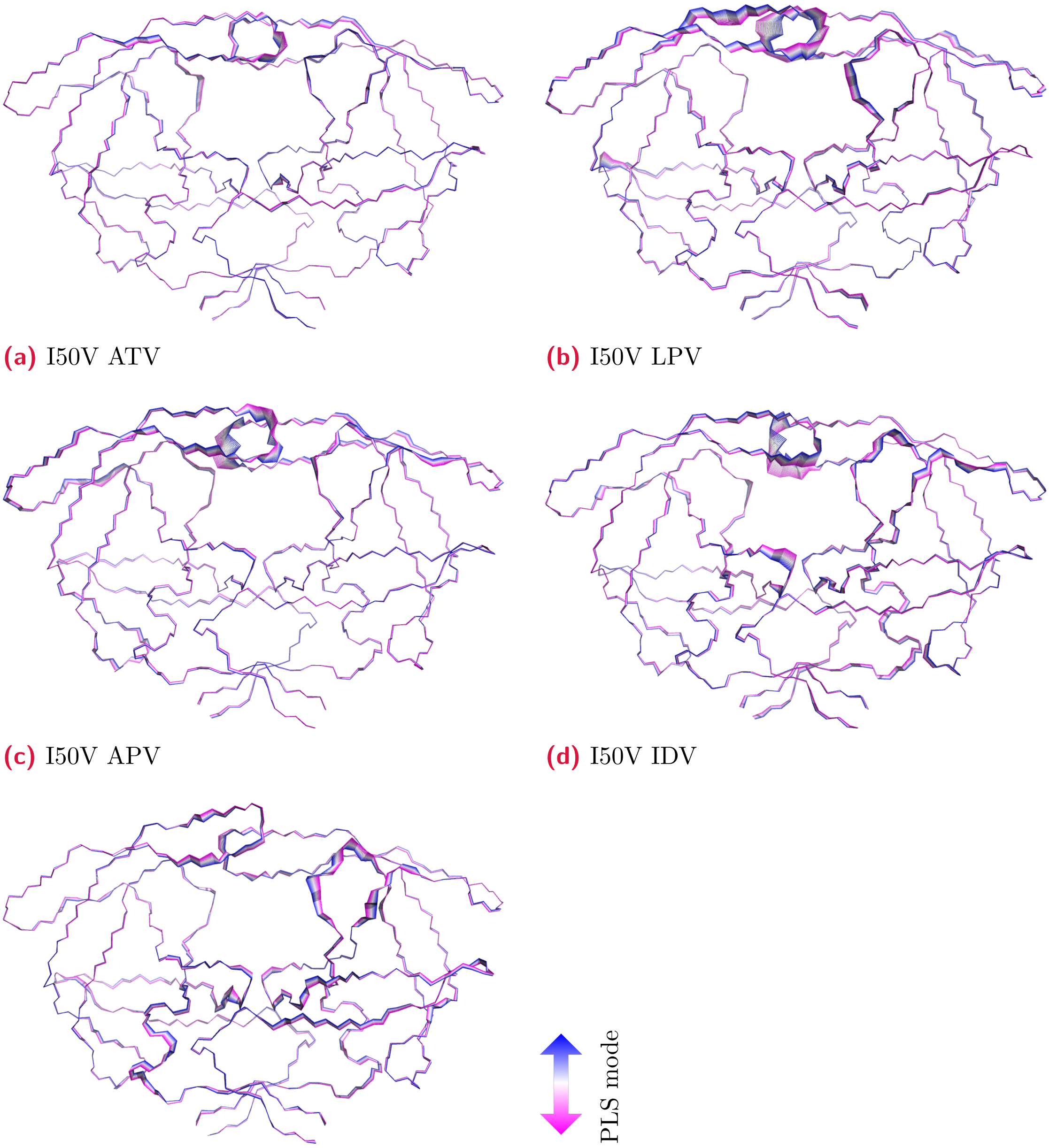
(g) I50V* IDV

(h) G48V SQV

(i) G48V/L90M SQV

(j) L90M SQV

Figure B.5: Correlation between predicted and true labels in the PLS models for the protein backbone trajectories from dataset 1 in cross validation for backbone (black), protein (green), side chain (blue), N+C$_\alpha$+C$_\beta$+C$_D$/S$_G$ (cyan), C$_\alpha$+C$_\beta$ (magenta), and inhibitor (red) heavy atoms.



(a) I50V ATV

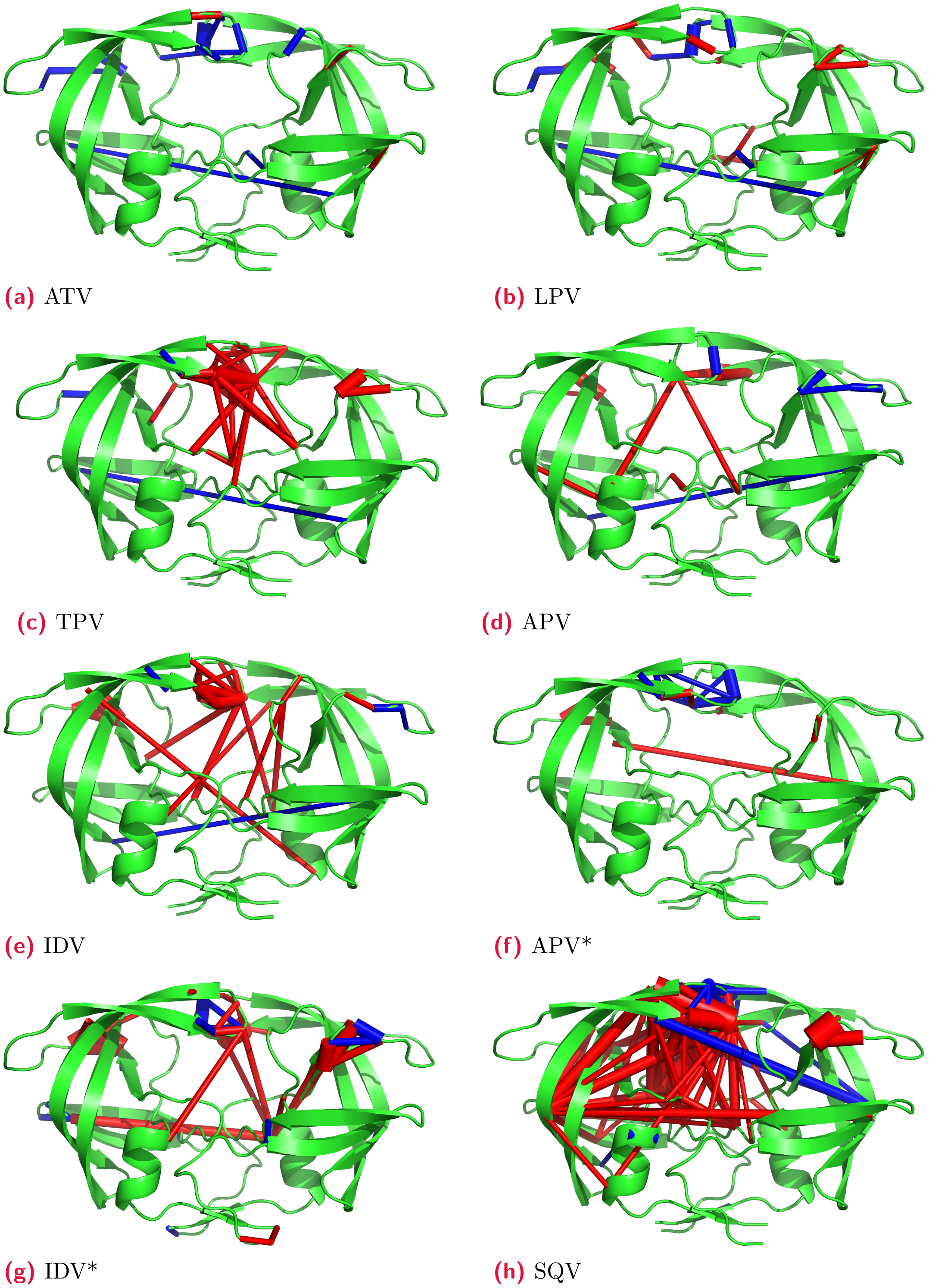
(b) I50V LPV

(c) I50V APV

(d) I50V IDV

(e) G48V/L90M SQV

Figure B.6: Interpolation between the extremes of the PLS models for the corresponding complexes for dataset 1. Blue-to-magenta bands correspond to the interpolation along the mode which relates the true label of simulation, wildtype or mutant, to the underlying differences in protein motions.


(a) ATV

(b) LPV

(c) TPV

(d) APV

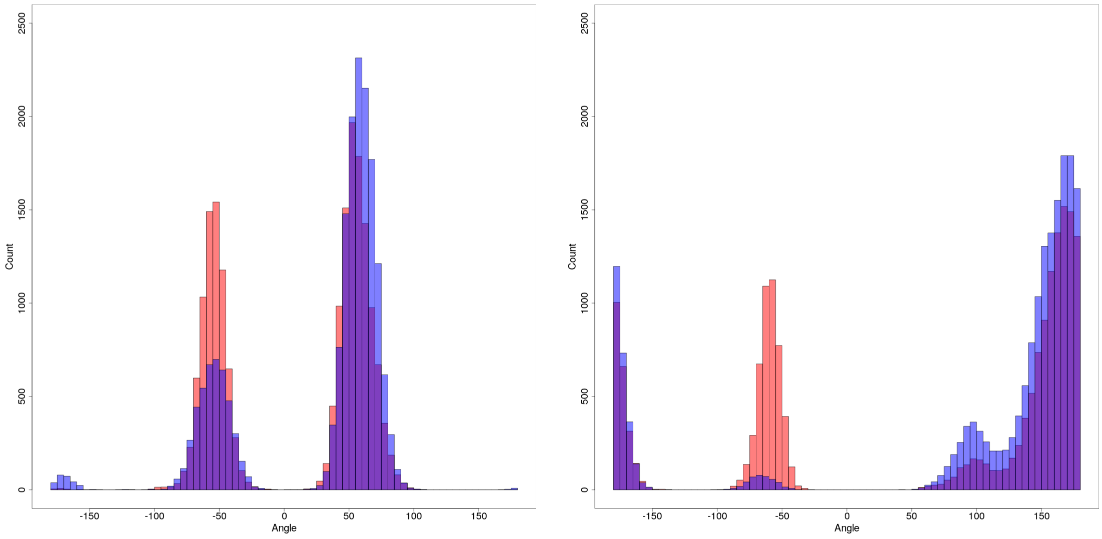
(e) IDV

(f) APV*

(g) IDV*

(h) SQV

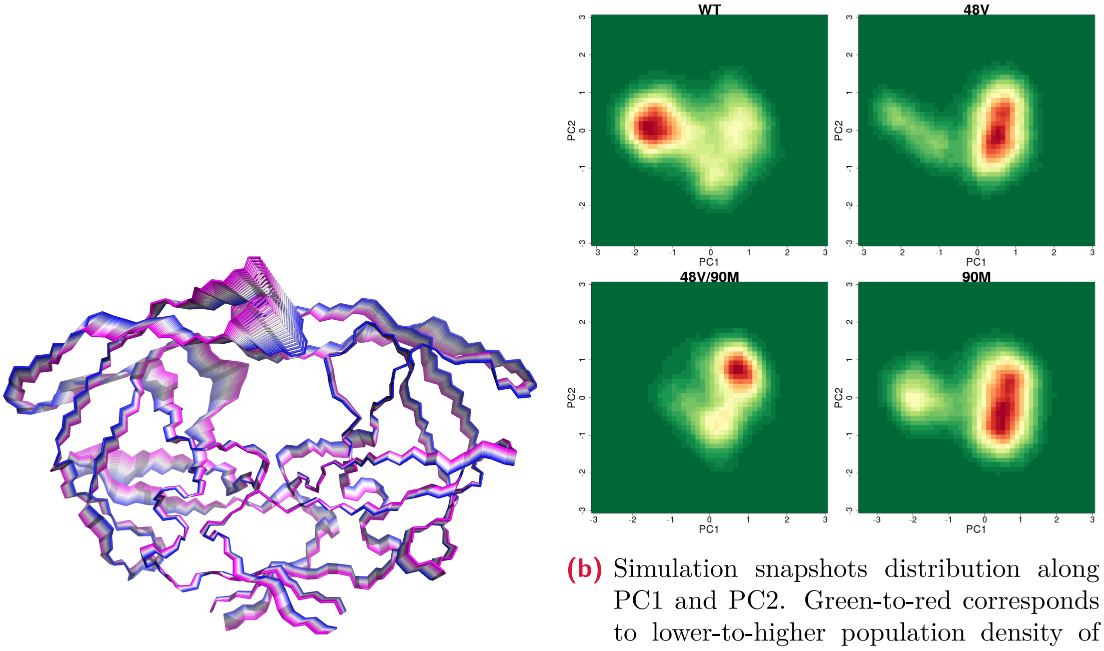
Figure B.7: Mutual information mapped onto the wildtype protease structures for dataset 1. Cylinders connecting residues represent differences in mutual information between those residues in apo and holo wildtype simulations, with the width of the cylinder proportional to the difference, red indicating higher mutual information in the holo complexes and blue in the apo complexes. This corresponds to the degree to which residue pairs exhibit differences in the correlation of their motions.



(a) χ_1

(b) χ_2

Figure B.8: Distribution of I84 side chain χ angles in WT (rose) and L90M (blue) mutant proteases in complex with SQV simulations from dataset 1.



(a) Motion between the extremes along PC1

(b) Simulation snapshots distribution along PC1 and PC2. Green-to-red corresponds to lower-to-higher population density of snapshots

Figure B.9: PCA of wildtype, G48V, G48V/L90M, and L90M protease complexes with SQV from dataset 1. Blue-to-magenta in (a) corresponds to minimum-to-maximum along the first principal component (PC) in (b).

Inhibitor	Genotype	Reference protonated aspartic acid	$\Delta\Delta G^{prot}_{WT}$	$\Delta\Delta G^{prot}_{MUT}$
APV	M46I	D25′	−1.86 ± 0.23	−2.32 ± 0.21
IDV	M46I	D25	1.15 ± 0.27	0.53 ± 0.45
APV	I84V	D25′	−1.67 ± 0.25	−0.31 ± 0.27
IDV	I84V	D25	1.67 ± 0.44	1.29 ± 0.38
LPV	I84V	D25′	−0.6 ± 0.35	−1.33 ± 0.3
SQV	I84V	D25	1.03 ± 0.19	2.04 ± 0.27
APV	N88S	D25	2.23 ± 0.29	−0.16 ± 0.24
IDV	N88S	D25	2.22 ± 0.51	1.42 ± 0.45

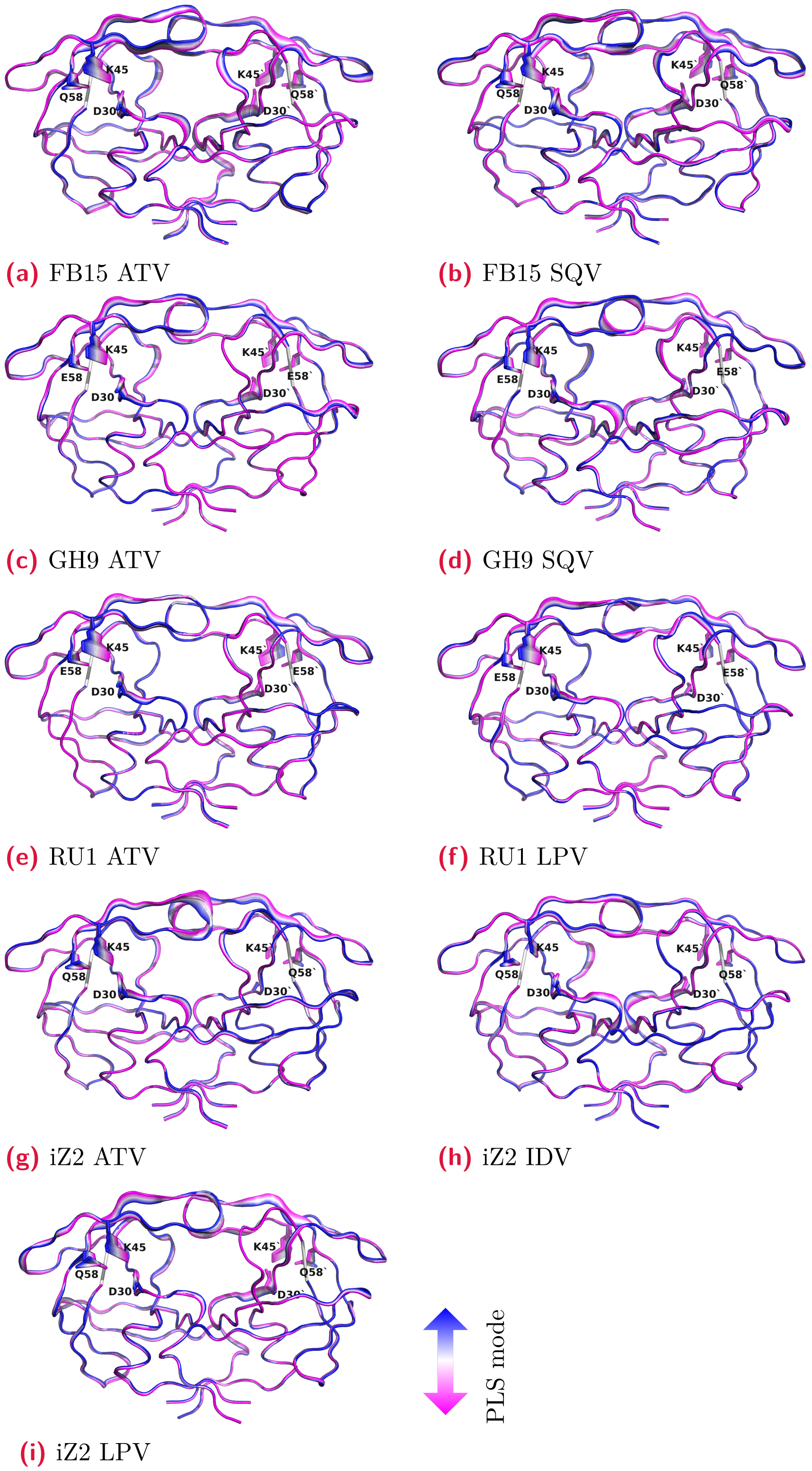
Table B.4: Change of the free energy of inhibitor binding upon switching the proton from the reference protonated active site residue to the active site residue on the opposite subunit for dataset 2. ± shows bootstrap error estimate, all values in kcal/mol.

Inhibitor	Genotype	Reference protonated aspartic acid	$\Delta\Delta G^{prot}_{WT}$	$\Delta\Delta G^{prot}_{MUT}$
ATV	FB15	D25′	−1.17 ± 0.31	−2.08 ± 0.36
SQV	FB15	D25	−0.08 ± 0.25	0.4 ± 0.37
ATV	GH9	D25′	−0.13 ± 0.21	−1.23 ± 0.32
SQV	GH9	D25	0.45 ± 0.21	−0.08 ± 0.31
ATV	RU1	D25′	−4.21 ± 0.38	−1.76 ± 0.47
LPV	RU1	D25′	−1.41 ± 0.32	−0.75 ± 0.37
ATV	iZ2	D25	−0.82 ± 0.34	−1.25 ± 0.4
IDV	iZ2	D25′	−0.95 ± 0.44	1.43 ± 0.31
LPV	iZ2	D25	0.56 ± 0.27	0.5 ± 0.75

Table B.5: Change of the free energy of inhibitor binding upon switching the proton from the reference protonated active site residue to the active site residue on the opposite subunit for dataset 3. ± shows bootstrap error estimate, all values in kcal/mol.

Inhibitor and mutation	ATV	LPV	TPV	APV	IDV	APV	IDV	SQV	SQV	SQV	Overall
combination	I50V	I50V	I50V	I50V	I50V	I50V*	I50V*	G48V	G48V/L90M	L90M	Mean
Jaccard Index	0.05	0.19	0.53	0.16	0.44	0.24	0.27	0.33	0.35	0.28	0.28

Table B.6: Jaccard index estimates between two difference matrices of mutual information for dataset 1: the first is apo vs. holo simulations of the wildtype protein and the second is for wildtype vs. mutant simulations for the holo protein.



(a) FB15 ATV

(b) FB15 SQV

(c) GH9 ATV

(d) GH9 SQV

(e) RU1 ATV

(f) RU1 LPV

(g) iZ2 ATV

(h) iZ2 IDV

(i) iZ2 LPV

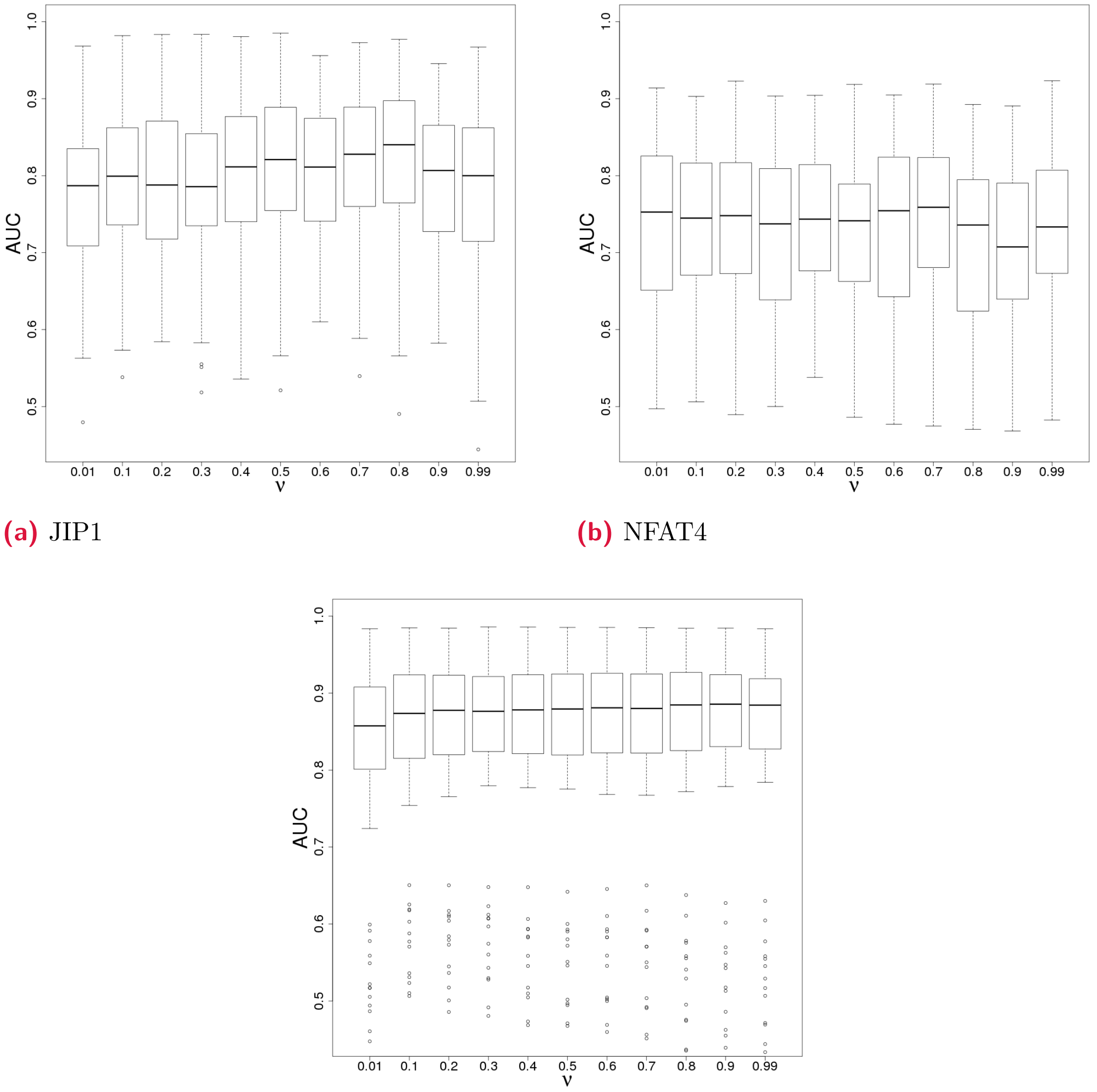
Figure B.10: Interpolation between the extremes of the PLS models for the corresponding complexes for dataset 3. Blue-to-magenta bands correspond to the interpolation along the mode as represented as cartoon for backbone and as sticks residues 30, 45, and 58. Mutated residue 76 is not part of the model and represented here as gray dash.

Residues	**Inhibitor**	**N88**	**S88**	**N88′**	**S88′**
D30/D30′	**APV**	$0.005 \pm 2 \times 10^{-4}$	0.61 ± 0.05	$5 \times 10^{-4} \pm 3 \times 10^{-7}$	0.66 ± 0.12
	IDV	$2 \times 10^{-4} \pm 2 \times 10^{-7}$	0.22 ± 0.008	$0.001 \pm 2 \times 10^{-6}$	0.55 ± 0.08
T31/T31′	**APV**	1.28 ± 0.004	0.21 ± 0.03	1.56 ± 0.003	0.13 ± 0.04
	IDV	1.26 ± 0.007	0.47 ± 0.03	1.35 ± 0.008	0.27 ± 0.04
T74/T74′	**APV**	0.8 ± 0.003	$3 \times 10^{-4} \pm 2 \times 10^{-7}$	0.85 ± 0.001	$2 \times 10^{-4} \pm 2 \times 10^{-7}$
	IDV	0.71 ± 0.001	$2 \times 10^{-4} \pm 2 \times 10^{-7}$	0.79 ± 0.001	$3 \times 10^{-4} \pm 9 \times 10^{-7}$

Table B.7: Average hydrogen bonds number between residues D30, T31, and T74 and N88 and S88 for wildtype and mutant complexes, respectively, in dataset 2. Columns 3 and 4 of the table corresponds to hydrogen bonds within monomer A of protease and columns 5 and 6 of the table corresponds to hydrogen bonds within monomer B of protease (residues marked with prime symbol). ± indicates standard error of bond frequency across independent simulations.

Inhibitor	Mutation	$\Delta\Delta G_{exp}$	$\Delta\Delta G_{total}$	$\Delta\Delta G_{theor}$
APV	I50V*	2.03	1.21(0.82)	$-5.07(7.1)^a$
IDV	I50V*	2.33	0.32(2.01)	$-4.40(6.73)^a$
APV	I50V*	2.03	1.21(0.82)	$1.49(0.54)^a$
IDV	I50V*	2.33	0.32(2.01)	$2.89(0.56)^a$
SQV	G48V	2.78	2.66(0.12)	$3.73(0.71)^b$
SQV	L90M	1.60	0.09(1.51)	$3.49(2.13)^b$
SQV	G48V/L90M	4.03	6.03(2)	$4.4(0.37)^b$

Table B.8: Experimental estimates of the change of the free energy of inhibitor binding upon mutation used the current study, $\Delta\Delta G_{exp}$, theoretical estimates from the current study, $\Delta\Delta G_{total}$, and computational predictions for the same mutations from the literature, $\Delta\Delta G_{theor}$, with unsigned error in the parenthesis. ^{a} taken from (R. Duan *et al.* 2015) and ^{b} from (Stoica *et al.* 2008). All values in kcal/mol.

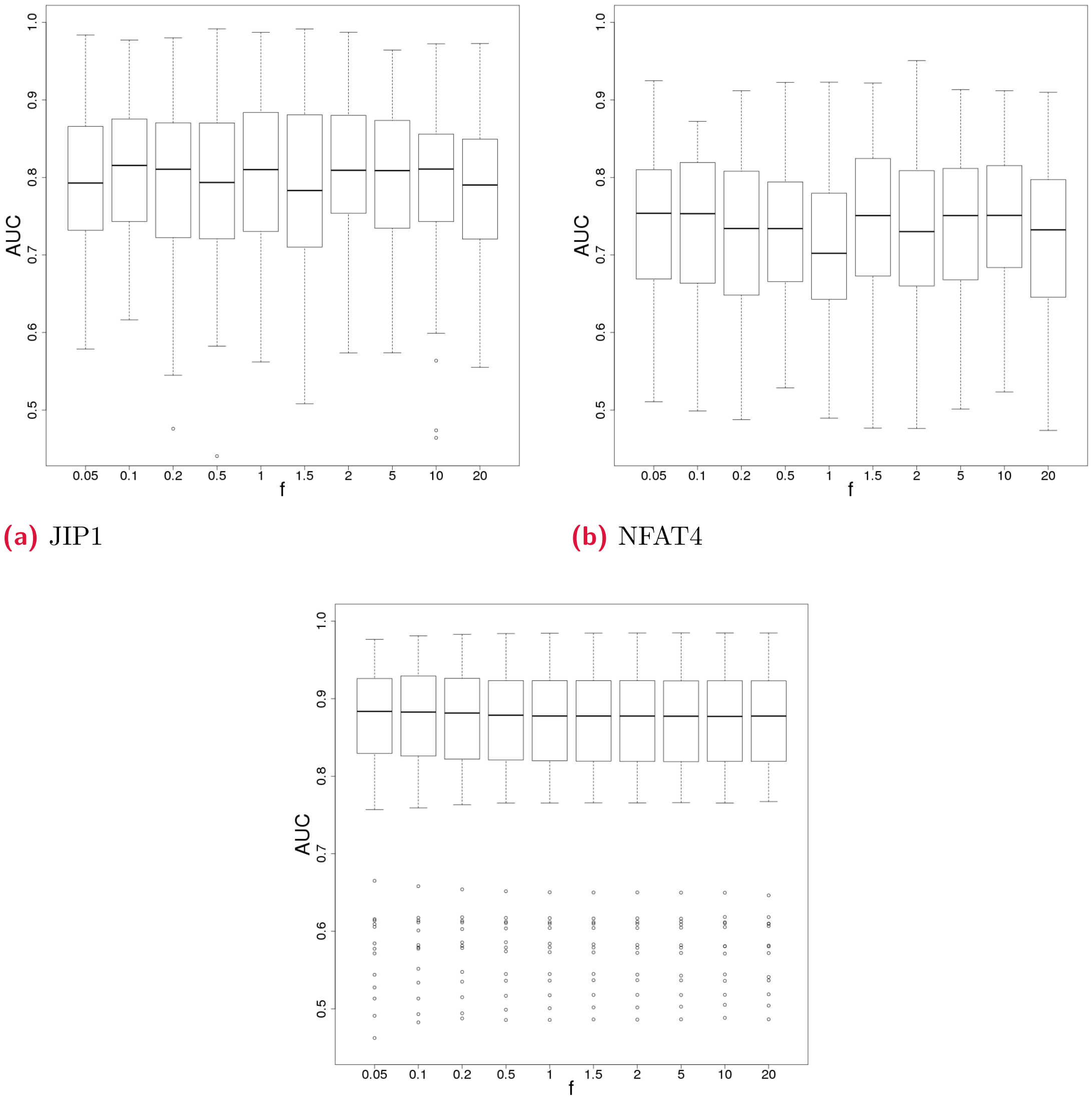


(a) JIP1

(b) NFAT4

(c) Greater MEF2A

Figure B.11: Performance of SVM models using RBF kernel with physical-chemical residues encoding for classification of MAPK D-motifs in terms of AUC for different ν parameter values.

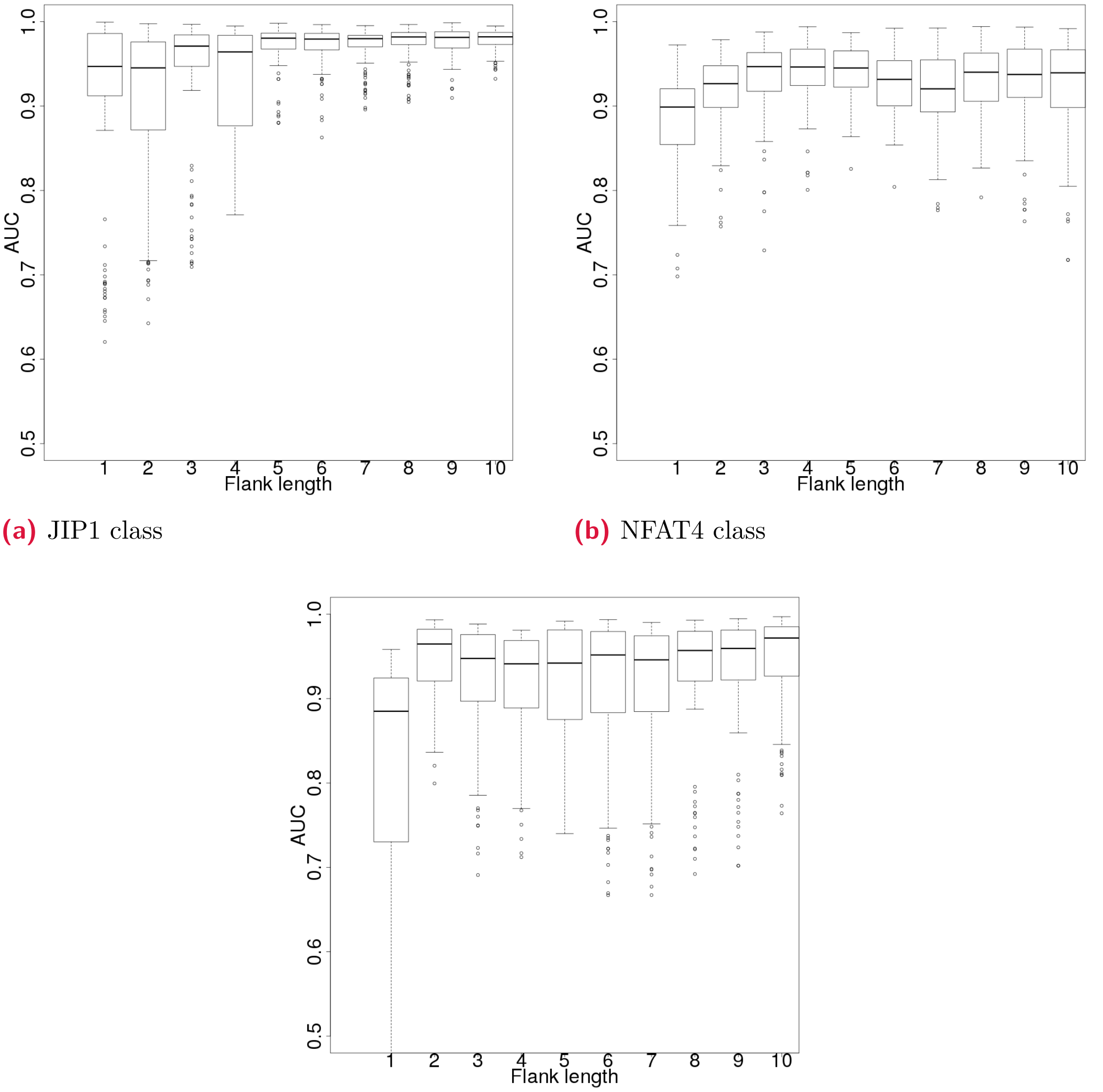


(a) JIP1

(b) NFAT4

(c) Greater MEF2A

Figure B.12: Performance of SVM models using RBF kernel with physical-chemical residues encoding for classification of MAPK D-motifs in terms of AUC for different kernel width $\gamma = \frac{1}{df}$ values.

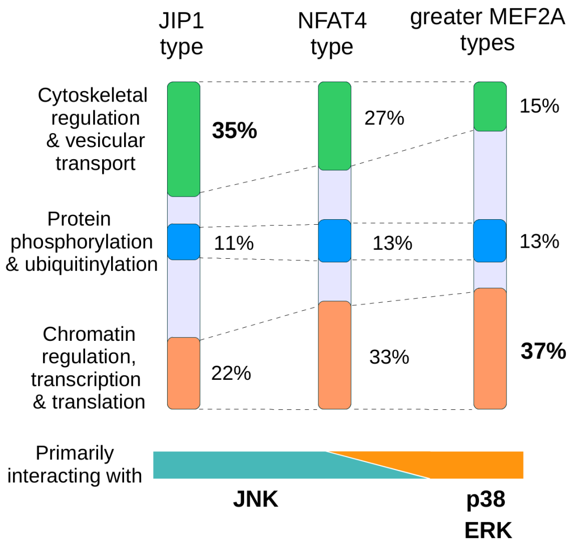


(a) JIP1 class

(b) NFAT4 class

(c) Greater MEF2A class

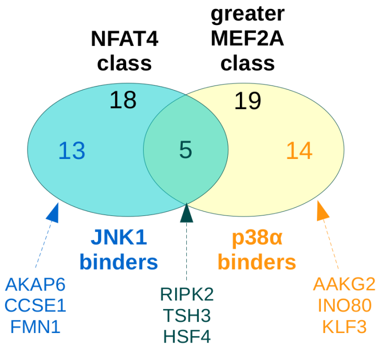
Figure B.13: Performance of the PSSM-based classification using D-motifs of different flank lengths.



(a) Comparison of functional grouping of hits among the 100 highest scoring JIP1, NFAT4 or greater MEF2A type motifs.

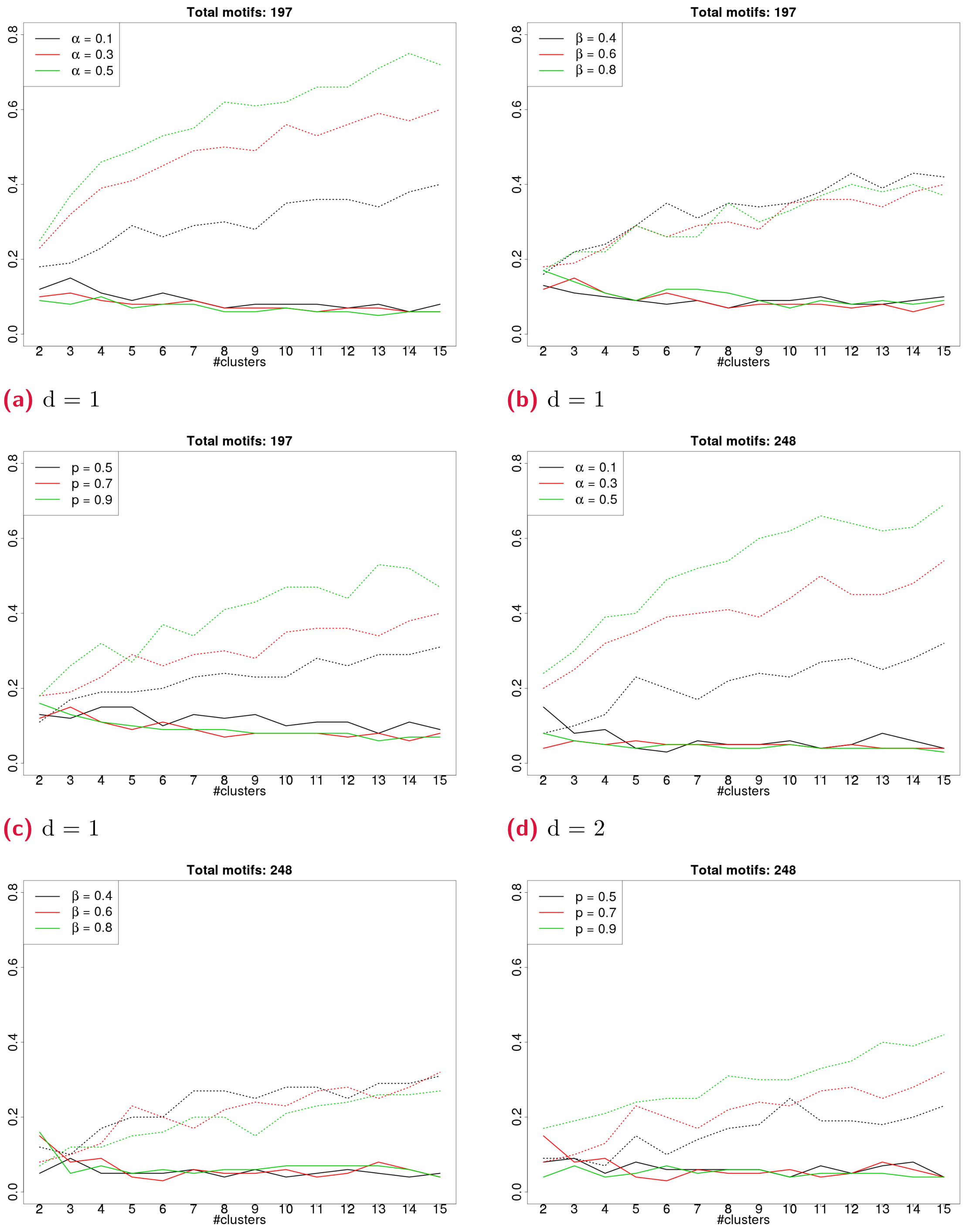
Family	Function	Protein with JIP1-like element	Protein with NFAT4-like element
Formin	Actin filament assembly	**FMNL1**	**FMN1** **FHOD3**
Smoothelin	Actin binding	**SMTL2**	**SMTN**
HECT-WW	E3 ubiquitin ligase	**HECW1**	**HECW2**
ATF/CREB	Transcription factor	**ATF7**	**ATF7**

(b) Selected examples of protein families in which one member carries a JIP1-type motif, but another closely related member has an NFAT4-type motif instead. Shown example motifs are either high-ranking predictions or were experimentally validated.



(c) Experimental proof for the suggested overlap between JNK1 and p38α partners. The upper numbers in this diagram represent the number of motifs (of either NFAT4 or greater MEF2A types) that were tested as positive in the dot-blot arrays. The lower row of numbers shows the numbers that selectively interact with JNK1 (blue), p38α (yellow) or with both (green) in the dot-blot arrays. For each case, a few characteristic examples are also given below.

Figure B.14: Comparisons of highest scoring 100 hits for JIP1, NFAT4, and greater MEF2A type motifs.



(a) d = 1

(b) d = 1

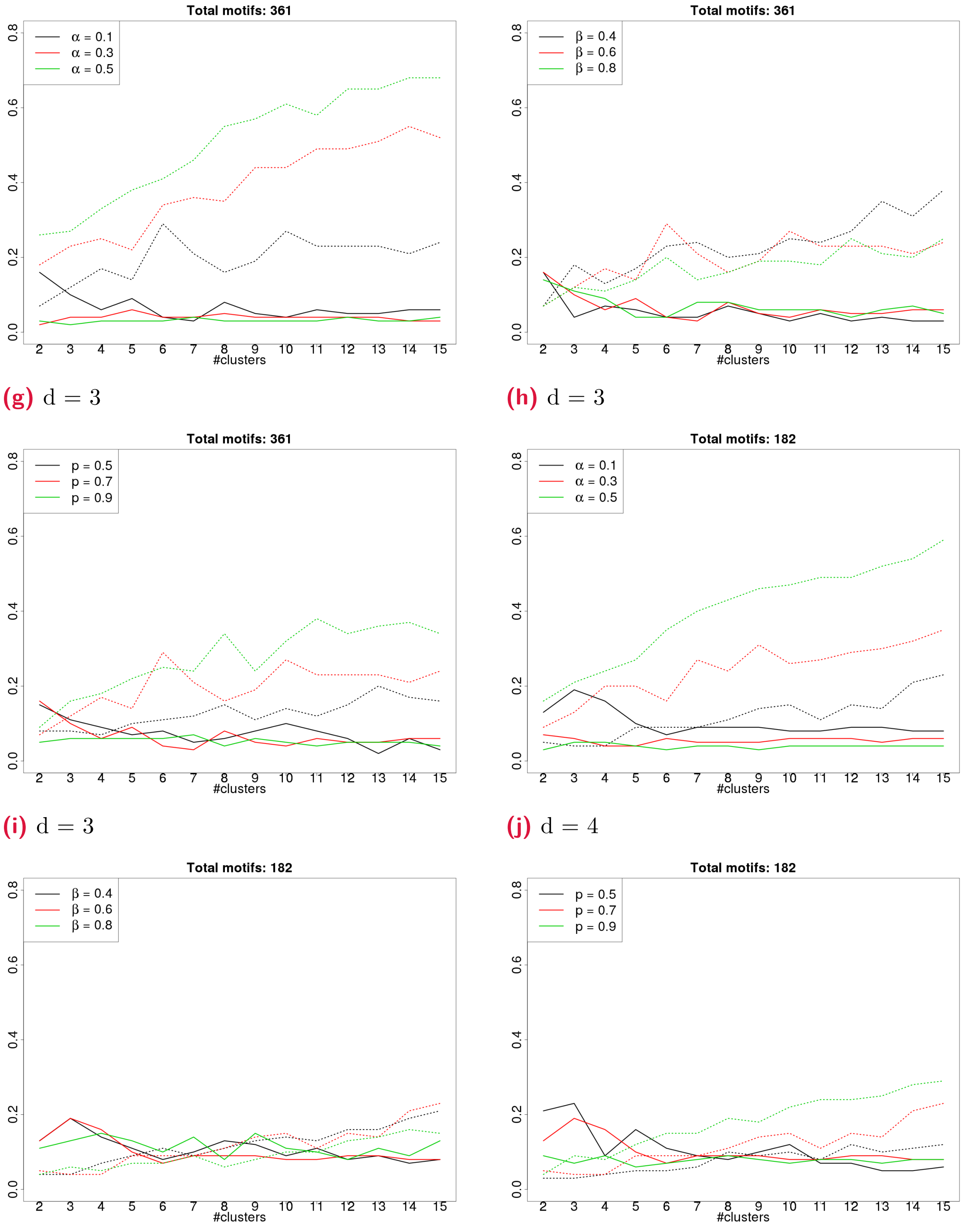
(c) d = 1

(d) d = 2

(e) d = 2

(f) d = 2

Figure B.15: The evaluation of performance of clustering of phosphorlytion site groups with different distances between the [ST] sites (d) in terms of silhouette value (solid lines) and proportion of peptides assigned to clusters (dashed lines) while varying separately α, β, and p parameters.



(g) d = 3

(h) d = 3

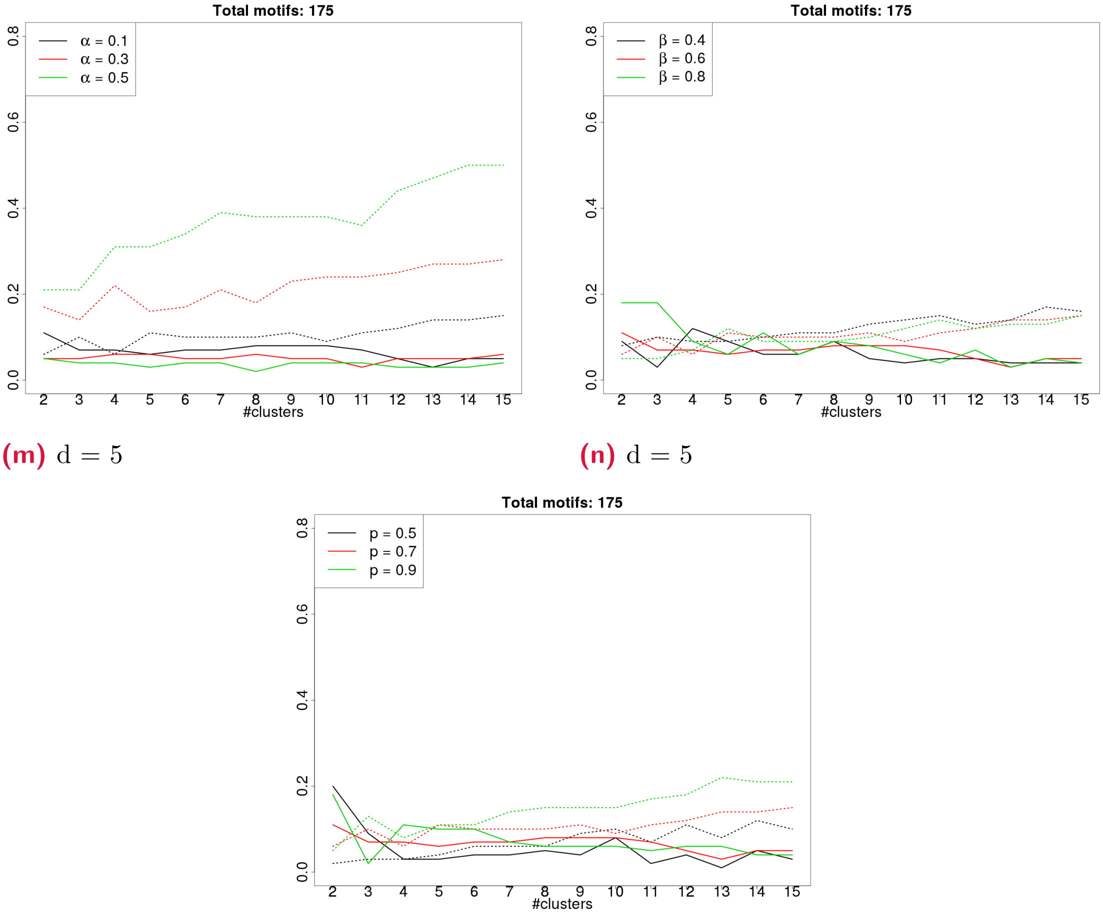
(i) d = 3

(j) d = 4

(k) d = 4

(l) d = 4

Figure B.15: The evaluation of performance of clustering of phosphorlytion site groups with different distances between the [ST] sites (d) in terms of silhouette value (solid lines) and proportion of peptides assigned to clusters (dashed lines) while varying separately α, β, and p parameters (cont.)



(m) d = 5

(n) d = 5

(o) d = 5

Figure B.15: The evaluation of performance of clustering of phosphorlytion site groups with different distances between the [ST] sites (d) in terms of silhouette value (solid lines) and proportion of peptides assigned to clusters (dashed lines) while varying separately α, β, and p parameters (cont.)

C List of Publications

Zeke, A., **Bastys, T.**, Alexa, A., Garai, Á., Mészáros, B., Kirsch, K., Dosztányi, Z., Kalinina, O. V., and Reményi, A. (2015). "Systematic discovery of linear binding motifs targeting an ancient protein interaction surface on MAP kinases." *Molecular Systems Biology* 11.11, p. 837. doi: 10.15252/msb.20156269.

Bastys, T., Gapsys, V., Doncheva, N. T., Kaiser, R., Groot, B. L. de, and Kalinina, O. V. (2018). "Consistent Prediction of Mutation Effect on Drug Binding in HIV-1 Protease Using Alchemical Calculations." *Journal of Chemical Theory and Computation* 14.7, pp. 3397–3408. doi: 10.1021/acs.jctc.7b01109.

Bastys, T., Gapsys, V., Doncheva, N. T., Walter, H., Kaiser, R., Groot, B. L. de, and Kalinina, O. V. "Mutations in Non-Active Site Residues of the HIV-1 Protease Influence Resistance and Sensitization Towards Protease Inhibitors." *Manuscript in preparation.*

References

Antiretroviral Drugs Used in the Treatment of HIV Infection. https://www.fda.gov/forpatients/illness/hivaids/treatment/ucm118915.htm. [Online; accessed 25-October-2018].

Chemaxon Calculator 5.3.8. http://www.chemaxon.com. ChemAxon, 2010.

PDB Current Holdings Breakdown. http://www.rcsb.org/pdb/statistics/holdings.do. [Online; accessed 14-August-2018].

The Stanford HIV Drug Resistance Database. https://hivdb.stanford.edu/dr-summary/resistance-notes/PI/. [Online; accessed 29-May-2018].

National Institute of Allergy and Infectious Diseases (2010). https://www.niaid.nih.gov/diseases-conditions/hiv-replication-cycle. [Online; accessed 12-August-2018]. Copyright notice: Licensed under CC BY 2.0.

Abram, M. E., Ferris, A. L., Shao, W., Alvord, W. G., and Hughes, S. H. (2010). "Nature, Position, and Frequency of Mutations Made in a Single Cycle of HIV-1 Replication." *Journal of Virology* 84.19, pp. 9864–9878. DOI: 10.1128/jvi.00915-10.

Abraham, M. J., Murtola, T., Schulz, R., P'all, S., Smith, J. C., Hess, B., and Lindahl, E. (2015). "GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism from Laptops to Supercomputers." *SoftwareX* 1–2, pp. 19–25. DOI: 10.1016/j.softx.2015.06.001.

Adachi, M., Ohhara, T., Kurihara, K., Tamada, T., Honjo, E., Okazaki, N., Arai, S., Shoyama, Y., Kimura, K., Matsumura, H., Sugiyama, S., Adachi, H., Takano, K., Mori, Y., Hidaka, K., Kimura, T., Hayashi, Y., Kiso, Y., and Kuroki, R. (2009). "Structure of HIV-1 Protease in Complex with Potent Inhibitor KNI-272 Determined by High-Resolution X-Ray and Neutron Crystallography." *Proceedings of the National Academy of Sciences* 106.12, pp. 4641–4646. DOI: 10.1073/pnas.0809400106.

Alcaro, S., Artese, A., Ceccherini-Silberstein, F., Ortuso, F., Perno, C. F., Sing, T., and Svicher, V. (2009). "Molecular Dynamics and Free Energy Studies on the Wild-Type and Mutated HIV-1 Protease Complexed with Four Approved Drugs: Mechanism of Binding and Drug Resistance." *Journal of Chemical Information and Modeling* 49.7, pp. 1751–1761. DOI: 10.1021/ci900012k.

Alexov, E., Mehler, E. L., Baker, N., M. Baptista, A., Huang, Y., Milletti, F., Erik Nielsen, J., Farrell, D., Carstensen, T., Olsson, M. H. M., Shen, J. K., Warwicker, J., Williams, S., and Word, J. M. (2011). "Progress in the Prediction of pKa Values in Proteins." *Proteins: Structure, Function, and Bioinformatics* 79.12, pp. 3260–3275. DOI: 10.1002/prot.23189.

Alexandropoulos, K. and Baltimore, D. (1996). "Coordinate Activation of c-Src by SH3-and SH2-Binding Sites on a Novel p130Cas-Related Protein, Sin." *Genes & Development* 10.11, pp. 1341–1355. DOI: 10.1101/gad.10.11.1341.

Allers, K., Hütter, G., Hofmann, J., Loddenkemper, C., Rieger, K., Thiel, E., and Schneider, T. (2011). "Evidence for the Cure of HIV Infection by CCR5Δ32/Δ32 Stem Cell Transplantation." *Blood* 117.10, pp. 2791–2799. DOI: 10.1182/blood-2010-09-309591.

Altman, M. D., Ali, A., Kumar Reddy, G. S. K., Nalam, M. N. L., Anjum, S. G., Cao, H., Chellappan, S., Kairys, V., Fernandes, M. X., Gilson, M. K., Schiffer, C. A., Rana, T. M., and Tidor, B. (2008). "HIV-1 Protease Inhibitors from Inverse Design in the Substrate Envelope Exhibit Subnanomolar Binding to Drug-Resistant Variants." *Journal of the American Chemical Society* 130.19, pp. 6099–6113. DOI: 10.1021/ja076558p.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215.3, pp. 403–410. DOI: 10.1006/jmbi.1990.9999.

Anderson, W. A. and Freeman, R. (1962). "Influence of a Second Radiofrequency Field on High-Resolution Nuclear Magnetic Resonance Spectra." *The Journal of Chemical Physics* 37.1, pp. 85–103. DOI: 10.1063/1.1732980.

Antiretroviral Therapy Cohort Collaboration (2017). "Survival of HIV-Positive Patients Starting Antiretroviral Therapy Between 1996 and 2013: A Collaborative Analysis of Cohort Studies." *The Lancet HIV* 4.8, e349–e356. DOI: 10.1016/s2352-3018(17)30066-8.

Arakawa, H. (2004). "Netrin-1 and Its Receptors in Tumorigenesis." *Nature Reviews Cancer* 4.12, pp. 978–987. DOI: 10.1038/nrc1504.

Baldwin, E. T., Bhat, T. N., Gulnik, S., Liu, B., Topol, I. A., Kiso, Y., Mimoto, T., Mitsuya, H., and Erickson, J. W. (1995). "Structure of HIV-1 Protease with KNI-272, a Tight-Binding Transition-State Analog Containing Allophenylnorstatine." *Structure* 3.6, pp. 581–590. DOI: 10.1016/s0969-2126(01)00192-7.

Bandyopadhyay, S., Chiang, C.-y., Srivastava, J., Gersten, M., White, S., Bell, R., Kurschner, C., Martin, C. H., Smoot, M., Sahasrabudhe, S., Barber, D. L., Chanda, S. K., and Ideker, T. (2010). "A Human MAP Kinase Interactome." *Nature Methods* 7.10, pp. 801–805. DOI: 10.1038/nmeth.1506.

Barraud, P., Paillart, J.-C., Marquet, R., and Tisné, C. (2008). "Advances in the Structural Understanding of Vif Proteins." *Current HIV research* 6.2, pp. 91–99. DOI: 10.2174/157016208783885056.

Barouch, D. H. *et al.* (2018). "Evaluation of a Mosaic HIV-1 Vaccine in a Multicentre, Randomised, Double-Blind, Placebo-Controlled, Phase 1/2a Clinical Trial (APPROACH) and in Rhesus Monkeys (NHP 13-19)." *The Lancet.* DOI: 10.1016/S0140-6736(18)31364-3.

Bardwell, L. (2006). "Mechanisms of MAPK Signalling Specificity." *Biochemical Society Transactions* 34.5, pp. 837–841. DOI: 10.1042/bst0340837.

Bastys, T., Gapsys, V., Doncheva, N. T., Kaiser, R., Groot, B. L. de, and Kalinina, O. V. (2018). "Consistent Prediction of Mutation Effect on Drug Binding in HIV-1 Protease Using Alchemical Calculations." *Journal of Chemical Theory and Computation* 14.7, pp. 3397–3408. DOI: 10.1021/acs.jctc.7b01109.

Bashford, D. and Karplus, M. (1990). "pKa's of Ionizable Groups in Proteins: Atomic Detail from a Continuum Electrostatic Model." *Biochemistry* 29.44, pp. 10219–10225. DOI: 10.1021/bi00496a010.

Bastys, T. (2012). "MAP Kinase Docking Motifs in HIV Proteins." Master's thesis. Universität des Saarlandes.

Bayly, C. I., Cieplak, P., Cornell, W. D., and Kollman, P. A. (1993). "A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model." *The Journal of Physical Chemistry* 97.40, pp. 10269–10280. DOI: 10.1021/j100142a004.

Beer (1852). "Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten." *Annalen der Physik* 162.5, pp. 78–88. DOI: 10.1002/andp.18521620505.

Bennett, C. H. (1976). "Efficient Estimation of Free Energy Differences from Monte Carlo Data." *Journal of Computational Physics* 22.2, pp. 245–268. DOI: 10.1016/0021-9991(76)90078-4.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I., and Bourne, P. E. (2000). "The Protein Data Bank." *Nucleic Acids Research* 28.1, pp. 235–242. DOI: 10.1093/nar/28.1.235.

Bhattacharyya, R. P., Reményi, A., Yeh, B. J., and Lim, W. A. (2006). "Domains, Motifs, and Scaffolds: The Role of Modular Interactions in the Evolution and Wiring of Cell Signaling Circuits." *Annual Review of Biochemistry* 75.1, pp. 655–680. DOI: 10.1146/annurev.biochem.75.103004.142710.

Biondi, R. M. and Nebreda, A. R. (2003). "Signalling Specificity of Ser/Thr Protein Kinases Through Docking-Site-Mediated Interactions." *Biochemical Journal* 372.1, pp. 1–13. DOI: 10.1042/bj20021641.

Born, M. and Oppenheimer, R. (1927). "Zur Quantentheorie der Molekeln." *Annalen der Physik* 389.20, pp. 457–484. DOI: 10.1002/andp.19273892002.

Bour, S., Schubert, U., and Strebel, K. (1995). "The Human Immunodeficiency Virus Type 1 Vpu Protein Specifically Binds to the Cytoplasmic Domain of CD4: Implications for the Mechanism of Degradation." *Journal of Virology* 69.3, pp. 1510–1520.

Bower, M. J., Cohen, F. E., and Dunbrack, R. L. (1997). "Prediction of Protein Side-Chain Rotamers from a Backbone-Dependent Rotamer Library: a new Homology Modeling Tool." *Journal of Molecular Biology* 267.5, pp. 1268–1282. DOI: 10.1006/jmbi.1997.0926.

Bragg, W. H. and Bragg, W. L. (1913). "The Reflection of X-rays by Crystals." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 88.605, pp. 428–438. DOI: 10.1098/rspa.1913.0040.

Brown, J. P., Couillard-Després, S., Cooper-Kuhn, C. M., Winkler, J., Aigner, L., and Kuhn, H. G. (2003). "Transient Expression of Doublecortin During Adult Neurogenesis." *Journal of Comparative Neurology* 467.1, pp. 1–10. DOI: 10.1002/cne.10874.

Bussi, G., Donadio, D., and Parrinello, M. (2007). "Canonical Sampling Through Velocity Rescaling." *The Journal of Chemical Physics* 126.1, p. 014101. DOI: 10.1063/1.2408420.

Cai, Y. and Schiffer, C. A. (2010). "Decomposing the Energetic Impact of Drug Resistant Mutations in HIV-1 Protease on Binding DRV." *Journal of Chemical Theory and Computation* 6.4, pp. 1358–1368. DOI: 10.1021/ct9004678.

Carlson, S. M., Chouinard, C. R., Labadorf, A., Lam, C. J., Schmelzle, K., Fraenkel, E., and White, F. M. (2011). "Large-Scale Discovery of ERK2 Substrates Identifies ERK-Mediated Transcriptional Regulation by ETV3." *Science Signaling* 4.196, rs11. DOI: 10.1126/scisignal.2002010.

Cartier, C., Deckert, M., Grangeasse, C., Trauger, R., Jensen, F., Bernard, A., Cozzone, A., Desgranges, C., and Boyer, V. (1997). “Association of ERK2 Mitogen-Activated Protein Kinase with Human Immunodeficiency Virus Particles.” *Journal of Virology* 71.6, pp. 4832–4837.

Chang, C.-I., Xu, B.-e., Akella, R., Cobb, M. H., and Goldsmith, E. J. (2002). “Crystal Structures of MAP Kinase p38 Complexed to the Docking Sites on Its Nuclear Substrate MEF2A and Activator MKK3b.” *Molecular cell* 9.6, pp. 1241–1249. DOI: 10.1016/s1097-2765(02)00525-7.

Champenois, K., Baras, A., Choisy, P., Ajana, F., Melliez, H., Bocket, L., and Yazdanpanah, Y. (2011). “Lopinavir/ritonavir resistance in patients infected with HIV-1: two divergent resistance pathways?” *Journal of Medical Virology* 83.10, pp. 1677–1681. DOI: 10.1002/jmv.22161.

Chang, M. W. and Torbett, B. E. (2011). “Accessory Mutations Maintain Stability in Drug-Resistant HIV-1 Protease.” en. *Journal of Molecular Biology* 410.4, pp. 756–760. DOI: 10.1016/j.jmb.2011.03.038.

Cha, S. (1975). “Tight-Binding Inhibitors-I: Kinetic Behavior.” *Biochemical Pharmacology* 24.23, pp. 2177–2185. DOI: 10.1016/0006-2952(75)90050-7.

Chelli, R., Marsili, S., Barducci, A., and Procacci, P. (2007). “Recovering the Crooks Equation for Dynamical Systems in the Isothermal-Isobaric Ensemble: A Strategy Based on the Equations of Motion.” *The Journal of Chemical Physics* 126.4, p. 044502. DOI: 10.1063/1.2424940.

Chen, J., Yang, M., Hu, G., Shi, S., Yi, C., and Zhang, Q. (2009). “Insights into the Functional Role of Protonation States in the HIV-1 Protease-BEA369 Complex: Molecular Dynamics Simulations and Free Energy Calculations.” *Journal of Molecular Modeling* 15.10, pp. 1245–1252. DOI: 10.1007/s00894-009-0452-y.

Chen, J., Zhang, S., Liu, X., and Zhang, Q. (2010). “Insights into Drug Resistance of Mutations D30N and I50V to HIV-1 Protease Inhibitor TMC-114: Free Energy Calculation and Molecular Dynamic Simulation.” *Journal of Molecular Modeling* 16.3, pp. 459–468. DOI: 10.1007/s00894-009-0553-7.

Chera, S., Ghila, L., Wenger, Y., and Galliot, B. (2011). “Injury-Induced Activation of the MAPK/CREB Pathway Triggers Apoptosis-Induced Compensatory Proliferation in Hydra Head Regeneration.” *Development, Growth & Differentiation* 53.2, pp. 186–201. DOI: 10.1111/j.1440-169x.2011.01250.x.

Chen, W.-K., Yeap, Y. Y., and Bogoyevitch, M. A. (2014). “The JNK1/JNK3 Interactome–Contributions by the JNK3 Unique N-Terminus and JNK Common Docking Site Residues.” *Biochemical and Biophysical Research Communications* 453.3, pp. 576–581. DOI: 10.1016/j.bbrc.2014.09.122.

Cheng, Y. and Prusoff, W. H. (1973). “Relationship Between the Inhibition Constant (KI) and the Concentration of Inhibitor Which Causes 50 per Cent Inhibition (I50) of an Enzymatic Reaction.” *Biochemical Pharmacology* 22.23, pp. 3099–3108. DOI: 10.1016/0006-2952(73)90196-2.

Chen, X. and Tropsha, A. (1995). “Relative Binding Free Energies of Peptide Inhibitors of HIV-1 Protease: The Influence of the Active Site Protonation State.” *Journal of Medicinal Chemistry* 38.1, pp. 42–48. DOI: 10.1021/jm00001a009.

Chinea, G., Padron, G., Hooft, R. W. W., Sander, C., and Vriend, G. (1995). “The Use of Position-Specific Rotamers in Model Building by Homology.” *Proteins: Structure, Function, and Bioinformatics* 23.3, pp. 415–421. DOI: 10.1002/prot.340230315.

Cihlar, T. and Ray, A. S. (2010). “Nucleoside and Nucleotide HIV Reverse Transcriptase Inhibitors: 25 Years After Zidovudine.” *Antiviral Research* 85.1. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects, pp. 39–58. DOI: 10.1016/j.antiviral.2009.09.014.

Colonno, R., Rose, R., McLaren, C., Thiry, A., Parkin, N., and Friborg, J. (2004). “Identification of I50L as the Signature Atazanavir (ATV)-Resistance Mutation in Treatment-Naive HIV-1-Infected Patients Receiving ATV-Containing Regimens.” *The Journal of Infectious Diseases* 189.10, pp. 1802–1810. DOI: 10.1086/386291.

Copeland, R. A. (2000). *Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis*. John Wiley & Sons. ISBN: 0-471-22063-9.

Copeland, R. A. (2013). *Evaluation of Enzyme Inhibitors in Drug Discovery: A Guide for Medicinal Chemists and Pharmacologists*. John Wiley & Sons. ISBN: 978-1-118-48813-3.

Coulombe, P. and Meloche, S. (2007). “Atypical Mitogen-Activated Protein Kinases: Structure, Regulation and Functions.” *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1773.8. Mitogen-Activated Protein Kinases: New Insights on Regulation, Function and Role in Human Disease, pp. 1376–1387. DOI: 10.1016/j.bbamcr.2006.11.001.

Courcelles, M., Frémin, C., Voisin, L., Lemieux, S., Meloche, S., and Thibault, P. (2014). "Phosphoproteome Dynamics Reveal Novel ERK1/2 MAP Kinase Substrates with Broad Spectrum of Functions." *Molecular Systems Biology* 9.1, p. 669. DOI: 10.1038/msb.2013.25.

Crooks, G. E. (1998). "Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems." *Journal of Statistical Physics* 90.5, pp. 1481–1487. DOI: 10.1023/A:1023208217925.

Crooks, G. E. (1999). "Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences." *Physical Review E* 60 (3), pp. 2721–2726. DOI: 10.1103/PhysRevE.60.2721.

Cusick, M. E., Klitgord, N., Vidal, M., and Hill, D. E. (2005). "Interactome: Gateway into Systems Biology." *Human Molecular Genetics* 14.suppl_2, R171–R181. DOI: 10.1093/hmg/ddi335.

Dajani, R., Fraser, E., Roe, S. M., Yeo, M., Good, V. M., Thompson, V., Dale, T. C., and Pearl, L. H. (2003). "Structural Basis for Recruitment of Glycogen Synthase Kinase 3β to the Axin-APC Scaffold Complex." *The EMBO Journal* 22.3, pp. 494–501. DOI: 10.1093/emboj/cdg068.

Dajas-Bailador, F., Jones, E. V., and Whitmarsh, A. J. (2008). "The JIP1 Scaffold Protein Regulates Axonal Development in Cortical Neurons." *Current Biology* 18.3, pp. 221–226. DOI: 10.1016/j.cub.2008.01.025.

D'arc, M., Ayouba, A., Esteban, A., Learn, G. H., Boué, V., Liegeois, F., Etienne, L., Tagg, N., Leendertz, F. H., Boesch, C., Madinda, N. F., Robbins, M. M., Gray, M., Cournil, A., Ooms, M., Letko, M., Simon, V. A., Sharp, P. M., Hahn, B. H., Delaporte, E., Mpoudi Ngole, E., and Peeters, M. (2015). "Origin of the HIV-1 Group O Epidemic in Western Lowland Gorillas." *Proceedings of the National Academy of Sciences* 112.11, E1343–1352. DOI: 10.1073/pnas.1502022112.

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T. J. (2012). "Attributes of Short Linear Motifs." *Molecular BioSystems* 8 (1), pp. 268–281. DOI: 10.1039/c1mb05231d.

De Nicola, G. F., Martin, E. D., Chaikuad, A., Bassi, R., Clark, J., Martino, L., Verma, S., Sicard, P., Tata, R., Atkinson, R. A., Knapp, S., Conte, M. R., and Marber, M. S. (2013). "Mechanism and Consequence of the Autoactivation of p38α Mitogen-Activated Protein Kinase Promoted by TAB1." *Nature Structural & Molecular Biology* 20.10, pp. 1182–1190. DOI: 10.1038/nsmb.2668.

Deng, N., Forli, S., He, P., Perryman, A., Wickstrom, L., Vijayan, R. S. K., Tiefenbrunn, T., Stout, D., Gallicchio, E., Olson, A. J., and Levy, R. M. (2015). "Distinguishing Binders from False Positives by Free Energy Calculations: Fragment Screening Against the Flap Site of HIV Protease." *The Journal of Physical Chemistry B* 119.3, pp. 976–988. DOI: 10.1021/jp506376z.

Deribe, Y. L., Pawson, T., and Dikic, I. (2010). "Post-Translational Modifications in Signal Integration." *Nature Structural & Molecular Biology* 17.6, pp. 666–672. DOI: 10.1038/nsmb.1842.

Derijard, B., Raingeaud, J., Barrett, T., Wu, I.-H., Han, J., Ulevitch, R. J., and Davis, R. J. (1995). "Independent Human MAP-Kinase Signal Transduction Pathways Defined by MEK and MKK Isoforms." *Science* 267.5198, pp. 682–685. DOI: 10.1126/science.7839144.

Dhanasekaran, N. and Reddy, E. P. (1998). "Signaling by Dual Specificity Kinases." *Oncogene* 17.11, pp. 1447–1455. DOI: 10.1038/sj.onc.1202251.

Dinkel, H., Van Roey, K., Michael, S., Davey, N. E., Weatheritt, R. J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kubań, M., Strumillo, M., Uyar, B., Budd, A., Altenberg, B., Seiler, M., Chemes, L. B., Glavina, J., Sánchez, I. E., Diella, F., and Gibson, T. J. (2013). "The Eukaryotic Linear Motif Resource ELM: 10 Years and Counting." *Nucleic Acids Research* 42.D1, pp. D259–D266. DOI: 10.1093/nar/gkt1047.

Dochi, T., Nakano, T., Inoue, M., Takamune, N., Shoji, S., Sano, K., and Misumi, S. (2014). "Phosphorylation of Human Immunodeficiency Virus Type 1 Capsid Protein at Serine 16, Required for Peptidyl-Prolyl Isomerase-Dependent Uncoating, Is Mediated by Virion-Incorporated Extracellular Signal-Regulated Kinase 2." *Journal of General Virology* 95.5, pp. 1156–1166. DOI: 10.1099/vir.0.060053-0.

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). "The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins." *Journal of Molecular Biology* 347.4, pp. 827–839. DOI: 10.1016/j.jmb.2005.01.071.

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. (2003). "A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations." *Journal of Computational Chemistry* 24.16, pp. 1999–2012. DOI: 10.1002/jcc.10349.

Duan, R., Lazim, R., and Zhang, D. (2015). “Understanding the Basis of I50V-Induced Affinity Decrease in HIV-1 Protease via Molecular Dynamics Simulations Using Polarized Force Field.” *Journal of Computational Chemistry* 36.25, pp. 1885–1892. DOI: 10.1002/jcc.24020.

Duch, A., Nadal, E. de, and Posas, F. (2012). “The p38 and Hog1 SAPKs Control Cell Cycle Progression in Response to Environmental Stresses.” *FEBS Letters* 586.18, pp. 2925–2931. DOI: 10.1016/j.febslet.2012.07.034.

Dunbrack, R. L. and Karplus, M. (1993). “Backbone-Dependent Rotamer Library for Proteins Application to Side-Chain Prediction.” *Journal of Molecular Biology* 230.2, pp. 543–574. DOI: 10.1006/jmbi.1993.1170.

Enslen, H., Brancho, D. M., and Davis, R. J. (2000). “Molecular Determinants That Mediate Selective Activation of p38 MAP Kinase Isoforms.” *The EMBO Journal* 19.6, pp. 1301–1311. DOI: 10.1093/emboj/19.6.1301.

Ermolieff, J., Lin, X., and Tang, J. (1997). “Kinetic Properties of Saquinavir-Resistant Mutants of Human Immunodeficiency Virus Type 1 Protease and Their Implications in Drug Resistance in Vivo.” *Biochemistry* 36.40, pp. 12364–12370. DOI: 10.1021/bi971072e.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). “A Smooth Particle Mesh Ewald Method.” *The Journal of Chemical Physics* 103.19, pp. 8577–8593. DOI: 10.1063/1.470117.

Fantz, D. A., Jacobs, D., Glossip, D., and Kornfeld, K. (2001). “Docking Sites on Substrate Proteins Direct Extracellular Signal-regulated Kinase to Phosphorylate Specific Residues.” *Journal of Biological Chemistry* 276.29, pp. 27256–27265. DOI: 10.1074/jbc.m102512200.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O. G., and Lemey, P. (2014). “The Early Spread and Epidemic Ignition of HIV-1 in Human Populations.” *Science* 346.6205, pp. 56–61. DOI: 10.1126/science.1256739.

Fernandes, N., Bailey, D. E., VanVranken, D. L., and Allbritton, N. L. (2007). “Use of Docking Peptides to Design Modular Substrates with High Efficiency for Mitogen-Activated Protein Kinase Extracellular Signal-Regulated Kinase.” *ACS Chemical Biology* 2.10, pp. 665–673. DOI: 10.1021/cb700158q.

Ferreon, J. C., Lee, C. W., Arai, M., Martinez-Yamout, M. A., Dyson, H. J., and Wright, P. E. (2009). “Cooperative Regulation of p53 by Modulation of Ternary Complex Formation with CBP/p300 and HDM2.” *Proceedings of the National Academy of Sciences* 106.16, pp. 6591–6596. DOI: 10.1073/pnas.0811023106.

Ferguson, D. M., Radmer, R. J., and Kollman, P. A. (1991). “Determination of the Relative Binding Free Energies of Peptide Inhibitors to the HIV-1 Protease.” *Journal of Medicinal Chemistry* 34.8, pp. 2654–2659. DOI: 10.1021/jm00112a048.

Feyfant, E., Sali, A., and Fiser, A. (2007). “Modeling Mutations in Protein Structures.” *Protein Science* 16.9, pp. 2030–2041. DOI: 10.1110/ps.072855507.

Finck, B. N. and Kelly, D. P. (2006). “PGC-1 Coactivators: Inducible Regulators of Energy Metabolism in Health and Disease.” *Journal of Clinical Investigation* 116.3, pp. 615–622. DOI: 10.1172/jci27794.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014). “Pfam: the protein families database.” *Nucleic Acids Research* 42.D1, pp. D222–D230. DOI: 10.1093/nar/gkt1223.

Flotho, A., Simpson, D. M., Qi, M., and Elion, E. A. (2004). “Localized Feedback Phosphorylation of Ste5p Scaffold by Associated MAPK Cascade.” *Journal of Biological Chemistry* 279.45, pp. 47391–47401. DOI: 10.1074/jbc.m405681200.

Frankel, A. D. and Young, J. A. (1998). “HIV-1: Fifteen Proteins and an RNA.” *Annual Review of Biochemistry* 67.3, pp. 1–25. DOI: 10.1146/annurev.biochem.67.1.1.

Freschi, L., Osseni, M., and Landry, C. R. (2014). “Functional Divergence and Evolutionary Turnover in Mammalian Phosphoproteomes.” *PLOS Genetics* 10.1, pp. 1–13. DOI: 10.1371/journal.pgen.1004062.

Frisch, M. J. *et al.* (2010). *Gaussian09 Revision C.01*. Gaussian Inc.: Wallingford, CT.

Fuhs, S. R. and Hunter, T. (2017). “pHisphorylation: the Emergence of Histidine Phosphorylation as a Reversible Regulatory Modification.” *Current Opinion in Cell Biology* 45, pp. 8–16. DOI: 10.1016/j.ceb.2016.12.010.

Gallay, P., Swingler, S., Song, J., Bushman, F., and Trono, D. (1995). “HIV Nuclear Import is Governed by the Phosphotyrosine-Mediated Binding of Matrix to the Core Domain of Integrase.” *Cell* 83.4, pp. 569–576. DOI: 10.1016/0092-8674(95)90097-7.

Ganiatsas, S., Kwee, L., Fujiwara, Y., Perkins, A., Ikeda, T., Labow, M. A., and Zon, L. I. (1998). "SEK1 Deficiency Reveals Mitogen-Activated Protein Kinase Cascade Crossregulation and Leads to Abnormal Hepatogenesis." *Proceedings of the National Academy of Sciences* 95.12, pp. 6881–6886. DOI: 10.1073/pnas.95.12.6881.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M., and Hahn, B. H. (1999). "Origin of HIV-1 in the Chimpanzee Pan troglodytes troglodytes." *Nature* 397.6718, pp. 436–441. DOI: 10.1038/17130.

Gapsys, V., Seeliger, D., and Groot, B. L. de (2012). "New Soft-Core Potential Function for Molecular Dynamics Based Alchemical Free Energy Calculations." *Journal of Chemical Theory and Computation* 8.7, pp. 2373–2382. DOI: 10.1021/ct300220p.

Gapsys, V. and Groot, B. L. de (2013). "Optimal Superpositioning of Flexible Molecule Ensembles." *Biophysical Journal* 104.1, pp. 196–207. DOI: 10.1016/j.bpj.2012.11.003.

Gapsys, V., Michielssens, S., Seeliger, D., and Groot, B. L. de (2014). "pmx: Automated Protein Structure and Topology Generation for Alchemical Perturbations." *Journal of Computational Chemistry* 36.5, pp. 348–354. DOI: 10.1002/jcc.23804.

Gapsys, V., Michielssens, S., Peters, J. H., Groot, B. L. de, and Leonov, H. (2015). "Calculation of Binding Free Energies." *Molecular Modeling of Proteins*. Ed. by A. Kukol. New York, NY: Springer New York, pp. 173–209. ISBN: 978-1-4939-1465-4. DOI: 10.1007/978-1-4939-1465-4_9.

Gapsys, V., Michielssens, S., Seeliger, D., and de Groot, B. L. (2016). "Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan." *Angewandte Chemie International Edition* 55.26, pp. 7364–7368. DOI: 10.1002/anie.201510054.

Garai, Á., Zeke, A., Gógl, G., Törő, I., Fördős, F., Blankenburg, H., Bárkai, T., Varga, J., Alexa, A., Emig, D., Albrecht, M., and Reményi, A. (2012). "Specificity of Linear Motifs That Bind to a Common Mitogen-Activated Protein Kinase Docking Groove." *Science Signaling* 5.245, ra74. DOI: 10.1126/scisignal.2003004.

Genheden, S. and Ryde, U. (2015). "The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities." *Expert Opinion on Drug Discovery* 10.5, pp. 449–461. DOI: 10.1517/17460441.2015.1032936.

Gerlits, O., Wymore, T., Das, A., Shen, C.-H., Parks, J. M., Smith, J. C., Weiss, K. L., Keen, D. A., Blakeley, M. P., Louis, J. M., Langan, P., Weber, I. T., and Kovalevsky, A. (2016). "Long-Range Electrostatics-Induced Two-Proton Transfer Captured by Neutron Crystallography in an Enzyme Catalytic Site." *Angewandte Chemie International Edition* 55.16, pp. 4924–4927. DOI: 10.1002/anie.201509989.

Gerstein, M., Sonnhammer, E. L., and Chothia, C. (1994). "Volume Changes in Protein Evolution." *Journal of Molecular Biology* 236.4, pp. 1067–1078. DOI: 10.1016/0022-2836(94)90012-4.

Giroud, C., Chazal, N., and Briant, L. (2011). "Cellular Kinases Incorporated into HIV-1 Particles: Passive or Active Passengers?" *Retrovirology* 8.1, p. 71. DOI: 10.1186/1742-4690-8-71.

Glatz, G., Gogl, G., Alexa, A., and Remenyi, A. (2013). "Structural Mechanism for the Specific Assembly and Activation of the Extracellular Signal Regulated Kinase 5 (ERK5) Module." *Journal of Biological Chemistry* 288.12, pp. 8596–8609. DOI: 10.1074/jbc.m113.452235.

Good, M., Tang, G., Singleton, J., Reményi, A., and Lim, W. A. (2009). "The Ste5 Scaffold Directs Mating Signaling by Catalytically Unlocking the Fus3 MAP Kinase for Activation." *Cell* 136.6, pp. 1085–1097. DOI: 10.1016/j.cell.2009.01.049.

Gordon, E. A., Whisenant, T. C., Zeller, M., Kaake, R. M., Gordon, W. M., Krotee, P., Patel, V., Huang, L., Baldi, P., and Bardwell, L. (2013). "Combining Docking Site and Phosphosite Predictions to Find New Substrates: Identification of Smoothelin-like-2 (SMTNL2) as a c-Jun N-terminal Kinase (JNK) Substrate." *Cellular Signalling* 25.12, pp. 2518–2529. DOI: 10.1016/j.cellsig.2013.08.004.

Grewal, S., Molina, D., and Bardwell, L. (2006). "Mitogen-Activated Protein Kinase (MAPK)-Docking Sites in MAPK Kinases Function as Tethers That Are Crucial for MAPK Regulation in vivo." *Cellular Signalling* 18.1, pp. 123–134. DOI: 10.1016/j.cellsig.2005.04.001.

Greenway, A., Azad, A., Mills, J., and McPhee, D. (1996). "Human Immunodeficiency Virus Type 1 Nef Binds Directly to Lck and Mitogen-Activated Protein Kinase, Inhibiting Kinase Activity." *Journal of Virology* 70.10, pp. 6701–6708.

Gsponer, J., Futschik, M. E., Teichmann, S. A., and Babu, M. M. (2008). "Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation." *Science* 322.5906, pp. 1365–1368. DOI: 10.1126/science.1163581.

Gupta, P., Singhal, P. K., Rajendrakumar, P., Padwad, Y., Tendulkar, A. V., Kalyanaraman, V., Schmidt, R. E., Srinivasan, A., and Mahalingam, S. (2011). "Mechanism of Host Cell MAPK/ERK-2 Incorporation into Lentivirus Particles: Characterization of the Interaction Between MAPK/ERK-2 and Proline-Rich-Domain Containing Capsid Region of Structural Protein Gag." *Journal of Molecular Biology* 410.4, pp. 681–697. DOI: 10.1016/j.jmb.2011.03.022.

Guvench, O. and MacKerell, A. D. (2008). "Comparison of Protein Force Fields for Molecular Dynamics Simulations." *Molecular Modeling of Proteins*. Ed. by A. Kukol. Totowa, NJ: Humana Press, pp. 63–88. ISBN: 978-1-59745-177-2. DOI: 10.1007/978-1-59745-177-2_4.

Hahn, M. W. and Wray, G. A. (2002). "The G-Value Paradox." *Evolution and Development* 4.2, pp. 73–75. DOI: 10.1046/j.1525-142X.2002.01069.x.

Hemelaar, J. (2012). "The Origin and Diversity of the HIV-1 Pandemic." *Trends in Molecular Medicine* 18.3, pp. 182–192. DOI: 10.1016/j.molmed.2011.12.001.

Henriet, S., Mercenne, G., Bernacchi, S., Paillart, J.-C., and Marquet, R. (2009). "Tumultuous Relationship Between the Human Immunodeficiency Virus Type 1 Viral Infectivity Factor (Vif) and the Human APOBEC-3G and APOBEC-3F Restriction Factors." *Microbiology and Molecular Biology Reviews* 73.2, pp. 211–232. DOI: 10.1128/mmbr.00040-08.

Henderson, G. J., Lee, S.-K., Irlbeck, D. M., Harris, J., Kline, M., Pollom, E., Parkin, N., and Swanstrom, R. (2012). "Interplay Between Single Resistance-Associated Mutations in the HIV-1 Protease and Viral Infectivity, Protease Activity, and Inhibitor Sensitivity." *Antimicrobial Agents and Chemotherapy* 56.2, pp. 623–633. DOI: 10.1128/aac.05549-11.

Heo, Y.-S., Kim, S.-K., Seo, C. I., Kim, Y. K., Sung, B.-J., Lee, H. S., Lee, J. I., Park, S.-Y., Kim, J. H., Hwang, K. Y., Hyun, Y. L., Jeon, Y. H., Ro, S., Cho, J. M., Lee, T. G., and Yang, C. H. (2004). "Structural Basis for the Selective Inhibition of JNK1 by the Scaffolding Protein JIP1 and SP600125." *The EMBO Journal* 23.11, pp. 2185–2195. DOI: 10.1038/sj.emboj.7600212.

Hess, B., Kutzner, C., Spoel, D. van der, and Lindahl, E. (2008). "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation." *Journal of Chemical Theory and Computation* 4.3, pp. 435–447. DOI: 10.1021/ct700301q.

Hess, B., Bekker, H., Berendsen, H. J., and Fraaije, J. G. (1997). "LINCS: A Linear Constraint Solver for Molecular Simulations." *Journal of Computational Chemistry* 18.12, pp. 1463–1472. DOI: 10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h.

Hirosumi, J., Tuncman, G., Chang, L., Görgün, C. Z., Uysal, K. T., Maeda, K., Karin, M., and Hotamisligil, G. S. (2002). "A Central Role for JNK in Obesity and Insulin Resistance." *Nature* 420.6913, pp. 333–336. DOI: 10.1038/nature01137.

Hirsch, V. M., Olmsted, R. A., Murphey-Corb, M., Purcell, R. H., and Johnson, P. R. (1989). "An African Primate Lentivirus (SIVsm) Closely Related to HIV-2." *Nature* 339.6223, pp. 389–392. DOI: 10.1038/339389a0.

Ho, D. D., Toyoshima, T., Mo, H., Kempf, D. J., Norbeck, D., Chen, C.-M., Wideburg, N. E., Burt, S. K., Erickson, J. W., and Singh, M. K. (1994). "Characterization of Human Immunodeficiency Virus Type 1 Variants with Increased Resistance to a C2-Symmetric Protease Inhibitor." *Journal of Virology* 68.3, pp. 2016–2020.

Hoffman, N. G., Schiffer, C. A., and Swanstrom, R. (2003). "Covariation of Amino Acid Positions in HIV-1 Protease." *Virology* 314.2, pp. 536–548. DOI: 10.1016/S0042-6822(03)00484-7.

Hong, L., Zhang, X. C., Hartsuck, J. A., and Tang, J. (2000). "Crystal Structure of an in Vivo HIV-1 Protease Mutant in Complex with Saquinavir: Insights into the Mechanisms of Drug Resistance." *Protein Science* 9.10, pp. 1898–1904. DOI: 10.1110/ps.9.10.1898.

Hope, T. J. (1999). "The Ins and Outs of HIV Rev." *Archives of Biochemistry and Biophysics* 365.2, pp. 186–191. DOI: 10.1006/abbi.1999.1207.

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., and Nakai, K. (2007). "WoLF PSORT: Protein Localization Predictor." *Nucleic Acids Research* 35.Web Server, W585–W587. DOI: 10.1093/nar/gkm259.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015). "PhosphoSitePlus, 2014: mutations, PTMs and Recalibrations." *Nucleic Acids Research* 43.D1, pp. D512–D520. DOI: 10.1093/nar/gku1267.

Hou, T. and Yu, R. (2007). "Molecular Dynamics and Free Energy Studies on the Wild-type and Double Mutant HIV-1 Protease Complexed with Amprenavir and Two Amprenavir-Related Inhibitors: Mechanism for Binding and Drug Resistance." *Journal of Medicinal Chemistry* 50.6, pp. 1177–1188. DOI: 10.1021/jm0609162.

Hou, T., McLaughlin, W. A., and Wang, W. (2008). "Evaluating the Potency of HIV-1 Protease Drugs to Combat Resistance." *Proteins: Structure, Function, and Bioinformatics* 71.3, pp. 1163–1174. DOI: 10.1002/prot.21808.

Hou, T., Zhang, W., Wang, J., and Wang, W. (2009). "Predicting Drug Resistance of the HIV-1 Protease Using Molecular Interaction Energy Components." *Proteins: Structure, Function, and Bioinformatics* 74.4, pp. 837–846. DOI: 10.1002/prot.22192.

Hou, T., Wang, J., Li, Y., and Wang, W. (2011). "Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations." *Journal of Chemical Information and Modeling* 51.1, pp. 69–82. DOI: 10.1021/ci100275a.

Hu, G.-D., Zhu, T., Zhang, S.-L., Wang, D., and Zhang, Q.-G. (2010). "Some Insights into Mechanism for Binding and Drug Resistance of Wild Type and I50V V82A and I84V Mutations in HIV-1 Protease with GRL-98065 Inhibitor from Molecular Dynamic Simulations." *European Journal of Medicinal Chemistry* 45.1, pp. 227–235. DOI: 10.1016/j.ejmech.2009.09.048.

Huang, J., Sun, B., Yao, Y., and Liu, J. (2017). "Fast and Reliable Thermodynamic Approach for Determining the Protonation State of Asp Dyad." *Journal of Chemical Information and Modeling* 57.9, pp. 2273–2280. DOI: 10.1021/acs.jcim.7b00207.

Hwang, C., Schürmann, D., Sobotha, C., Boffito, M., Sevinsky, H., Ray, N., Ravindran, P., Xiao, H., Keicher, C., Hüser, A., Krystal, M., Dicker, I. B., Grasela, D., and Lataillade, M. (2017). "Antiviral Activity, Safety, and Exposure-Response Relationships of GSK3532795, a Second-Generation Human Immunodeficiency Virus Type 1 Maturation Inhibitor, Administered as Monotherapy or in Combination with Atazanavir with or Without Ritonavir in a Phase 2a Randomized, Dose-Ranging, Controlled Trial (AI468002)." *Clinical Infectious Diseases* 65.3, pp. 442–452. DOI: 10.1093/cid/cix239.

Hyland, L. J., Tomaszek, T. A. J., and Meek, T. D. (1991). "Human Immunodeficiency Virus-1 Protease. 2. Use of pH Rate Studies and Solvent Kinetic Isotope Effects to Elucidate Details of Chemical Mechanism." *Biochemistry* 30.34, pp. 8454–8463. DOI: 10.1021/bi00098a024.

Iacob, S. A. and Iacob, D. G. (2017). "Ibalizumab Targeting CD4 Receptors, An Emerging Molecule in HIV Therapy." *Frontiers in Microbiology* 8, p. 2323. DOI: 10.3389/fmicb.2017.02323.

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004). "The Importance of Intrinsic Disorder for Protein Phosphorylation." *Nucleic Acids Research* 32.3, pp. 1037–1049. DOI: 10.1093/nar/gkh253.

Ilari, A. and Savino, C. (2008). "Protein Structure Determination by X-Ray Crystallography." *Bioinformatics: Data, Sequence Analysis and Evolution.* Ed. by J. M. Keith. Totowa, NJ: Humana Press, pp. 63–87. ISBN: 978-1-60327-159-2. DOI: 10.1007/978-1-60327-159-2_3.

International Human Genome Sequencing Consortium (2001). "Initial Sequencing and Analysis of the Human Genome." *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062.

Jacqué, J.-M., Mann, A., Enslen, H., Sharova, N., Brichacek, B., Davis, R. J., and Stevenson, M. (1998). "Modulation of HIV-1 Infectivity by MAPK, a Virion–Associated Kinase." *The EMBO Journal* 17.9, pp. 2607–2618. DOI: 10.1093/emboj/17.9.2607.

Jacobs, D., Glossip, D., Xing, H., Muslin, A. J., and Kornfeld, K. (1999). "Multiple Docking Sites on Substrate Proteins form a Modular System That Mediates Recognition by ERK MAP Kinase." *Genes & Development* 13.2, pp. 163–175. DOI: 10.1101/gad.13.2.163.

Japour, A. J., Mayers, D., Johnson, V., Kuritzkes, D., Beckett, L., Arduino, J., Lane, J., Black, R., Reichelderfer, P., and D'Aquila, R. (1993). "Standardized Peripheral Blood Mononuclear Cell Culture Assay for Determination of Drug Susceptibilities of Clinical Human Immunodeficiency Virus Type 1 Isolates. The RV-43 Study Group, the AIDS Clinical Trials Group Virology Committee Resistance Working Group." *Antimicrobial Agents and Chemotherapy* 37.5, pp. 1095–1101. DOI: 10.1128/aac.37.5.1095.

Jarzynski, C. (1997). "Nonequilibrium Equality for Free Energy Differences." *Physical Review Letters* 78 (14), pp. 2690–2693. DOI: 10.1103/PhysRevLett.78.2690.

Jensen, L. J., Julien, P., Kuhn, M., Mering, C. von, Muller, J., Doerks, T., and Bork, P. (2008). "eggNOG: Automated Construction and Annotation of Orthologous Groups of Genes." *Nucleic Acids Research* 36.Database, pp. D250–D254. DOI: 10.1093/nar/gkm796.

Jessen, H., Allen, T. M., and Streeck, H. (2014). "How a Single Patient Influenced HIV Research - 15-Year Follow-up." *New England Journal of Medicine* 370.7, pp. 682–683. DOI: 10.1056/NEJMc1308413.

Johnson, L. N. and Lewis, R. J. (2001). “Structural Basis for Control by Phosphorylation.” *Chemical Reviews* 101.8, pp. 2209–2242. DOI: 10.1021/cr000225s.

Johnson, G. L. and Lapadat, R. (2002). “Mitogen-Activated Protein Kinase Pathways Mediated by ERK, JNK, and p38 Protein Kinases.” *Science* 298.5600, pp. 1911–1912. DOI: 10.1126/science.1072682.

Johnson, S. A. and Hunter, T. (2005). “Kinomics: Methods for Deciphering the Kinome.” *Nature Methods* 2.1, pp. 17–25. DOI: 10.1038/nmeth731.

Johnson, E. C. B., Malito, E., Shen, Y., Pentelute, B., Rich, D., Florián, J., Tang, W.-J., and Kent, S. B. (2007). “Insights from Atomic-Resolution X-Ray Structures of Chemically Synthesized HIV-1 Protease in Complex with Inhibitors.” *Journal of Molecular Biology* 373.3, pp. 573–586. DOI: 10.1016/j.jmb.2007.07.054.

Joint United Nations Programme on HIV/AIDS (UNAIDS) (2017a). *Right to Health.* http://www.unaids.org/sites/default/files/media_asset/RighttoHealthReport_Full_web%20%20Nov.pdf. [Online; accessed 12-June-2018].

Joint United Nations Programme on HIV/AIDS (UNAIDS) (2017b). *UNAIDS Data 2017.* http://www.unaids.org/sites/default/files/media_asset/20170720_Data_book_2017_en.pdf.

Jorgensen, W. L. and Tirado-Rives, J. (1996). “Monte Carlo vs Molecular Dynamics for Conformational Sampling.” *The Journal of Physical Chemistry* 100.34, pp. 14508–14513. DOI: 10.1021/jp960880x.

Josefsson, L., King, M. S., Makitalo, B., Brännström, J., Shao, W., Maldarelli, F., Kearney, M. F., Hu, W.-S., Chen, J., Gaines, H., Mellors, J. W., Albert, J., Coffin, J. M., and Palmer, S. E. (2011). “Majority of CD4+ T Cells from Peripheral Blood of HIV-1–Infected Individuals Contain Only One HIV DNA Molecule.” *Proceedings of the National Academy of Sciences* 108.27, pp. 11199–11204. DOI: 10.1073/pnas.1107729108.

Josefsson, L., Palmer, S., Faria, N. R., Lemey, P., Casazza, J., Ambrozak, D., Kearney, M., Shao, W., Kottilil, S., Sneller, M., Mellors, J., Coffin, J. M., and Maldarelli, F. (2013). “Single Cell Analysis of Lymph Node Tissue from HIV-1 Infected Patients Reveals That the Majority of CD4+ T-Cells Contain One HIV-1 DNA Molecule.” *PLoS Pathogens* 9.6, e1003432. DOI: 10.1371/journal.ppat.1003432.

Joung, I. S. and Cheatham, T. E. (2008). “Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations.” *The Journal of Physical Chemistry B* 112.30, pp. 9020–9041. DOI: 10.1021/jp8001614.

Kalra, P., Reddy, T. V., and Jayaram, B. (2001). “Free Energy Component Analysis for Drug Design: A Case Study of HIV-1 Protease-Inhibitor Binding.” *Journal of Medicinal Chemistry* 44.25, pp. 4325–4338. DOI: 10.1021/jm010175z.

Käll, L., Krogh, A., and Sonnhammer, E. L. (2007). “Advantages of Combined Transmembrane Topology and Signal Peptide Prediction-The Phobius Web Server.” *Nucleic Acids Research* 35.Web Server, W429–W432. DOI: 10.1093/nar/gkm256.

Kannangai, R., David, S., and Sridharan, G. (2012). “Human Immunodeficiency Virus Type-2-A Milder, Kinder Virus: An Update.” *Indian Journal of Medical Microbiology* 30.1, pp. 6–15. DOI: 10.4103/0255-0857.93014.

Kao, S.-Y., Calman, A. F., Luciw, P. A., and Peterlin, B. M. (1987). “Anti-Termination of Transcription Within the Long Terminal Repeat of HIV-1 by tat Gene Product.” *Nature* 330.6147, pp. 489–493. DOI: 10.1038/330489a0.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). “kernlab – An S4 Package for Kernel Methods in R.” *Journal of Statistical Software* 11.9, pp. 1–20. DOI: 10.18637/jss.v011.i09.

Kar, P. and Knecht, V. (2012a). “Energetic Basis for Drug Resistance of HIV-1 Protease Mutants Against Amprenavir.” *Journal of Computer-Aided Molecular Design* 26.2, pp. 215–232. DOI: 10.1007/s10822-012-9550-5.

Kar, P. and Knecht, V. (2012b). “Origin of Decrease in Potency of Darunavir and Two Related Antiviral Inhibitors against HIV-2 Compared to HIV-1 Protease.” *The Journal of Physical Chemistry B* 116.8, pp. 2605–2614. DOI: 10.1021/jp211768n.

Keele, B. F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Ngole, E. M., Bienvenue, Y., Delaporte, E., Brookfield, J. F. Y., Sharp, P. M., Shaw, G. M., Peeters, M., and Hahn, B. H. (2006). “Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1.” *Science* 313.5786, pp. 523–526. DOI: 10.1126/science.1126531.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R., Wyckoff, H., and Phillips, D. C. (1958). “A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis.” *Nature* 181.4610, pp. 662–666. DOI: 10.1038/181662a0.

Khan, F. I., Wei, D.-Q., Gu, K.-R., Hassan, M. I., and Tabrez, S. (2016). “Current Updates on Computer Aided Protein Modeling and Designing.” *International Journal of Biological Macromolecules* 85, pp. 48–62. DOI: 10.1016/j.ijbiomac.2015.12.072.

Khoury, G. A., Baliban, R. C., and Floudas, C. A. (2011). "Proteome-Wide Post-Translational Modification Statistics: Frequency Analysis and Curation of the Swiss-Prot Database." *Scientific Reports* 1.1, p. 90. DOI: 10.1038/srep00090.

Kim, E. K. and Choi, E.-J. (2015). "Compromised MAPK Signaling in Human Diseases: An Update." *Archives of Toxicology* 89.6, pp. 867–882. DOI: 10.1007/s00204-015-1472-2.

King, N. *et al.* (2008). "The Genome of the Choanoflagellate Monosiga Brevicollis and the Origin of Metazoans." *Nature* 451.7180, pp. 783–788. DOI: 10.1038/nature06617.

Kirkwood, J. G. (1935). "Statistical Mechanics of Fluid Mixtures." *The Journal of Chemical Physics* 3.5, pp. 300–313. DOI: 10.1063/1.1749657.

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E. (2000). "Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models." *Accounts of Chemical Research* 33.12, pp. 889–897. DOI: 10.1021/ar000033j.

Kosako, H., Yamaguchi, N., Aranami, C., Ushiyama, M., Kose, S., Imamoto, N., Taniguchi, H., Nishida, E., and Hattori, S. (2009). "Phosphoproteomics Reveals New ERK MAP Kinase Targets and Links ERK to Nucleoporin-Mediated Nuclear Transport." *Nature Structural & Molecular Biology* 16.10, pp. 1026–1035. DOI: 10.1038/nsmb.1656.

Kovalevsky, A. Y., Tie, Y., Liu, F., Boross, P. I., Wang, Y. F., Leshchenko, S., Ghosh, A. K., Harrison, R. W., and Weber, I. T. (2006). "Effectiveness of Nonpeptide Clinical Inhibitor TMC-114 on HIV-1 Protease with Highly Drug resistant Mutations D30N, I50V, and L90M." *Journal of Medicinal Chemistry* 49.4, pp. 1379–1387. DOI: 10.1021/jm050943c.

Kožíšek, M., Bray, J., Řezáčová, P., Šašková, K., Brynda, J., Pokorná, J., Mammano, F., Rulíšek, L., and Konvalinka, J. (2007). "Molecular Analysis of the HIV-1 Resistance Development: Enzymatic Activities, Crystal Structures, and Thermodynamics of Nelfinavir-Resistant HIV Protease Mutants." *Journal of Molecular Biology* 374.4, pp. 1005–1016. DOI: 10.1016/j.jmb.2007.09.083.

Kräusslich, H.-G. and Wimmer, E. (1988). "Viral Proteinases." *Annual Review of Biochemistry* 57.1, pp. 701–754. DOI: 10.1146/annurev.bi.57.070188.003413.

Krivobokova, T., Briones, R., Hub, J. S., Munk, A., and Groot, B. L. de (2012). "Partial Least-Squares Functional Mode Analysis: Application to the Membrane Proteins AQP1, Aqy1, and CLC-ec1." *Biophysical Journal* 103.4, pp. 786–796. DOI: 10.1016/j.bpj.2012.07.022.

Kuhn, B., Gerber, P., Schulz-Gasch, T., and Stahl, M. (2005). "Validation and Use of the MM-PBSA Approach for Drug Discovery." *Journal of Medicinal Chemistry* 48.12, pp. 4040–4048. DOI: 10.1021/jm049081q.

Kumar, P., Chimenti, M. S., Pemble, H., Schönichen, A., Thompson, O., Jacobson, M. P., and Wittmann, T. (2012). "Multisite Phosphorylation Disrupts Arginine-Glutamate Salt Bridge Networks Required for Binding of Cytoplasmic Linker-associated Protein 2 (CLASP2) to End-binding Protein 1 (EB1)." *Journal of Biological Chemistry* 287.21, pp. 17050–17064. DOI: 10.1074/jbc.m111.316661.

Lang, T., Yu, L., Tu, Q., Jiang, J., Chen, Z., Xin, Y., Liu, G., and Zhao, S. (2000). "Molecular Cloning, Genomic Organization, and Mapping of PRKAG2, a Heart Abundant $\gamma 2$ Subunit of 5 -AMP-Activated Protein Kinase, to Human Chromosome 7q36." *Genomics* 70.2, pp. 258–263. DOI: 10.1006/geno.2000.6376.

Larder, B., Kemp, S., and Harrigan, P. (1995). "Potential Mechanism for Sustained Antiretroviral Efficacy of AZT-3TC Combination Therapy." *Science* 269.5224, pp. 696–699. DOI: 10.1126/science.7542804.

Laughlin, J. D., Nwachukwu, J. C., Figuera-Losada, M., Cherry, L., Nettles, K. W., and LoGrasso, P. V. (2012). "Structural Mechanisms of Allostery and Autoinhibition in JNK Family Kinases." *Structure* 20.12, pp. 2174–2184. DOI: 10.1016/j.str.2012.09.021.

Le Goff, C., Rogers, C., Le Goff, W., Pinto, G., Bonnet, D., Chrabieh, M., Alibeu, O., Nistchke, P., Munnich, A., Picard, C., and Cormier-Daire, V. (2016). "Heterozygous Mutations in MAP3K7, Encoding TGF-β-Activated Kinase 1, Cause Cardiospondylocarpofacial Syndrome." *The American Journal of Human Genetics* 99.2, pp. 407–413. DOI: 10.1016/j.ajhg.2016.06.005.

Lee, Y. H., Giraud, J., Davis, R. J., and White, M. F. (2003). "c-Jun N-terminal Kinase (JNK) Mediates Feedback Inhibition of the Insulin Signaling Cascade." *Journal of Biological Chemistry* 278.5, pp. 2896–2902. DOI: 10.1074/jbc.m208359200.

Lee, K. S., Park, J.-E., Kang, Y. H., Kim, T.-S., and Bang, J. K. (2014). "Mechanisms Underlying Plk1 Polo-Box Domain-Mediated Biological Processes and Their Physiological Significance." *Molecules and Cells* 37.4, pp. 286–294. DOI: 10.14348/molcells.2014.0002.

Lengauer, T., Pfeifer, N., and Kaiser, R. (2014). "Personalized HIV Therapy to Control Drug Resistance." *Drug Discovery Today: Technologies* 11. Drug Resistance, pp. 57–64. DOI: 10.1016/j.ddtec.2014.02.004.

Li, H., Robertson, A. D., and Jensen, J. H. (2005). "Very Fast Empirical Prediction and Rationalization of Protein pKa Values." *Proteins: Structure, Function, and Bioinformatics* 61.4, pp. 704–721. DOI: 10.1002/prot.20660.

Li, H., Xing, X., Ding, G., Li, Q., Wang, C., Xie, L., Zeng, R., and Li, Y. (2009). "SysPTM: A Systematic Resource for Proteomic Research on Post-translational Modifications." *Molecular & Cellular Proteomics* 8.8, pp. 1839–1849. DOI: 10.1074/mcp.M900030-MCP200.

Li, D., Han, J.-G., Chen, H., Li, L., Zhao, R.-N., Liu, G., and Duan, Y. (2012). "Insights into the Structural Function of the Complex of HIV-1 Protease with TMC-126: Molecular Dynamics Simulations and Free-Energy Calculations." *Journal of Molecular Modeling* 18.5, pp. 1841–1854. DOI: 10.1007/s00894-011-1205-2.

Li, D., Zhang, Y., Zhao, R.-N., Fan, S., and Han, J.-G. (2014). "Investigation on the Mechanism for the Binding and Drug Resistance of Wild Type and Mutations of G86 Residue in HIV-1 Protease Complexed with Darunavir by Molecular Dynamic Simulation and Free Energy Calculation." *Journal of Molecular Modeling* 20.2, p. 2122. DOI: 10.1007/s00894-014-2122-y.

Liang, S. and Grishin, N. V. (2002). "Side-Chain Modeling with an Optimized Scoring Function." *Protein Science* 11.2, pp. 322–331. DOI: 10.1110/ps.24902.

Lim, W. A., Richards, F. M., and Fox, R. O. (1994). "Structural Determinants of Peptide-Binding Orientation and of Sequence Specificity in SH3 Domains." *Nature* 372.6504, pp. 375–379. DOI: 10.1038/372375a0.

Lindahl, E. (2015). "Molecular Dynamics Simulations." *Molecular Modeling of Proteins*. Ed. by A. Kukol. New York, NY: Springer New York, pp. 3–26. ISBN: 978-1-4939-1465-4. DOI: 10.1007/978-1-4939-1465-4_1.

Liu, F., Boross, P. I., Wang, Y. F., Tozser, J., Louis, J. M., Harrison, R. W., and Weber, I. T. (2005). "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, pp. 789–800. DOI: 10.1016/j.jmb.2005.09.095.

Livingstone, C., Patel, G., and Jones, N. (1995). "ATF-2 Contains a Phosphorylation-Dependent Transcriptional Activation Domain." *The EMBO Journal* 14.8, pp. 1785–1797. DOI: 10.1002/j.1460-2075.1995.tb07167.x.

Long, Y., Sou, W. H., Yung, K. W. Y., Liu, H., Wan, S. W. C., Li, Q., Zeng, C., Law, C. O. K., Chan, G. H. C., Lau, T. C. K., and Ngo, J. C. K. (2018). "Distinct Mechanisms Govern the Phosphorylation of Different SR Protein Splicing Factors." *Journal of Biological Chemistry*. DOI: 10.1074/jbc.RA118.003392.

Los Alamos National Laboratory (2017). *HIV-1 Gene Map*. https://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html. [Online; accessed 12-June-2018]. Copyright notice: For Scientific and Technical Information Only. ©Copyright Triad National Security, LLC. All Rights Reserved. For All Information Unless otherwise indicated, this information has been authored by an employee or employees of the Triad National Security, LLC, operator of the Los Alamos National Laboratory with the U.S. Department of Energy. The U.S. Government has rights to use, reproduce, and distribute this information. The public may copy and use this information without charge, provided that this Notice and any statement of authorship are reproduced on all copies. Neither the Government nor Triad makes any warranty, express or implied, or assumes any liability or responsibility for the use of this information.

Louis, J. M., Zhang, Y., Sayer, J. M., Wang, Y.-F., Harrison, R. W., and Weber, I. T. (2011). "The L76V Drug Resistance Mutation Decreases the Dimer Stability and Rate of Autoprocessing of HIV-1 Protease by Reducing Internal Hydrophobic Contacts." en. *Biochemistry* 50.21, pp. 4786–4795. DOI: 10.1021/bi200033z.

Lovell, S. C., Word, J. M., Richardson, J. S., and Richardson, D. C. (2000). "The Penultimate Rotamer Library." *Proteins: Structure, Function, and Bioinformatics* 40.3, pp. 389–408. DOI: 10.1002/1097-0134(20000815)40:3<389::aid-prot50>3.3.co;2-u.

Lowe, E. D., Tews, I., Cheng, K. Y., Brown, N. R., Gul, S., Noble, M. E. M., Gamblin, S. J., and Johnson, L. N. (2002). "Specificity Determinants of Recruitment Peptides Bound to Phospho-CDK2/Cyclin A." *Biochemistry* 41.52, pp. 15625–15634. DOI: 10.1021/bi0268910.

Lu, K. P., Liou, Y.-C., and Zhou, X. Z. (2002). "Pinning Down Proline-Directed Phosphorylation Signaling." *Trends in Cell Biology* 12.4, pp. 164–172. DOI: 10.1016/S0962-8924(02)02253-5.

Lupas, A., Van Dyke, M., and Stock, J. (1991). "Predicting Coiled Coils from Protein Sequences." *Science* 252.5009, pp. 1162–1164. DOI: 10.1126/science.252.5009.1162.

Ma, W., Shang, Y., Wei, Z., Wen, W., Wang, W., and Zhang, M. (2010). "Phosphorylation of DCC by ERK2 Is Facilitated by Direct Docking of the Receptor P1 Domain to the Kinase." *Structure* 18.11, pp. 1502–1511. DOI: 10.1016/j.str.2010.08.011.

Machuqueiro, M. and Baptista, A. M. (2006). "Constant-pH Molecular Dynamics with Ionic Strength Effects: Protonation-Conformation Coupling in Decalysine." *The Journal of Physical Chemistry B* 110.6, pp. 2927–2933. DOI: 10.1021/jp056456q.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2017). *cluster: Cluster Analysis Basics and Extensions.* R package version 2.0.6.

Maeyer, M. D., Desmet, J., and Lasters, I. (1997). "All in One: A Highly Detailed Rotamer Library Improves Both Accuracy and Speed in the Modelling of Sidechains by Dead-End Elimination." *Folding and Design* 2.1, pp. 53–66. DOI: 10.1016/S1359-0278(97)00006-0.

Mahalingam, B., Boross, P., Wang, Y.-F., Louis, J. M., Fischer, C. C., Tozser, J., Harrison, R. W., and Weber, I. T. (2002). "Combining Mutations in HIV-1 Protease to Understand Mechanisms of Resistance." en. *Proteins: Structure, Function, and Genetics* 48.1, pp. 107–116. DOI: 10.1002/prot.10140.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). "The Protein Kinase Complement of the Human Genome." *Science* 298.5600, pp. 1912–1934. DOI: 10.1126/science.1075762.

Mansky, L. M. and Temin, H. M. (1995). "Lower in vivo Mutation Rate of Human Immunodeficiency Virus Type 1 than that Predicted from the Fidelity of Purified Reverse Transcriptase." *Journal of Virology* 69.8, pp. 5087–5094.

Margerison, E. S., Maguire, M., Pillay, D., Cane, P., and Elston, R. C. (2008). "The HIV-1 Protease Substitution K55R: A Protease-Inhibitor-Associated Substitution Involved in Restoring Viral Replication." en. *Journal of Antimicrobial Chemotherapy* 61.4, pp. 786–791. DOI: 10.1093/jac/dkm545.

Markley, J. L., Bahrami, A., Eghbalnia, H. R., Peterson, F. C., Tyler, R. C., Ulrich, E. L., Westler, W. M., and Volkman, B. F. (2009). "Macromolecular Structure Determination by NMR Spectroscopy." *Structural Bioinformatics.* Ed. by J. Gu and P. E. Bourne. Wiley-Blackwell. Chap. 5, pp. 93–142. ISBN: 978-0-470-18105-8.

Markham, A. (2018). "Ibalizumab: First Global Approval." *Drugs* 78.7, pp. 781–785. DOI: 10.1007/s40265-018-0907-5.

Maschera, B., Darby, G., Palú, G., Wright, L. L., Tisdale, M., Myers, R., Blair, E. D., and Furfine, E. S. (1996). "Human Immunodeficiency Virus. Mutations in the Viral Protease That Confer Resistance to Saquinavir Increase the Dissociation Rate Constant of the Protease-Saquinavir Complex." *Journal of Biological Chemistry* 271.52, pp. 33231–33235. DOI: 10.1074/jbc.271.52.33231.

Mayers, D. L., McCutchan, F. E., Sanders-Buell, E. E., Merritt, L. I., Dilworth, S., Fowler, A. K., Marks, C. A., Ruiz, N. M., Richman, D. D., and Roberts, C. R. (1992). "Characterization of HIV Isolates Arising After Prolonged Zidovudine Therapy." *Journal of Acquired Immune Deficiency Syndromes* 5.8, pp. 749–759. DOI: 10.1097/00126334-199208000-00001.

Mazzei, L., Ciurli, S., and Zambelli, B. (2014). "Hot Biological Catalysis: Isothermal Titration Calorimetry to Characterize Enzymatic Reactions." *Journal of Visualized Experiments* 86, p. 51487. DOI: 10.3791/51487.

McClendon, C. L., Friedland, G., Mobley, D. L., Amirkhani, H., and Jacobson, M. P. (2009). "Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles." *Journal of Chemical Theory and Computation* 5.9, pp. 2486–2502. DOI: 10.1021/ct9001812.

McColl, D. J. and Chen, X. (2010). "Strand Transfer Inhibitors of HIV-1 Integrase: Bringing IN a New Era of Antiretroviral Therapy." *Antiviral Research* 85.1. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects, pp. 101–118. DOI: 10.1016/j.antiviral.2009.11.004.

McCarrick, M. A. and Kollman, P. A. (1999). "Predicting Relative Binding Affinities of Non-Peptide HIV Protease Inhibitors with Free Energy Perturbation Calculations." *Journal of Computer-Aided Molecular Design* 13.2, pp. 109–121.

McGee Jr, T. D., Edwards, J., and Roitberg, A. E. (2014). "pH-REMD Simulations Indicate That the Catalytic Aspartates of HIV-1 Protease Exist Primarily in a Monoprotonated State." *The Journal of Physical Chemistry B* 118.44, pp. 12577–12585. DOI: 10.1021/jp504011c.

Medders, K. E. and Kaul, M. (2011). "Mitogen-Activated Protein Kinase p38 in HIV Infection and Associated Brain Injury." *Journal of Neuroimmune Pharmacology* 6.2, pp. 202–215. DOI: 10.1007/s11481-011-9260-0.

Meher, B. R. and Wang, Y. (2012). "Interaction of I50V Mutant and I50L/A71V Double Mutant HIV-Protease with Inhibitor TMC114 (Darunavir): Molecular Dynamics Simulation and Binding Free Energy Studies." *The Journal of Physical Chemistry B* 116.6, pp. 1884–1900. DOI: 10.1021/jp2074804.

Merk, A., Bartesaghi, A., Banerjee, S., Falconieri, V., Rao, P., Davis, M. I., Pragani, R., Boxer, M. B., Earl, L. A., Milne, J. L., and Subramaniam, S. (2016). “Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery.” *Cell* 165.7, pp. 1698–1707. DOI: 10.1016/j.cell.2016.05.040.

Mészáros, B., Simon, I., and Dosztányi, Z. (2009). “Prediction of Protein Binding Regions in Disordered Proteins.” *PLoS Computational Biology* 5.5, e1000376. DOI: 10.1371/journal.pcbi.1000376.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2018). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group, TU Wien.* R package version 1.7-0.

Mittal, S., Bandaranayake, R. M., King, N. M., Prabu-Jeyabalan, M., Nalam, M. N., Nalivaika, E. A., Yilmaz, N. K., and Schiffer, C. A. (2013). “Structural and Thermodynamic Basis of Amprenavir/Darunavir and Atazanavir Resistance in HIV-1 Protease with Mutations at Residue 50.” *Journal of Virology* 87.8, pp. 4176–4184. DOI: 10.1128/jvi.03486-12.

Mooney, L. M. and Whitmarsh, A. J. (2003). “Docking Interactions in the c-Jun N-terminal Kinase Pathway.” *Journal of Biological Chemistry* 279.12, pp. 11843–11852. DOI: 10.1074/jbc.m311841200.

Muslin, A. J. (2008). “MAPK Signalling in Cardiovascular Health and Disease: Molecular Mechanisms and Therapeutic Targets.” *Clinical Science* 115.7, pp. 203–218. DOI: 10.1042/cs20070430.

Muzammil, S., Armstrong, A. A., Kang, L. W., Jakalian, A., Bonneau, P. R., Schmelmer, V., Amzel, L. M., and Freire, E. (2007). “Unique Thermodynamic Response of Tipranavir to Human Immunodeficiency Virus Type 1 Protease Drug Resistance Mutations.” *Journal of Virology* 81.10, pp. 5144–5154. DOI: 10.1128/jvi.02706-06.

Neduva, V. and Russell, R. B. (2005). “Linear Motifs: Evolutionary Interaction Switches.” *FEBS Letters* 579.15, pp. 3342–3345. DOI: 10.1016/j.febslet.2005.04.005.

Nelson, W. D., Blakely, S. E., Nesmelov, Y. E., and Thomas, D. D. (2005). “Site-Directed Spin Labeling Reveals a Conformational Switch in the Phosphorylation Domain of Smooth Muscle Myosin.” *Proceedings of the National Academy of Sciences* 102.11, pp. 4000–4005. DOI: 10.1073/pnas.0401664102.

Ngo, J. C. K., Chakrabarti, S., Ding, J.-H., Velazquez-Dones, A., Nolen, B., Aubol, B. E., Adams, J. A., Fu, X.-D., and Ghosh, G. (2005). “Interplay Between SRPK and Clk/Sty Kinases in Phosphorylation of the Splicing Factor ASF/SF2 Is Regulated by a Docking Motif in ASF/SF2.” *Molecular Cell* 20.1, pp. 77–89. DOI: 10.1016/j.molcel.2005.08.025.

Ngo, S. T., Mai, B. K., Hiep, D. M., and Li, M. S. (2015). “Estimation of the Binding Free Energy of AC1NX476 to HIV-1 Protease Wild Type and Mutations Using Free Energy Perturbation Method.” *Chemical Biology & Drug Design* 86.4, pp. 546–558. DOI: 10.1111/cbdd.12518.

Nijhuis, M., Wensing, A. M. J., Bierman, W. F. W., Jong, D. de, Kagan, R., Fun, A., Jaspers, C. A. J. J., Schurink, K. A. M., Agtmael, M. A. van, and Boucher, C. A. B. (2009). “Failure of Treatment with First-Line Lopinavir Boosted with Ritonavir Can Be Explained by Novel Resistance Pathways with Protease Mutation 76V.” en. *The Journal of Infectious Diseases* 200.5, pp. 698–709. DOI: 10.1086/605329.

Nijhuis, M., Schuurman, R., Jong, D. de, Erickson, J., Gustchina, E., Albert, J., Schipper, P., Gulnik, S., and Boucher, C. A. (1999). “Increased Fitness of Drug Resistant HIV-1 Protease as a Result of Acquisition of Compensatory Mutations During Suboptimal Therapy.” *AIDS* 13.17, pp. 2349–2359. DOI: 10.1097/00002030-199912030-00006.

Nishi, H., Hashimoto, K., and Panchenko, A. R. (2011). “Phosphorylation in Protein-Protein Binding: Effect on Stability and Function.” *Structure* 19.12, pp. 1807–1815. DOI: 10.1016/j.str.2011.09.021.

Nishi, H., Fong, J. H., Chang, C., Teichmann, S. A., and Panchenko, A. R. (2013). “Regulation of Protein-Protein Binding by Coupling Between Phosphorylation and Intrinsic Disorder: Analysis of Human Protein Complexes.” *Molecular BioSystems* 9 (7), pp. 1620–1626. DOI: 10.1039/c3mb25514j.

Nussinov, R., Tsai, C.-J., Xin, F., and Radivojac, P. (2012). “Allosteric Post-Translational Modification Codes.” *Trends in Biochemical Sciences* 37.10, pp. 447–455. DOI: 10.1016/j.tibs.2012.07.001.

Ode, H., Neya, S., Hata, M., Sugiura, W., and Hoshino, T. (2006). “Computational Simulations of HIV-1 Proteases Multi-Drug Resistance due to Nonactive Site Mutation L90M.” *Journal of the American Chemical Society* 128.24, pp. 7887–7895. DOI: 10.1021/ja060682b.

Ode, H., Matsuyama, S., Hata, M., Hoshino, T., Kakizawa, J., and Sugiura, W. (2007). “Mechanism of Drug Resistance Due to N88S in CRF01_AE HIV-1 Protease, Analyzed by Molecular Dynamics Simulations.” en. *Journal of Medicinal Chemistry* 50.8, pp. 1768–1777. DOI: 10.1021/jm061158i.

Old, W. M., Shabb, J. B., Houel, S., Wang, H., Couts, K. L., Yen, C.-y., Litman, E. S., Croy, C. H., Meyer-Arendt, K., Miranda, J. G., Brown, R. A., Witze, E. S., Schweppe, R. E., Resing, K. A., and Ahn, N. G. (2009). “Functional Proteomics Identifies Targets of Phosphorylation by B-Raf Signaling in Melanoma.” *Molecular Cell* 34.1, pp. 115–131. DOI: 10.1016/j.molcel.2009.03.007.

Olson, B. L., Hock, M. B., Ekholm-Reed, S., Wohlschlegel, J. A., Dev, K. K., Kralli, A., and Reed, S. I. (2008). "SCFCdc4 Acts Antagonistically to the PGC-1α Transcriptional Coactivator by Targeting it for Ubiquitin-Mediated Proteolysis." *Genes & Development* 22.2, pp. 252–264. DOI: 10.1101/gad.1624208.

Ott, D. E., Coren, L. V., Chertova, E. N., Gagliardi, T. D., Nagashima, K., Sowder, R. C., Poon, D. T. K., and Gorelick, R. J. (2003). "Elimination of Protease Activity Restores Efficient Virion Production to a Human Immunodeficiency Virus Type 1 Nucleocapsid Deletion Mutant." en. *Journal of Virology* 77.10, pp. 5547–5556. DOI: 10.1128/JVI.77.10.5547-5556.2003.

Overhauser, A. W. (1953). "Polarization of Nuclei in Metals." *Physical Review* 92 (2), pp. 411–415. DOI: 10.1103/PhysRev.92.411.

Palella, F. J., Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer, J., Satten, G. A., Aschman, D. J., and Holmberg, S. D. (1998). "Declining Morbidity and Mortality among Patients with Advanced Human Immunodeficiency Virus Infection." *New England Journal of Medicine* 338.13, pp. 853–860. DOI: 10.1056/nejm199803263381301.

Pan, C., Olsen, J. V., Daub, H., and Mann, M. (2009). "Global Effects of Kinase Inhibitors on Signaling Networks Revealed by Quantitative Phosphoproteomics." *Molecular & Cellular Proteomics* 8.12, pp. 2796–2808. DOI: 10.1074/mcp.m900285-mcp200.

Parinello, M. and Rahman, A. (1981). "Canonical Sampling Through Velocity Rescaling." *Journal of Applied Physics* 52.12, pp. 7182–7190. DOI: 10.1063/1.2408420.

Pazhanisamy, S., Stuver, C. M., Cullinan, A. B., Margolin, N., Rao, B., and Livingston, D. J. (1996). "Kinetic Characterization of Human Immunodeficiency Virus Type-1 Protease-resistant Variants." *Journal of Biological Chemistry* 271.30, pp. 17979–17985. DOI: 10.1074/jbc.271.30.17979.

Pearlman, A., Loke, J., Caignec, C. L., White, S., Chin, L., Friedman, A., Warr, N., Willan, J., Brauer, D., Farmer, C., Brooks, E., Oddoux, C., Riley, B., Shajahan, S., Camerino, G., Homfray, T., Crosby, A. H., Couper, J., David, A., Greenfield, A., Sinclair, A., and Ostrer, H. (2010). "Mutations in MAP3K1 Cause 46,XY Disorders of Sex Development and Implicate a Common Signal Transduction Pathway in Human Testis Determination." *The American Journal of Human Genetics* 87.6, pp. 898–904. DOI: 10.1016/j.ajhg.2010.11.003.

Pearlman, D. A. (2005). "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, pp. 7796–7807. DOI: 10.1021/jm050306m.

Pellicena, P., Stowell, K. R., and Miller, W. T. (1998). "Enhanced Phosphorylation of Src Family Kinase Substrates Containing SH2 Domain Binding Sites." *Journal of Biological Chemistry* 273.25, pp. 15325–15328. DOI: 10.1074/jbc.273.25.15325.

Perryman, A. L., Lin, J. H., and McCammon, J. A. (2004). "HIV-1 Protease Molecular Dynamics of a Wild-Type and of the V82F/I84V Mutant: Possible Contributions to Drug Resistance and a Potential New Target Site for Drugs." *Protein Science* 13.4, pp. 1108–1123. DOI: 10.1110/ps.03468904.

Perryman, A. L., Lin, J. H., and McCammon, J. A. (2006). "Restrained Molecular Dynamics Simulations of HIV-1 Protease: The First Step in Validating a New Target for Drug Design." *Biopolymers* 82.3, pp. 272–284. DOI: 10.1002/bip.20497.

Petropoulos, C. J., Parkin, N. T., Limoli, K. L., Lie, Y. S., Wrin, T., Huang, W., Tian, H., Smith, D., Winslow, G. A., Capon, D. J., and Whitcomb, J. M. (2000). "A Novel Phenotypic Drug Susceptibility Assay for Human Immunodeficiency Virus Type 1." *Antimicrobial Agents and Chemotherapy* 44.4, pp. 920–928. DOI: 10.1128/aac.44.4.920-928.2000.

Petersen, T. N., Brunak, S., Heijne, G. von, and Nielsen, H. (2011). "SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions." *Nature Methods* 8.10, pp. 785–786. DOI: 10.1038/nmeth.1701.

Plantier, J.-C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., Robertson, D. L., and Simon, F. (2009). "A New Human Immunodeficiency Virus Derived from Gorillas." *Nature Medicine* 15.8, pp. 871–872. DOI: 10.1038/nm.2016.

Ponder, J. W. and Richards, F. M. (1987). "Tertiary Templates for Proteins: Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes." *Journal of Molecular Biology* 193.4, pp. 775–791. DOI: 10.1016/0022-2836(87)90358-5.

Prabu-Jeyabalan, M., Nalivaika, E., and Schiffer, C. A. (2002). "Substrate Shape Determines Specificity of Recognition for HIV-1 Protease: Analysis of Crystal Structures of Six Substrate Complexes." *Structure* 10.3, pp. 369–381. DOI: 10.2210/pdb1kjf/pdb.

Prado, J. G., Wrin, T., Beauchaine, J., Ruiz, L., Petropoulos, C. J., Frost, S. D., Clotet, B., Richard, T. D., and Martinez-Picado, J. (2002). "Amprenavir-Resistant HIV-1 Exhibits Lopinavir Cross-Resistance and Reduced Replication Capacity." *AIDS* 16.7, pp. 1009–1017. DOI: 10.1097/00002030-200205030-00007.

Qu, C., Li, W., Shao, Q., Dwyer, T., Huang, H., Yang, T., and Liu, G. (2013). "c-Jun N-terminal Kinase 1 (JNK1) Is Required for Coordination of Netrin Signaling in Axon Guidance." *Journal of Biological Chemistry* 288.3, pp. 1883–1895. DOI: 10.1074/jbc.m112.417881.

Ragland, D. A., Nalivaika, E. A., Nalam, M. N. L., Prachanronarong, K. L., Cao, H., Bandaranayake, R. M., Cai, Y., Kurt-Yilmaz, N., and Schiffer, C. A. (2014). "Drug Resistance Conferred by Mutations Outside the Active Site through Alterations in the Dynamic and Structural Ensemble of HIV-1 Protease." en. *Journal of the American Chemical Society* 136.34, pp. 11956–11963. DOI: 10.1021/ja504096m.

Ragland, D. A., Whitfield, T. W., Lee, S.-K., Swanstrom, R., Zeldovich, K. B., Kurt-Yilmaz, N., and Schiffer, C. A. (2017). "Elucidating the Interdependence of Drug Resistance from Combinations of Mutations." *Journal of Chemical Theory and Computation* 13.11, pp. 5671–5682. DOI: 10.1021/acs.jctc.7b00601.

Rao, B., Tilton, R., and Singh, U. (1992). "Free Energy Perturbation Studies on Inhibitor Binding to HIV-1 Proteinase." *Journal of the American Chemical Society* 114.12, pp. 4447–4452. DOI: 10.1021/ja00038a001.

Rao, B. and Murcko, M. (1994). "Reversed Stereochemical Preference in Binding of Ro 31-8959 to HIV-1 Proteinase: A Free Energy Perturbation Analysis." *Journal of Computational Chemistry* 15.11, pp. 1241–1253. DOI: 10.1002/jcc.540151106.

Rao, B., Kim, E., and Murcko, M. (1996). "Calculation of Solvation and Binding Free Energy Differences Between VX-478 and Its Analogs by Free Energy Perturbation and AMSOL Methods." *Journal of Computer-Aided Molecular Design* 10.1, pp. 23–30. DOI: 10.1007/bf00124462.

Rao, B. and Murcko, M. (1996). "Free Energy Perturbation Studies on Binding of A-74704 and its Diester Analog to HIV-1 Protease." *Protein Engineering, Design and Selection* 9.9, pp. 767–771. DOI: 10.1093/protein/9.9.767.

Rätsch, G. and Sonnenburg, S. (2004). "Accurate Splice Site Prediction for Caenorhabditis elegans." *Kernel Methods in Computational Biology*. MIT Press series on Computational Molecular Biology. MIT Press, pp. 277–298.

Rätsch, G., Sonnenburg, S., and Schölkopf, B. (2005). "RASE: Recognition of Alternatively Spliced Exons in C.elegans." *Bioinformatics* 21.suppl_1, pp. i369–i377. DOI: 10.1093/bioinformatics/bti1053.

Reddy, M. R., Viswanadhan, V. N., and Weinstein, J. N. (1991). "Relative Differences in the Binding Free Energies of Human Immunodeficiency Virus 1 Protease Inhibitors: a Thermodynamic Cycle-Perturbation Approach." *Proceedings of the National Academy of Sciences* 88.22, pp. 10287–10291. DOI: 10.1073/pnas.88.22.10287.

Reddy, M. R. and Erion, M. D. (1998). "Structure-Based Drug Design Approaches for Predicting Binding Affinities of Hiv1 Protease Inhibitors." *Journal of Enzyme Inhibition* 14.1, pp. 1–14. DOI: 10.3109/14756369809036542.

Regan, C. P., Li, W., Boucher, D. M., Spatz, S., Su, M. S., and Kuida, K. (2002). "Erk5 Null Mice Display Multiple Extraembryonic Vascular and Embryonic Cardiovascular Defects." *Proceedings of the National Academy of Sciences* 99.14, pp. 9248–9253. DOI: 10.1073/pnas.142293999.

Reményi, A., Good, M. C., Bhattacharyya, R. P., and Lim, W. A. (2005). "The Role of Docking Interactions in Mediating Signaling Input, Output, and Discrimination in the Yeast MAPK Network." *Molecular Cell* 20.6, pp. 951–962. DOI: 10.1016/j.molcel.2005.10.030.

Resch, W., Ziermann, R., Parkin, N., Gamarnik, A., and Swanstrom, R. (2002). "Nelfinavir-Resistant, Amprenavir-Hypersusceptible Strains of Human Immunodeficiency Virus Type 1 Carrying an N88S Mutation in Protease Have Reduced Infectivity, Reduced Replication Capacity, and Reduced Fitness and Process the Gag Polyprotein Precursor Aberrantly." *Journal of Virology* 76.17, pp. 8659–8666. DOI: 10.1128/JVI.76.17.8659-8666.2002.

Rhee, S.-Y., Taylor, J., Fessel, W. J., Kaufman, D., Towner, W., Troia, P., Ruane, P., Hellinger, J., Shirvani, V., Zolopa, A., and Shafer, R. W. (2010). "HIV-1 Protease Mutations and Protease Inhibitor Cross-Resistance." *Antimicrobial Agents and Chemotherapy* 54.10, pp. 4253–4261. DOI: 10.1128/AAC.00574-10.

Rizzo, R. C., Wang, D.-P., Tirado-Rives, J., and Jorgensen, W. L. (2000). "Validation of a Model for the Complex of HIV-1 Reverse Transcriptase with Sustiva through Computation of Resistance Profiles." *Journal of the American Chemical Society* 122.51, pp. 12898–12900. DOI: 10.1021/ja003113r.

Rom, W. N. and Markowitz, S. B. (2007). *Environmental and Occupational Medicine*. 4th ed. Lippincott Williams & Wilkins. ISBN: 978-0-781-76299-1.

Roy, S., Delling, U., Chen, C.-H., Rosen, C., and Sonenberg, N. (1990). "A Bulge Structure in HIV-1 TAR RNA Is Required for Tat Binding and Tat-Mediated Trans-Activation." *Genes & Development* 4.8, pp. 1365–1373. DOI: 10.1101/gad.4.8.1365.

Rubin, G. M. (2001). “The Draft Sequences: Comparing Species.” *Nature* 409.6822, p. 820. DOI: 10.1038/35057277.

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. (1977). “Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes.” *Journal of Computational Physics* 23.3, pp. 327–341. DOI: 10.1016/0021-9991(77)90098-5.

Sabio, G., Kennedy, N. J., Cavanagh-Kyros, J., Jung, D. Y., Ko, H. J., Ong, H., Barrett, T., Kim, J. K., and Davis, R. J. (2009). “Role of Muscle c-Jun NH2-terminal Kinase 1 in Obesity-Induced Insulin Resistance.” *Molecular and Cellular Biology* 30.1, pp. 106–115. DOI: 10.1128/mcb.01162-09.

Sadiq, S. K., Wright, D. W., Kenway, O. A., and Coveney, P. V. (2010). “Accurate Ensemble Molecular Dynamics Binding Free Energy Ranking of Multidrug-Resistant HIV-1 Proteases.” *Journal of Chemical Information and Modeling* 50.5, pp. 890–905. DOI: 10.1021/ci100007w.

Sali, A. and Blundell, T. L. (1993). “Comparative Protein Modelling by Satisfaction of Spatial Restraints.” *Journal of Molecular Biology* 234.3, pp. 779–815. DOI: 10.1006/jmbi.1993.1626.

Santos, A. F. and Soares, M. A. (2010). “HIV Genetic Diversity and Drug Resistance.” *Viruses* 2.2, pp. 503–531. DOI: 10.3390/v2020503.

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., and Overington, J. P. (2016). “A Comprehensive Map of Molecular Drug Targets.” *Nature Reviews Drug Discovery* 16.1, pp. 19–34. DOI: 10.1038/nrd.2016.230.

Sarafianos, S. G., Marchand, B., Das, K., Himmel, D. M., Parniak, M. A., Hughes, S. H., and Arnold, E. (2009). “Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition.” *Journal of Molecular Biology* 385.3, pp. 693–713. DOI: 10.1016/j.jmb.2008.10.071.

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., and Platt, J. C. (2000). “Support Vector Method for Novelty Detection.” *Advances in Neural Information Processing Systems*, pp. 582–588.

Schrager, J. A., Der Minassian, V., and Marsh, J. W. (2002). “HIV Nef Increases T Cell ERK MAP Kinase Activity.” *Journal of Biological Chemistry* 277.8, pp. 6137–6142. DOI: 10.1074/jbc.m107322200.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). “SWISS-MODEL: An Automated Protein Homology-Modeling Server.” *Nucleic Acids Research* 31.13, pp. 3381–3385. DOI: 10.1093/nar/gkg520.

Schweiger, R. and Linial, M. (2010). “Cooperativity Within Proximal Phosphorylation Sites is Revealed from Large-Scale Proteomics Data.” *Biology Direct* 5.1, p. 6. DOI: 10.1186/1745-6150-5-6.

Schock, H. B., Garsky, V. M., and Kuo, L. C. (1996). “Mutational Anatomy of an HIV-1 Protease Variant Conferring Cross-resistance to Protease Inhibitors in Clinical Trials Compensatory Modulations of Binding and Activity.” *Journal of Biological Chemistry* 271.50, pp. 31957–31963. DOI: 10.1074/jbc.271.50.31957.

Schlick, T. (2002). *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 978-1-4419-6350-5. DOI: 10.1007/978-1-4419-6351-2.

Schmitt, L. (2005). *Biochemie. Eine Einführung für Mediziner und Naturwissenschaftler*. Vol. 117. 26. WILEY-VCH Verlag, pp. 4025–4025. DOI: 10.1002/ange.200385274.

Schutten, M. (2006). “Resistance Assays.” *Antiretroviral Resistance in Clinical Practice*. Ed. by A. M. Geretti. Mediscript. Chap. 5. ISBN: 978-0-955-16690-7.

Sharrocks, A. D., Yang, S.-H., and Galanis, A. (2000). “Docking Domains and Substrate-Specificity Determination for MAP Kinases.” *Trends in Biochemical Sciences* 25.9, pp. 448–453. DOI: 10.1016/s0968-0004(00)01627-3.

Sharp, P. M. and Hahn, B. H. (2011). “Origins of HIV and the AIDS Pandemic.” *Cold Spring Harbor Perspectives in Medicine* 1.1, a006841. DOI: 10.1101/cshperspect.a006841.

Shen, C. H., Wang, Y. F., Kovalevsky, A. Y., Harrison, R. W., and Weber, I. T. (2010). “Amprenavir Complexes with HIV-1 Protease and Its Drug-Resistant Mutants Altering Hydrophobic Clusters.” *FEBS Journal* 277.18, pp. 3699–3714. DOI: 10.1111/j.1742-4658.2010.07771.x.

Shirts, M. R., Bair, E., Hooker, G., and Pande, V. S. (2003). “Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods.” *Physical Review Letters* 91.14, p. 140601. DOI: 10.1103/physrevlett.91.140601.

Shi, S.-H., Chen, J.-Z., Hu, G.-D., Yi, C.-H., Zhang, S.-L., and Zhang, Q.-G. (2009). “Molecular Insight into the Interaction Mechanisms of Inhibitors BEC and BEG with HIV-1 Protease by Using MM-PBSA Method and Molecular Dynamics Simulation.” *Journal of Molecular Structure: THEOCHEM* 913.1, pp. 22–27. DOI: 10.1016/j.theochem.2009.07.010.

Silva, A. W. S. da and Vranken, W. F. (2012). “ACPYPE-AnteChamber PYthon Parser interfacE.” *BMC Research Notes* 5.1, p. 367. DOI: 10.1186/1756-0500-5-367.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). "ROCR: Visualizing Classifier Performance in R." *Bioinformatics* 21.20, p. 7881. DOI: 10.1093/bioinformatics/bti623.

Singh, G. P. (2015). "Association Between Intrinsic Disorder and Serine/Threonine Phosphorylation in Mycobacterium tuberculosis." *PeerJ* 3, e724. DOI: 10.7717/peerj.724.

Smith, J. A., Poteet-Smith, C. E., Lannigan, D. A., Freed, T. A., Zoltoski, A. J., and Sturgill, T. W. (2000). "Creation of a Stress-Activated p90 Ribosomal S6 Kinase the Carboxyl-terminal Tail of the MAPK-Activated Protein Kinases Dictates the Signal Transduction Pathway in Which They Function." *Journal of Biological Chemistry* 275.41, pp. 31588–31593. DOI: 10.1074/jbc.m005892200.

Smith, R. E., Lovell, S. C., Burke, D. F., Montalvao, R. W., and Blundell, T. L. (2007). "Andante: Reducing Side-Chain Rotamer Search Space During Comparative Modeling Using Environment-Specific Substitution Probabilities." *Bioinformatics* 23.9, pp. 1099–1105. DOI: 10.1093/bioinformatics/btm073.

Søndergaard, C. R., Olsson, M. H., Rostkowski, M., and Jensen, J. H. (2011). "Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values." *Journal of Chemical Theory and Computation* 7.7, pp. 2284–2295. DOI: 10.1021/ct200133y.

Sonnenburg, S. *et al.* (2017). *shogun-toolbox/shogun: Shogun 6.1.3*. DOI: 10.5281/zenodo.1067840.

Spielmann, M., Kakar, N., Tayebi, N., Leettola, C., Nürnberg, G., Sowada, N., Lupiáñez, D. G., Harabula, I., Flöttmann, R., Horn, D., Chan, W. L., Wittler, L., Yilmaz, R., Altmüller, J., Thiele, H., Bokhoven, H. van, Schwartz, C. E., Nürnberg, P., Bowie, J. U., Ahmad, J., Kubisch, C., Mundlos, S., and Borck, G. (2016). "Exome Sequencing and CRISPR/Cas Genome Editing Identify Mutations of ZAK as a Cause of Limb Defects in Humans and Mice." *Genome Research* 26.2, pp. 183–191. DOI: 10.1101/gr.199430.115.

Spinelli, S., Liu, Q. Z., Alzari, P. M., Hirel, P., and Poljak, R. J. (1991). "The Three-Dimensional Structure of the Aspartyl Protease from the HIV-1 Isolate BRU." *Biochimie* 73.11, pp. 1391–1396. DOI: 10.1016/0300-9084(91)90169-2.

Srivastava, H. K. and Sastry, G. N. (2012). "Molecular Dynamics Investigation on a Series of HIV Protease Inhibitors: Assessing the Performance of MM-PBSA and MM-GBSA Approaches." *Journal of Chemical Information and Modeling* 52.11, pp. 3088–3098. DOI: 10.1021/ci300385h.

Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., and Case, D. A. (1998). "Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices." *Journal of the American Chemical Society* 120.37, pp. 9401–9409. DOI: 10.1021/ja981844+.

Stanley, B. J., Ehrlich, E. S., Short, L., Yu, Y., Xiao, Z., Yu, X.-F., and Xiong, Y. (2008). "Structural Insight into the Human Immunodeficiency Virus Vif SOCS Box and Its Role in Human E3 Ubiquitin Ligase Assembly." *Journal of Virology* 82.17, pp. 8656–8663. DOI: 10.1128/jvi.00767-08.

Stock, A. M., Robinson, V. L., and Goudreau, P. N. (2000). "Two-Component Signal Transduction." *Annual Review of Biochemistry* 69.1, pp. 183–215. DOI: 10.1146/annurev.biochem.69.1.183.

Stoica, I., Sadiq, S. K., and Coveney, P. V. (2008). "Rapid and Accurate Prediction of Binding Free Energies for Saquinavir-Bound HIV-1 Proteases." *Journal of the American Chemical Society* 130.8, pp. 2639–2648. DOI: 10.1021/ja0779250.

Tanoue, T., Adachi, M., Moriguchi, T., and Nishida, E. (2000). "A Conserved Docking Motif in MAP Kinases Common to Substrates, Activators and Regulators." *Nature Cell Biology* 2.2, pp. 110–116. DOI: 10.1038/35000065.

Tanoue, T., Maeda, R., Adachi, M., and Nishida, E. (2001). "Identification of a Docking Groove on ERK and p38 MAP Kinases that Regulates the Specificity of Docking Interactions." *The EMBO Journal* 20.3, pp. 466–479. DOI: 10.1093/emboj/20.3.466.

Tawa, G. J., Topol, I. A., Burt, S. K., and Erickson, J. W. (1998). "Calculation of Relative Binding Free Energies of Peptidic Inhibitors to HIV-1 Protease and Its I84V Mutant." *Journal of the American Chemical Society* 120.34, pp. 8856–8863. DOI: 10.1021/ja9733090.

The UK Collaborative Group on HIV Drug Resistance, U. C. S. G. (2005). "Long Term Probability of Detection of HIV-1 Drug Resistance After Starting Antiretroviral Therapy in Routine Clinical Practice." *AIDS* 19.5, pp. 487–494. DOI: 10.1097/01.aids.0000162337.58557.3d.

The UniProt Consortium (2016). "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 45.D1, pp. D158–D169. DOI: 10.1093/nar/gkw1099.

Thomsen, M. C. F. and Nielsen, M. (2012). "Seq2Logo: A Method for Construction and Visualization of Amino Acid Binding Motifs and Sequence Profiles Including Sequence Weighting, Pseudo Counts and Two-Sided Representation of Amino Acid Enrichment and Depletion." *Nucleic Acids Research* 40.W1, W281–W287. DOI: 10.1093/nar/gks469.

Thomas Jr, C. A. (1971). “The Genetic Organization of Chromosomes.” *Annual Review of Genetics* 5.1, pp. 237–256. DOI: 10.1146/annurev.ge.05.120171.001321.

Tibshirani, R., Walther, G., and Hastie, T. (2001). “Estimating the Number of Clusters in a Data Set via the Gap Statistic.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 411–423. DOI: 10.1111/1467-9868.00293.

Tilton, J. C. and Doms, R. W. (2010). “Entry Inhibitors in the Treatment of HIV-1 Infection.” *Antiviral Research* 85.1. Twenty-five Years of Antiretroviral Drug Development: Progress and Prospects, pp. 91–100. DOI: 10.1016/j.antiviral.2009.07.022.

Tompa, P., Davey, N. E., Gibson, T. J., and Babu, M. M. (2014). “A Million Peptide Motifs for the Molecular Biologist.” *Molecular Cell* 55.2, pp. 161–169. DOI: 10.1016/j.molcel.2014.05.032.

Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J.-H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri S Lasky, L. A., Sander, C., Boone, C., Bader, G. D., and Sidhu, S. S. (2008). “A Specificity Map for the PDZ Domain Family.” *PLoS Biology* 6.9, e239. DOI: 10.1371/journal.pbio.0060239.

Tong, L. (2002). “Viral Proteases.” *Chemical Reviews* 102.12, pp. 4609–4626. DOI: 10.1021/cr010184f.

Toschi, E., Bacigalupo, I., Strippoli, R., Chiozzini, C., Cereseto, A., Falchi, M., Nappi, F., Sgadari, C., Barillari, G., Mainiero, F., and Ensoli, B. (2006). “HIV-1 Tat Regulates Endothelial Cell Cycle Progression via Activation of the Ras/ERK MAPK Signaling Pathway.” *Molecular Biology of the Cell* 17.4, pp. 1985–1994. DOI: 10.1091/mbc.e05-08-0717.

Tropsha, A. and Hermans, J. (1992). “Application of free Energy Simulations to the Binding of a Transition-State-Analogue Inhibitor to HTV Protease.” *Protein Engineering, Design and Selection* 5.1, pp. 29–33. DOI: 10.1093/protein/5.1.29.

Tseng, G. C. and Wong, W. H. (2005). “Tight Clustering: A Resampling-Based Approach for Identifying Stable and Tight Patterns in Data.” *Biometrics* 61.1, pp. 10–16. DOI: 10.1111/j.0006-341x.2005.031032.x.

Tseng, G. C. and Wong, W. H. (2012). *tightClust: Tight Clustering*. R package version 1.0. URL: https://CRAN.R-project.org/package=tightClust.

Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. (1991). “A New Approach to the Rapid Determination of Protein Side Chain Conformations.” *Journal of Biomolecular Structure and Dynamics* 8.6, pp. 1267–1289. DOI: 10.1080/07391102.1991.10507882.

Tzoupis, H., Leonis, G., Megariotis, G., Supuran, C. T., Mavromoustakos, T., and Papadopoulos, M. G. (2012). “Dual Inhibitors for Aspartic Proteases HIV-1 PR and Renin: Advancements in AIDS-Hypertension-Diabetes Linkage via Molecular Dynamics, Inhibition Assays, and Binding Free Energy Calculations.” *Journal of Medicinal Chemistry* 55.12, pp. 5784–5796. DOI: 10.1021/jm300180r.

Udier-Blagović, M., Tirado-Rives, J., and Jorgensen, W. L. (2003). “Validation of a Model for the Complex of HIV-1 Reverse Transcriptase with Nonnucleoside Inhibitor TMC125.” *Journal of the American Chemical Society* 125.20, pp. 6016–6017. DOI: 10.1021/ja034308c.

Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., Mbanya, D., Kaptué, L., Ndembi, N., Gürtler, L., Devare, S., and Brennan, C. A. (2011). “Confirmation of Putative HIV-1 Group P in Cameroon.” *Journal of Virology* 85.3, pp. 1403–1407. DOI: 10.1128/jvi.02005-10.

Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B. F., Liu, W., Loul, S., Butel, C., Liegeois, F., Bienvenue, Y., Ngolle, E. M., Sharp, P. M., Shaw, G. M., Delaporte, E., Hahn, B. H., and Peeters, M. (2006). “Human Immunodeficiency Viruses: SIV Infection in Wild Gorillas.” *Nature* 444.7116, p. 164. DOI: 10.1038/444164a.

Venkatarajan, M. S. and Braun, W. (2001). “New Quantitative Descriptors of Amino Acids Based on Multidimensional Scaling of a Large Number of Physical-Chemical Properties.” *Journal of Molecular Modeling* 7.12, pp. 445–453. DOI: 10.1007/s00894-001-0058-5.

Vendrely, R. and Vendrely, C. (1948). “La teneur du noyau cellulaire en acide désoxyribonucléique à travers les organes, les individus et les espèces animales.” *Experientia* 4.11, pp. 434–436. DOI: 10.1007/bf02144998.

Vermeiren, H., Craenenbroeck, E. V., Alen, P., Bacheler, L., Picchio, G., and Lecocq, P. (2007). “Prediction of HIV-1 Drug Susceptibility Phenotype from the Viral Genotype Using Linear Regression Modeling.” *Journal of Virological Methods* 145.1, pp. 47–55. DOI: 10.1016/j.jviromet.2007.05.009.

Vitari, A. C., Thastrup, J., Rafiqi, F. H., Deak, M., Morrice, N. A., Karlsson, H. K., and Alessi, D. R. (2006). "Functional Interactions of the SPAK/OSR1 Kinases with Their Upstream Activator WNK1 and Downstream Substrate NKCC1." *Biochemical Journal* 397.1, pp. 223–231. DOI: 10.1042/bj20060220.

Vlahopoulos, S. and Zoumpourlis, V. (2004). "JNK: A Key Modulator of Intracellular Signaling." *Biochemistry (Moscow)* 69.8, pp. 844–854. DOI: 10.1023/b:biry.0000040215.02460.45.

Vlastaridis, P., Kyriakidou, P., Chaliotis, A., Van de Peer, Y., Oliver, S. G., and Amoutzias, G. D. (2017). "Estimating the Total Number of Phosphoproteins and Phosphorylation Sites in Eukaryotic Proteomes." *GigaScience* 6.2, giw015. DOI: 10.1093/gigascience/giw015.

Volkmann, N. and Hanein, D. (2009). "Electron Microscopy in the Context of Structural Systems Biology." *Structural Bioinformatics*. Ed. by J. Gu and P. E. Bourne. Wiley-Blackwell. Chap. 6, pp. 143–170. ISBN: 978-0-470-18105-8.

Wade, E. M., Daniel, P. B., Jenkins, Z. A., McInerney-Leo, A., Leo, P., Morgan, T., Addor, M. C., Adès, L. C., Bertola, D., Bohring, A., Carter, E., Cho, T. J., Duba, H. C., Fletcher, E., Kim, C. A., Krakow, D., Morava, E., Neuhann, T., Superti-Furga, A., Veenstra-Knol, I., Wieczorek, D., Wilson, L. C., Hennekam, R. C., Sutherland-Smith, A. J., Strom, T. M., Wilkie, A. O., Brown, M. A., Duncan, E. L., Markie, D. M., and Robertson, S. P. (2016). "Mutations in MAP3K7 That Alter the Activity of the TAK1 Signaling Complex Cause Frontometaphyseal Dysplasia." *The American Journal of Human Genetics* 99.2, pp. 392–406. DOI: 10.1016/j.ajhg.2016.05.024.

Walsh, C. T., Garneau-Tsodikova, S., and Gatto, G. J. (2005). "Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications." *Angewandte Chemie International Edition* 44.45, pp. 7342–7372. DOI: 10.1002/anie.200501023.

Wang, D.-P., Rizzo, R. C., Tirado-Rives, J., and Jorgensen, W. L. (2001). "Antiviral Drug Design: Computational Analyses of the Effects of the L100I Mutation for HIV-RT on the Binding of NNRTIs." *Bioorganic & Medicinal Chemistry Letters* 11.21, pp. 2799–2802. DOI: 10.1016/S0960-894X(01)00510-8.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). "Development and Testing of a General Amber Force Field." *Journal of Computational Chemistry* 25.9, pp. 1157–1174. DOI: 10.1002/jcc.20035.

Wang, Y. X., Freedberg, D. I., Yamazaki, T., Wingfield, P. T., Stahl, S. J., Kaufman, J. D., Kiso, Y., and Torchia, D. A. (1996). "Solution NMR Evidence That the HIV-1 Protease Catalytic Aspartyl Groups Have Different Ionization States in the Complex Formed with the Asymmetric Drug KNI-272." *Biochemistry* 35.31, pp. 9945–9950. DOI: 10.1021/bi961268z.

Weber, I. T., Waltman, M. J., Mustyakimov, M., Blakeley, M. P., Keen, D. A., Ghosh, A. K., Langan, P., and Kovalevsky, A. Y. (2013). "Joint X-Ray/Neutron Crystallographic Study of HIV-1 Protease with Clinical Inhibitor Amprenavir: Insights for Drug Design." *Journal of Medicinal Chemistry* 56.13, pp. 5631–5635. DOI: 10.1021/jm400684f.

Wensing, A. M., Maarseveen, N. M. van, and Nijhuis, M. (2010). "Fifteen Years of HIV Protease Inhibitors: Raising the Barrier to Resistance." *Antiviral Research* 85.1, pp. 59–74. DOI: 10.1016/j.antiviral.2009.10.003.

Whisenant, T. C., Ho, D. T., Benz, R. W., Rogers, J. S., Kaake, R. M., Gordon, E. A., Huang, L., Baldi, P., and Bardwell, L. (2010). "Computational Prediction and Experimental Verification of New MAP Kinase Docking Sites and Substrates Including Gli Transcription Factors." *PLoS Computational Biology* 6.8, e1000908. DOI: 10.1371/journal.pcbi.1000908.

Wiesmann, F., Vachta, J., Ehret, R., Walter, H., Kaiser, R., Stürmer, M., Tappe, A., Däumer, M., Berg, T., Naeth, G., Braun, P., and Knechten, H. (2011). "The L76V Mutation in HIV-1 Protease Is Potentially Associated with Hypersusceptibility to Protease Inhibitors Atazanavir and Saquinavir: Is There a Clinical Advantage?" *AIDS Research and Therapy* 8.1, p. 7. DOI: 10.1186/1742-6405-8-7.

Wikipedia contributors (2018). *MAPK/ERK pathway — Wikipedia, The Free Encyclopedia.* [Online; accessed 25-November-2018]. Public domain. URL: https://commons.wikimedia.org/wiki/File:MAPKpathway.jpg.

Wittayanarakul, K., Hannongbua, S., and Feig, M. (2008). "Accurate Prediction of Protonation State as a Prerequisite for Reliable MM-PB(GB)SA Binding Free Energy Calculations of HIV-1 Protease Inhibitors." *Journal of Computational Chemistry* 29.5, pp. 673–685. DOI: 10.1002/jcc.20821.

Wong-Sam, A., Wang, Y.-F., Zhang, Y., Ghosh, A. K., Harrison, R. W., and Weber, I. T. (2018). "Drug Resistance Mutation L76V Alters Nonpolar Interactions at the Flap–Core Interface of HIV-1 Protease." *ACS Omega* 3.9, pp. 12132–12140. DOI: 10.1021/acsomega.8b01683.

Woods, R. and Chappelle, R. (2000). "Restrained Electrostatic Potential Atomic Partial Charges for Condensed-Phase Simulations of Carbohydrates." *Journal of Molecular Structure: THEOCHEM* 527.1–3, pp. 149–156. DOI: 10.1016/S0166-1280(00)00487-5.

World Health Organization (2016). *Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection: Recommendations for a Public Health Approach. 2nd edition.* ISBN: 978 92 4 154968 4.

World Health Organization (2017). *HIV Drug Resistance Report 2017*. 2nd. ISBN: 978-92-4-151283-1.

Wright, D. W., Hall, B. A., Kenway, O. A., Jha, S., and Coveney, P. V. (2014). "Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors." *Journal of Chemical Theory and Computation* 10.3, pp. 1228–1241. DOI: 10.1021/ct4007037.

Xiang, Z. and Honig, B. (2001). "Extending the Accuracy Limits of Prediction for Side-Chain Conformations." *Journal of Molecular Biology* 311.2, pp. 421–430. DOI: 10.1006/jmbi.2001.4865.

Xin, F. and Radivojac, P. (2012). "Post-Translational Modifications Induce Significant Yet Not Extreme Changes to Protein Structure." *Bioinformatics* 28.22, pp. 2905–2913. DOI: 10.1093/bioinformatics/bts541.

Yamashita, H., Endo, S., Wako, H., and Kidera, A. (2001). "Sampling Efficiency of Molecular Dynamics and Monte Carlo Method in Protein Simulation." *Chemical Physics Letters* 342.3, pp. 382–386. DOI: 10.1016/S0009-2614(01)00613-3.

Yamazaki, T., Nicholson, L. K., Torchia, D. A., Wingfield, P., Stahl, S. J., Kaufman, J. D., Eyermann, C. J., Hodge, C. N., Lam, P. Y. S., Ru, Y., Jadhav, P. K., Chang, C. H. C., and Weber, P. C. (1994). "NMR and X-Ray Evidence That the HIV Protease Catalytic Aspartyl Groups Are Protonated in the Complex Formed by the Protease and a Non-Peptide Cyclic Urea-Based Inhibitor." *Journal of the American Chemical Society* 116.23, pp. 10791–10792. DOI: 10.1021/ja00102a057.

Yang, X. and Gabuzda, D. (1998). "Mitogen-Activated Protein Kinase Phosphorylates and Regulates the HIV-1 Vif Protein." *Journal of Biological Chemistry* 273.45, pp. 29879–29887. DOI: 10.1074/jbc.273.45.29879.

Yang, X. and Gabuzda, D. (1999). "Regulation of Human Immunodeficiency Virus Type 1 Infectivity by the ERK Mitogen-Activated Protein Kinase Signaling Pathway." *Journal of Virology* 73.4, pp. 3460–3466.

Yilmaz, N. K., Swanstrom, R., and Schiffer, C. A. (2016). "Improving Viral Protease Inhibitors to Counter Drug Resistance." *Trends in Microbiology* 24.7, pp. 547–557. DOI: 10.1016/j.tim.2016.03.010.

Young, T. P., Parkin, N. T., Stawiski, E., Pilot-Matias, T., Trinh, R., Kempf, D. J., and Norton, M. (2010). "Prevalence, Mutation Patterns, and Effects on Protease Inhibitor Susceptibility of the L76V Mutation in HIV-1 Protease." *Antimicrobial Agents and Chemotherapy* 54.11, pp. 4903–4906. DOI: 10.1128/aac.00906-10.

Yu, Y., Wang, J., Shao, Q., Shi, J., and Zhu, W. (2015). "Effects of Drug-Resistant Mutations on the Dynamic Properties of HIV-1 Protease and Inhibition by Amprenavir and Darunavir." *Scientific Reports* 5.1, p. 10517. DOI: 10.1038/srep10517.

Zeke, A., Bastys, T., Alexa, A., Garai, Á., Mészáros, B., Kirsch, K., Dosztányi, Z., Kalinina, O. V., and Reményi, A. (2015). "Systematic discovery of linear binding motifs targeting an ancient protein interaction surface on MAP kinases." *Molecular Systems Biology* 11.11, p. 837. DOI: 10.15252/msb.20156269.

Zhang, J., Zhou, B., Zheng, C.-F., and Zhang, Z.-Y. (2003). "A Bipartite Mechanism for ERK2 Recognition by its Cognate Regulators and Substrates." *Journal of Biological Chemistry* 278.32, pp. 29901–29912. DOI: 10.1074/jbc.m303909200.

Zhang, Y.-Y., Wu, J.-W., and Wang, Z.-X. (2011). "A Distinct Interaction Mode Revealed by the Crystal Structure of the Kinase p38α with the MAPK Binding Domain of the Phosphatase MKP5." *Science Signaling* 4.204, ra88–ra88. DOI: 10.1126/scisignal.2002241.

Zhou, T., Sun, L., Humphreys, J., and Goldsmith, E. J. (2006). "Docking Interactions Induce Exposure of Activation Loop in the MAP Kinase ERK2." *Structure* 14.6, pp. 1011–1019. DOI: 10.1016/j.str.2006.04.006.

Ziermann, R., Limoli, K., Das, K., Arnold, E., Petropoulos, C. J., and Parkin, N. T. (2000). "A Mutation in Human Immunodeficiency Virus Type 1 Protease, N88S, That Causes in vitro Hypersensitivity to Amprenavir." *Journal of Virology* 74.9, pp. 4414–4419. DOI: 10.1128/JVI.74.9.4414-4419.2000.

Zwanzig, R. W. (1954). "High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases." *The Journal of Chemical Physics* 22.8, pp. 1420–1426. DOI: 10.1063/1.1740409.